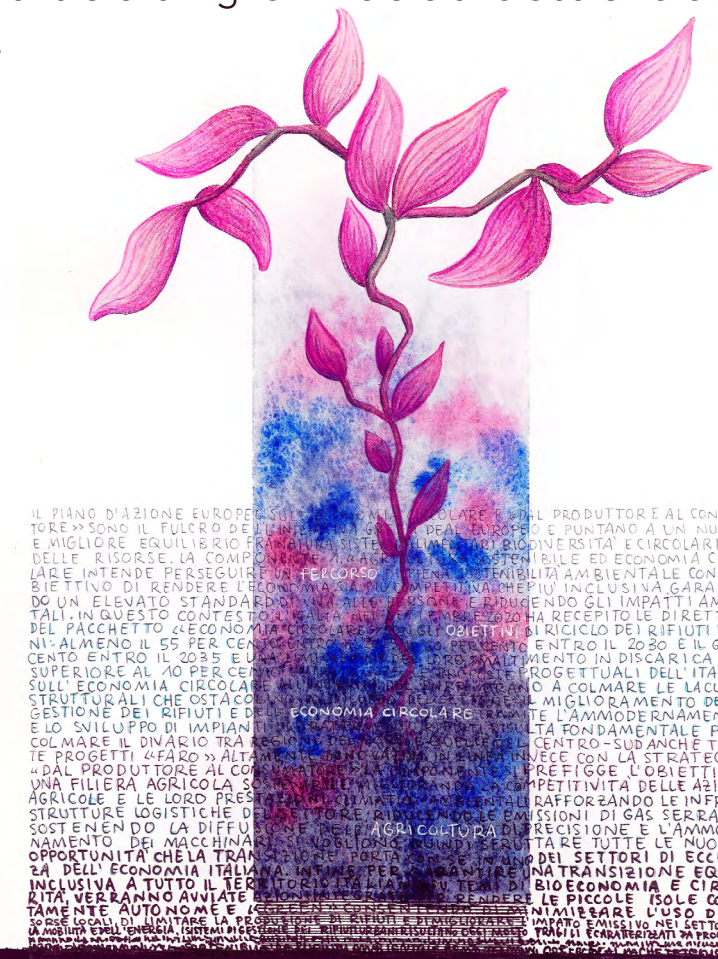
COLLANA PISIA

La governance dei dati pubblici
Testi, contesti e politiche pubbliche

a cura di
Maria Stella Righettini e Stefano Sbalchiero



PADOVA UP
PADOVA UNIVERSITY PRESS

PISIA

Collana del Master in Innovazione, progettazione e valutazione delle politiche e dei servizi

Prima edizione 2022 Padova University Press

Titolo originale *La governance dei dati pubblici. Testi, contesti e politiche pubbliche*

Università degli Studi di Padova
via 8 Febbraio 2, Padova
www.padovauniversitypress.it

Progetto grafico: Padova University Press
Impaginazione: Padova University Press

In copertina: disegno di Giulia David

ISBN 978-88-6938-318-2

La governance dei dati pubblici. Testi, contesti e politiche pubbliche

Come usare i dati testuali a supporto della capacità di policy, della capacità amministrativa e della qualità dei servizi pubblici

a cura di

Maria Stella Righettini
Stefano Sbalchiero

PADOVA UP

Indice

INTRODUZIONE
L'approccio *text-as-data* (TasD) per le politiche pubbliche e le pubbliche amministrazioni

Maria Stella Righettini[1]

L'approccio *text-as-data*, (TasD) per migliorare la governance, le politiche pubbliche e i servizi

Il testo può essere considerato come «il più pervasivo e certamente il più persistente artefatto del comportamento politico», amministrativo e giudiziario (Monroe et al. 2008, p. 351). Nelle moderne forme di governance multilivello, gli attori, pubblici e privati, che collaborano alla produzione di politiche pubbliche e di servizi di rilevanza pubblica, producono grandi quantità di testi che racchiudono decisioni formali e informali di varia natura: politica, amministrativa, giurisdizionale. Il formato testuale caratterizza la produzione decisionale delle istituzioni di governo nazionale, regionale e locale, delle istituzioni comunitarie, delle autorità indipendenti di regolazione, delle amministrazioni sanitarie, delle imprese, *profit* e non *profit*, che gestiscono servizi di pubblica utilità e che partecipano ai bandi di fornitura di servizi alla persona. A questi si aggiungono i dati testuali prodotti dai soggetti che, a vario titolo, partecipano in nome

[1] Professoressa associata di Governance e Valutazione delle politiche pubbliche e Valutazione delle performance e dei servizi presso il Dipartimento di Scienze Politiche, Giuridiche e Studi Internazionali dell'Università degli Studi di Padova e Direttrice del Master Interregionale di II Livello "Innovazione, progettazione e valutazione delle politiche e dei servizi. Agenda 2030 - PISIA" presso l'Università degli Studi di Padova.

e per conto della società civile ai processi di policy e decisionali (associazioni, comitati e altro). La crescente complessità dei processi decisionali in contesti sempre più mutevoli rende necessario disporre di strumenti adeguati all'analisi, gestione e valutazione delle politiche anche e soprattutto nella prospettiva della sostenibilità (Agenda 2030). La sottoscrizione dell'Agenda dell'ONU per la sostenibilità impegna i governi anche ad un miglioramento della governance della sostenibilità. Da qui l'esigenza di ampliare e rafforzare la capacità di analisi e utilizzo dei dati e di sottoporre una varietà di fonti e testi all'analisi testuale, sia a supporto di un policy making più integrato, efficiente ed efficace, e sia in funzione della valutazione degli effetti e degli impatti prodotti dalle politiche sulla società. Sempre più pervasivi processi di digitalizzazione hanno accresciuto la disponibilità di grandi corpora di dati testuali digitalizzati (*big data*), ed hanno fatto registrare un'esplosione di metodi e tecniche per raccogliere e analizzare questo tipo di dati.

Dal punto di vista metodologico, trattare in modo automatizzato i testi come dati con tecniche di *machine learning* facilita il confronto e la comparazione, diacronica e sincronica, tra decisioni dello stesso tipo, tra differenti assetti istituzionali, tra contesti giuridico-economici e culturali di riferimento. Questa opportunità impone tuttavia alcune riflessioni di fondo, sia sulla semantica, ovvero sul significato diverso che lo stesso termine può assumere in contesti differenti, sia sulla grammatica e sulla struttura differente dei linguaggi sottoposti ad analisi. In particolare, nell'ambito della scienza politica e dell'analisi delle politiche l'approccio *text-as-data* (TasD), ovvero al trattamento dei testi come dati, è stato utilizzato in ricerche comparate in vari ambiti di policy (Gilardi et al. 2020). Pur affondando le radici nell'analisi testuale di tipo classico, che ha intenti prevalentemente descrittivi ed è basata sulla struttura del testo e le sue regole formali, l'approccio TasD ha cercato di coniugare la capacità di classificazione e mappatura dei contenuti testuali con le principali teorie e framework analitici politologici. Il fine è quello di testare la validità di tali teorie e approcci e fornire una spiegazione plausibile del cambiamento in atto nei sistemi politici, amministrativi e nei processi di policy guardando agli output e outcome del sistema. L'analisi testuale si è occupata sia dei cambiamenti di politics rinvenibili nei contenuti dei discorsi politici dentro le sedi istituzionali (parlamenti e governi) e nella società (partiti, manifesti elettorali, comunicazione politica), (Slapin et al. 2008) e sia dei cambiamenti di policy. L'utilizzo dei TasD nella policy analysis ha permesso di esplorare e tracciare l'evoluzione dei valori di cui sono

portatori i policy maker in settori specifici di intervento, i contenuti dei processi di cambiamento favoriti da specifici strumenti di policy (consultazioni), (Righettini 2021) e l'analisi degli output e outcome delle politiche regolative (Righettini et al. 2017 a, b).

L'approccio TasD si avvale di tecniche automatizzate di elaborazione che consentono di processare grandi moli di dati evitando difformità ed errori di lettura e mappatura che dipendono dalle operazioni manuali e dalla discrezionalità dei ricercatori ed operatori coinvolti. Alcuni software – quelli che saranno presentati nel capitolo 1 sono fra questi – permettono anche di processare dati testuali contenuti in banche dati più piccole e più diffuse nel settore pubblico. Con l'analisi automatizzata, il contributo del ricercatore in carne ed ossa non scompare, ma viene valorizzato nelle fasi iniziali di creazione del corpus e codificazione dei dati (*coding*), e nella fase finale di interpretazione dei risultati. Gli algoritmi, creati dai ricercatori in accordo con i tecnici e con i policy maker, fanno il resto.

La scienza politica e l'analisi delle politiche pubbliche forniscono ai decisori pubblici un contributo rilevante all'individuazione dei problemi, alla formulazione di ipotesi di lavoro e all'individuazione delle domande di ricerca che guidano i bisogni di conoscenza nelle varie fasi del policy making: l'esplorazione e individuazione delle banche dati disponibili; la creazione del corpus (standardizzazione, lemmatizzazione, ecc.); la codificazione dei contenuti dei testi analizzati secondo le dimensioni ritenute più rilevanti (valori, attori, problemi, processi, strumenti, beneficiari e risultati).

Lo sviluppo del text mining e la sua applicazione alle decisioni pubbliche hanno contribuito allo sviluppo della comparazione tra casi, che, assieme allo studio di caso, è uno dei metodi più utilizzati nella ricerca di policy. Alcuni ambiti di ricerca accademica sono già fortemente sviluppati. È il caso, ad esempio, degli studi sulle agende parlamentari – *The comparative Agendas Project* (Jones et al. 2005). Questo filone di ricerche analizza diacronicamente e in modo comparato i contenuti delle leggi approvate dai parlamenti nazionali. Esso è passato da un'iniziale classificazione (*coding*) manuale dei testi, molto costosa in termini di tempo e limitata ad alcuni paesi, a procedure automatizzate che hanno permesso grandi avanzamenti, sia per il numero di paesi coinvolti nella ricerca e sia per la riduzione dei costi. Un altro filone di ricerche in crescita è quello che si ispira alla teoria di Elinor Ostrom della *Institutional Grammar* (IGRI), o grammatica delle istituzioni (Ostrom 2005). Questo filone accademico di studi considera le regole come dati e si occupa in particolare di analizzare

le dinamiche o i *pattern* regolativi attraverso la struttura dei testi legislativi o della regolazione indipendente, allo scopo di favorire l'apprendimento di policy. Guardando alla produzione di testi regolativi si può tentare di stabilire una relazione tra struttura del testo, modalità e processi decisionali ed esiti della regolazione stessa. Si ritiene, ad esempio, che quanto più la regolazione sia complessa nella struttura grammaticale e del testo, tanto più essa sia di difficile implementazione (Siddiki et al. 2022). L'ambizione dei ricercatori è che da un'analisi di questo tipo si possa partire per affinare la capacità di definizione dei problemi di policy, di individuazione degli strumenti attuativi fornita da decisori e stakeholder lungo il ciclo di policy e, infine, migliorare processi e risultati della regolazione in una prospettiva multilivello. L'approccio TasD può essere applicato a testi prodotti da istituzioni di vario tipo e lungo tutte le fasi del ciclo di policy: agenda setting, consultazioni, adozione legislativa o amministrativa e valutazione. Può essere analizzato, ad esempio, il grado di salienza/rilevanza attribuita dai policy maker a determinati problemi (policy issue) nel corso del tempo, non solo attraverso testi di decisioni vere e proprie, ma anche tramite documenti di rendicontazione come relazioni annuali o report di varia natura (Righettini et al. 2017a, 2021). La salienza/rilevanza di un certo tema può essere inoltre messa in relazione alla variazione di altri fattori: tipo di soggetto incaricato della decisione, luogo in cui la decisione è assunta, tempo e altri fattori di interesse. La salienza pubblica di un tema è inoltre rilevante per capire quanto questa condizioni comportamenti e decisioni pubbliche su un dato problema a scapito di altri (*bias*), e ne influenzi gli esiti (Sunstein 2002, Leonard et al. 2008). Recentemente, è stato analizzato come il *turn-over* dei presidenti delle autorità di regolazione nel settore delle telecomunicazioni influisse nel medio periodo sul livello di salienza attribuito al tema della 'tutela del consumatore' (*consumer protection*), (Righettini et al. 2021) e come, nonostante una regolazione europea uniforme che attribuisce al tema una rilevanza crescente, le autorità di regolazione nazionale avessero un'attenzione condizionata da fattori 'locali' e 'personali', quali i presidenti in carica. Altro sviluppo dell'approccio TasD si è avuto nell'ambito degli studi sulla *diffusione di policy (policy diffusion)*. Tali ricerche rispondono alla seguente domanda: come una policy (innovativa) che si sviluppa in un certo ambito territoriale si diffonde e viene trasferita in altri contesti territoriali e di governo? (Gilardi et al. 2021, Gilardi et al. 2018). Le ricerche studiano caratteri comuni di specifiche regolazioni statali e/o regionali e analizzano come il processo di diffusione faciliti un adattamento della definizione del proble-

ma oggetto di regolazione (ad esempio divieto di fumo), degli strumenti e delle modalità di implementazione (governance) ai vari contesti culturali ed amministrativi. L'analisi testuale permette di andare oltre l'apparente 'convergenza' dei governi sulla medesima missione regolativa e permette di capire come e con quali differenze i policy maker adottano e implementano una certa regolazione. Un più recente studio ha adottato tecniche di *topic detection* (cfr. capitolo introduttivo Sbalchiero) per sviluppare una metanalisi, cioè un'analisi di secondo livello effettuata sulla letteratura, che cerca di individuare sia le differenze delle risposte fornite dai governi alla crisi Covid-19, e sia le ragioni per cui alcuni strumenti sono stati preferiti ad altri (Capano et al. 2020). La ricerca confronta il mix di strumenti adottati dai paesi per contrastare la pandemia analizzando la legislazione di emergenza e i provvedimenti governativi adottati tra dicembre 2019 e aprile 2020. La digitalizzazione e l'uso dei social media hanno permesso di utilizzare l'approccio TasD nell'ambito della teoria del *policy feedback* per analizzare le reazioni pubbliche a determinate decisioni (leggi, regolazioni), a fatti/eventi politico istituzionali (elezioni), (Flores 2017) e per valutare il gradimento relativo a modificazione di servizio (digitalizzazione). L'analisi testuale è stata applicata, infine, anche a corpora testuali molto specifici, non inerenti direttamente ai processi decisionali, ma relativi a fenomeni socialmente rilevanti per le politiche pubbliche. È il caso, ad esempio, dei 'manuali di cittadinanza' redatti negli Stati Uniti tra il 1921 e il 1996 da gruppi della società civile che si occupano dell'educazione civica degli immigrati (Goodman 2021) per capire come evolve il concetto di 'buon americano' nel tempo.

Dati testuali e produzione di valore pubblico. Riesplorare la miniera

Quando si tratta di dati testuali, le scienze politiche ed amministrative si trovano di fronte a una vera e propria miniera di risorse disponibili, ancora in gran parte da esplorare, utilizzabili per migliorare la qualità delle decisioni e dei processi. L'interesse per i processi di digitalizzazione e la produzione di moli massicce di testo digitalizzato sta lentamente affermandosi come opportunità offerta ai policy maker per accrescere la capacità di comprensione, trattamento e correzione delle decisioni pubbliche nelle varie fasi del ciclo di policy. Tuttavia, la difficoltà di uscire dalla logica meramente descrittiva dei dati numerici e degli indicatori statistici costituisce ancora un potente freno allo sviluppo dell'analisi testuale qua-

li-quantitativa nei contesti di policy. Siamo al corrente di quanti dibattiti legislativi, testi di legge, consultazioni, atti ispettivi, sentenze delle corti, documenti di programmazione, atti amministrativi, rapporti annuali e valutazioni sono stati prodotti nel tempo, ma troppo poco sappiamo ancora sull'evoluzione del loro contenuto e nulla possiamo apprendere dal cambiamento che tale contenuto racchiude e custodisce.

Lo scopo del volume è duplice: fornire un quadro analitico legato alla governance dei dati nel settore pubblico, e illustrare esempi di applicazioni concrete dell'approccio TasD che vadano oltre la retorica tecnocratica della digitalizzazione. Ciò allo scopo di favorire un diverso approccio al tema dei dati testuali disponibili presso governi, pubbliche amministrazioni, magistrature, imprese pubbliche e organizzazioni della società civile, e contribuire così a migliorare la qualità delle decisioni, delle politiche e dei servizi pubblici.

Chi scrive ritiene che, contrariamente a quanto generalmente si crede e si auspica, la valorizzazione dei dati pubblici e dei *big data* nei processi decisionali e nella p.a. non dipenda unicamente da competenze informatiche o di *data science* (Maciejewski 2017), bensì dall'integrazione di competenze tecniche con competenze applicative di policy design e dallo sviluppo di *soft skill* che favoriscano l'innovazione. Tale integrazione, tanto evocata quanto scarsamente coltivata nei corsi di formazione delle p.a., favorisce lo sviluppo di *competenze di metodo* tra gli operatori pubblici, modifica lo sguardo sulle risorse disponibili e il valore che esse racchiudono.

La capacità di sviluppare competenze di analisi testuale è troppo spesso compromessa da processi di digitalizzazione e applicazione di algoritmi di AI alle attività pubbliche che prescindono dalla conoscenza della complessità dei processi decisionali e dalla capacità di mettere in atto strategie che ne facilitino il miglioramento (Dente 2011). Si può immaginare ed auspicare per il futuro che le pubbliche amministrazioni coltivino al proprio interno unità di specialisti (TasD *Unit*) dediti all'analisi dei testi prodotti all'interno dei propri processi legislativi e di policy, anche in chiave comparativa.

Il potenziale dell'approccio TasD è particolarmente evidente se si guarda alla sua applicazione non tanto o soltanto nell'ambito degli studi accademici, quanto alle sue applicazioni all'interno della pubblica amministrazione (Maciejewski 2017). Applicazioni di successo nelle pubbliche amministrazioni riguardano in particolare tre ambiti: la capacità di accrescere l'accuratezza delle decisioni verso l'esterno; l'accelerazione

e l'interoperabilità dei flussi informativi interni alle organizzazioni, per migliorarne le performance; la riduzione degli oneri amministrativi per stakeholder e cittadini e la riduzione di ingiustificate diseguaglianze di trattamento. Utilizzando le parole del premio Nobel Kahneman (2021), possiamo dire che attraverso l'analisi del contenuto delle decisioni è possibile riconoscere e misurare il 'rumore' prodotto dalle decisioni organizzative, ovvero la varianza con cui lo stesso caso viene trattato nell'applicazione di normative tecniche standardizzate e regolate dalla legge.

La difformità interpretativa, per quanto sia fisiologica, ed entro certe soglie accettabile, se troppo elevata può creare problemi di reputazione e legittimazione all'organizzazione stessa verso l'esterno. Nel caso delle pubbliche amministrazioni – siano esse amministrative, giudiziarie e/o sanitarie, solo per fare alcuni esempi – l'elevata varianza di trattamento/decisione dello stesso caso può alimentare un tasso elevato di ricorsi giurisdizionali, con conseguente allungamento dei tempi, la perdita di efficacia dell'intervento o anche la perdita di fiducia dei cittadini nelle istituzioni, nella loro trasparenza ed equità (Kahneman et al. 2021). Ciò vale anche nell'ambito dei bandi pubblici (nazionali ed europei) e dei processi di reclutamento e verifica delle conoscenze dei candidati dei concorsi pubblici (si pensi ai concorsi per magistratura). Misurare le cause di varianza/difformità dei contenuti delle decisioni sul medesimo oggetto può supportare interventi/riforme che mirano ad un maggiore coordinamento ed efficientamento del sistema e ad accrescerne la reputazione.

Nell'ambito delle pubbliche amministrazioni il TasD *approach* può essere utilizzato in tre modi: per ricostruire l'evoluzione storica delle decisioni e imparare da quello che è stato fatto e correggere eventuali errori o distorsioni; per migliorare la qualità delle decisioni *real-time*, cioè nel momento in cui i processi decisionali si svolgono e la tempistica delle decisioni è particolarmente importante; infine, a scopi predittivi, per migliorare la comprensione delle cause dei problemi, dei probabili esiti futuri e per migliorare efficacia ed efficienza delle decisioni e il design delle politiche (Maciejewski 2017). L'analisi testuale consentirebbe, infine, anche un approccio applicativo alle politiche di *better regulation*, alla valutazione dell'impatto regolativo, individuando la distanza tra contenuti dei processi di consultazione e contenuti della regolazione per mettere a fuoco i miglioramenti della regolazione stessa come esito dei processi partecipativi.

Data Governance e apprendimento di policy

Per *governance* intendiamo l'insieme degli assetti organizzativi, delle

regole e dei processi attraverso i quali un ente, istituzione o impresa, in modo condiviso e non gerarchico, coinvolge altri soggetti (stakeholder e beneficiari), formula e cerca di raggiungere la propria missione istituzionale. La governance è uno stile di governo caratterizzato dalla collaborazione costruttiva tra attori che hanno obiettivi e poste in gioco non sempre convergenti.

> «La data governance può essere definita come un sistema condiviso di produzione di politiche/servizi che considera i dati come un asset strategico e assegna obiettivi, individua strumenti e metodi in merito alla raccolta, trattamento, accesso, condivisione e utilizzo adeguato dei dati (anche testuali) per lo svolgimento di funzioni istituzionali (pubbliche o aziendali)» (Alhassan et al. 2016).

Una recente rassegna bibliografica degli articoli scientifici pubblicati sull'argomento rivela che il tema della data governance è più frequentemente associato all'analisi del sistema di regole e responsabilità che governano le scelte di policy dei dati di un ente o organizzazione, dei processi e delle procedure alla base dell'utilizzo dei dati, assieme al tema della qualità dei dati stessi.

Le dimensioni della data governance che rilevano ai fini non soltanto della ricerca e degli approfondimenti possibili, ma anche dal punto di vista funzionale, sono pertanto le seguenti:

- la qualità della raccolta, relativa ai metodi utilizzati;
- la qualità dell'archiviazione dei dati (fisica o digitale);
- l'interoperabilità dei dati, internamente a un'organizzazione o tra organizzazioni di diversa natura;
- le regole di accesso ai dati e la facilità con cui ciò è consentito (*data protection*);
- il grado di trasparenza sui processi di raccolta e utilizzo dei dati, in una parola il sistema di *accountability* istituzionale dell'ente e i diritti tutelati;
- l'attendibilità dei dati, ovvero la fiducia pubblica di cui godono i dati e le elaborazioni che essi rendono possibili (accesso, sicurezza, riservatezza, sostenibilità);
- le finalità pubbliche perseguite attraverso i dati, relative a funzionalità interne e/o ai bisogni soddisfatti nell'ambito delle stesse (Cerrillo-Martínez et al. 2021).

Il termine *governance* differisce da quello di *management* (*data management*), e si riferisce alle decisioni che vanno prese in merito non solo alla gestione (raccolta, trattamento, accesso e condivisione), ma anche

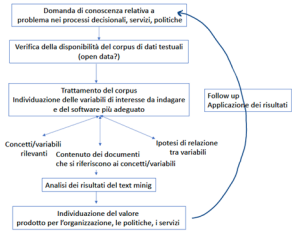
alle finalità pubbliche da perseguire attraverso la gestione dei dati (procedure di utilizzo, trattamento e qualità dei dati adeguata al fine, alla missione di policy o istituzionale).

Nel sistema pubblico, la data governance attiene all'innovazione dei processi decisionali e valutativi, al sistema di *accountability* istituzionale dell'ente, alla qualità e attendibilità dei dati, alle finalità pubbliche perseguite, ai bisogni soddisfatti e ai diritti tutelati (accesso, sicurezza, trasparenza, riservatezza, sostenibilità) attraverso le politiche.

Sempre più spesso parliamo dei dati raccolti dalle istituzioni pubbliche relativi a dati personali o a fenomeni sociali che rilevano per le decisioni di policy. È questo, ad esempio, il caso dei dati sanitari e dei dati relativi al mercato del lavoro. Meno frequentemente si affronta il tema della grande mole di dati direttamente prodotti dalle istituzioni pubbliche e delle conoscenze da essi ricavabili: è questo, ad esempio, il caso della regolazione, degli appalti, o delle decisioni amministrative.

I dati prodotti e di cui dispongono le istituzioni pubbliche e le p.a. possono essere di due tipi, numerici e testuali. I dati numerici riassumono e illustrano tendenze quantitative (in valore assoluto o percentuale) di un fenomeno: si prenda ad esempio la produzione legislativa distribuita lungo le varie legislature o il numero di autorizzazioni rilasciate da un comune. I dati testuali sono invece riferiti al formato e ai contenuti dei vari documenti prodotti dalle pubbliche amministrazioni: piani, programmi, progetti, delibere, documenti di valutazione, comunicazione interna e/o esterna, ordinanze e sentenze. La rappresentazione numerica delle attività del settore pubblico è prevalente rispetto alla rappresentazione dei contenuti testuali e della loro evoluzione, soprattutto se consideriamo corpora di testi prodotti e/o archiviati e riferibili a periodi medio-lunghi.

Fig. 1 Dai testi ai contesti amministrativi, di policy e di servizio, alla produzione di valore.



Se guardiamo ai recenti sviluppi delle p.a. durante la pandemia, possiamo rilevare un interesse crescente alla disponibilità di dati per formulare e reindirizzare le politiche, per valutare le performance sociali,

ambientali ed economiche. Grazie all'Agenda 2030 ed al suo recepimento a tutti i livelli di governo, prima e durante la pandemia, si assiste alla produzione di un volume crescente di dati statistici – numero di piani e delibere approvati, indagini di *customer* e sentenze – adottati in una data materia o da un determinato ufficio. Tuttavia, anche se i dati testuali rappresentano l'output per antonomasia delle pubbliche amministrazioni, l'analisi del loro contenuto resta sporadica e scarsamente considerata come un potenziale *driver* di miglioramento delle performance amministrative e dei risultati di policy dell'ente.

I dati testuali sono una delle modalità, forse la principale, attraverso cui le istituzioni di governo e le organizzazioni lavorano e comunicano tra di loro, intrattengono rapporti con il mondo esterno e stabiliscono modalità di interazione con cittadini e stakeholder.

La governance dei dati testuali può generare *apprendimento* (*human learning*) in varie direzioni: lo studio del contenuto delle decisioni passate e l'analisi automatizzata dei testi può accrescere la conoscenza delle caratteristiche ritenute più salienti degli output testuali prodotti dagli enti pubblici, permette di individuare eventuali distorsioni nel sistema decisionale, rafforza la capacità di correzione delle decisioni future. L'approccio TasD può, inoltre, accrescere la capacità di confronto e coordinamento tra decisori che devono concorrere al raggiungimento della medesima missione; può favorire una maggiore capacità di valorizzazione delle attività e dei processi organizzativi in funzione della revisione della missione istituzionale; può accrescere la capacità di mappatura e comprensione dei bisogni e/o interessi di altri attori (stakeholder e/o beneficiari) coinvolti nel policy making; può accrescere la capacità dell'organizzazione di comunicare più efficacemente sia all'interno e sia all'esterno (comunicazione istituzionale).

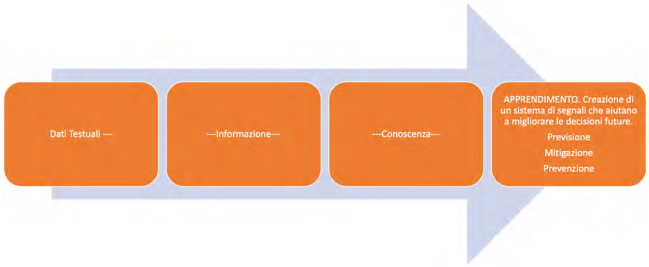
Il miglioramento della governance dei dati testuali può esercitare effetti sia sul sistema di relazioni interistituzionali, tra organi e livelli di governo, sia tra uffici all'interno della stessa organizzazione, e sia sul rapporto tra istituzioni, organizzazioni della società civile, stakeholder e cittadini.

Tali miglioramenti non sono un fatto meramente tecnico, legato cioè all'introduzione di software di analisi testuale e al loro utilizzo, ma dipendono dalla *capacità* degli operatori di *apprendere* come l'utilizzo dei software possa migliorare la funzionalità della decisione o del servizio. Come lo strumento possa essere calato nel contesto amministrativo e di policy favorendo la diffusione di nuove modalità operative e migliori risultati per l'ente e per i cittadini.

Questo volume è una guida all'analisi testuale secondo il framework di policy e le sue molteplici applicazioni e declinazioni nel settore pubblico, e risponde a una logica operativa, fornendo spunti dalle applicazioni effettuate da operatori pubblici nell'ambito dei *project-work* del Master PISIA.

Le applicazioni si sono avvalse del contributo fornito dalla ricerca scientifica in quest'ambito, e il volume cerca di andare oltre la mera illustrazione delle singole esperienze applicative. Esso fornisce un quadro interpretativo e metodologico del potenziale innovativo dello strumento, unitamente alle istruzioni per rendere l'utilizzo di tali strumenti più accessibile per progettare il miglioramento delle performance delle politiche e della capacità amministrativa.

Fig. 2 Dai dati testuali all'apprendimento amministrativo e di policy.



Alcune (ma non esclusive) chiavi interpretative (framework) adottabili per supportare la formulazione delle domande e l'applicazione dell'analisi testuale sono le seguenti:

- miglioramento della qualità della programmazione, ossia analizzare i documenti di programmazione come dati per individuare le caratteristiche del design di politiche '*problem based*' e '*place based*'. A questo scopo è utile interrogarsi e indagare non solo sulle differenze e/o convergenze che caratterizzano le decisioni pubbliche e le loro più dirette manifestazioni (piani, documenti, decisioni, report), ma anche sulle attività a supporto della programmazione in un certo ambito di intervento, su un certo tema e all'interno di un dato territorio. Capire da cosa siano determinate le differenze nei processi e negli output della programmazione può essere utile per mettere a fuoco le implicazioni sui processi di implementazione delle decisioni stesse e sui risultati raggiunti. In quest'ambito rientrano le ricerche sui programmi multilivello nei vari settori di policy e sul loro coordinamento con Agenda

2030, sui processi partecipativi, sui tavoli di concertazione, sulle piattaforme a supporto della programmazione;
- miglioramento della qualità della regolazione (*better regulation*) e *regulatory refit*. La policy di *better regulation* si basa sull'assunto che la regolazione (leggi e delibere) debba esse efficace ed efficiente, riducendo oneri a carico dei destinatari (imprese e cittadini) e semplificando le procedure in modo da rendere l'adeguamento e il cambiamento richiesto ai comportamenti comprensibile e quindi facile da implementare e controllare. L'analisi testuale può ricostruire la grammatica della regolazione, può esaminare corpora molto ampi di testi regolativi (ad esempio europei, nazionali e regionali) individuando dimensioni ritenute rilevanti del design regolativo e confrontando la presenza/assenza di tali dimensioni con i risultati attesi e con quelli effettivamente raggiunti a distanza di tempo;
- orientamento dal lato della domanda di policy o di servizio, ossia interrogarsi sui contenuti delle domande provenienti dai cittadini e non solo sul numero delle domande stesse. In questo ambito rientrano le ricerche sui social, di *sentiment*, sugli strumenti (digitali e non), di *customer*, le analisi sui contenuti dei processi partecipativi;
- estrazione di valore dalle banche dati esistenti e costruzione di nuove banche dati digitalizzate per migliorare l'accesso e l'utilizzo di informazioni, la conoscenza dei processi decisionali e dei servizi offerti, e favorire decisioni informate (*evidence-based*);
- valutazione della progettazione e degli impatti delle politiche e dei servizi su stakeholder e beneficiari ultimi. Questo approccio considera l'analisi testuale applicata, per esempio, ai documenti di valutazione (ex ante, in itinere ed ex post), che hanno spesso una natura scarsamente strutturata, e contengono informazioni rilevanti per i policy maker soprattutto per individuare distorsioni, errori e possibili miglioramenti dei processi e delle decisioni. L'analisi testuale ex ante può servire, ad esempio, alla migliore selezione di decisioni e progetti finanziabili. La valutazione in itinere ad analizzare i contenuti di report intermedi relativi alle implementazioni dei progetti. Infine, la valutazione ex post favorisce l'analisi testuale delle rendicontazioni e la rispondenza di tali rendicontazioni a criteri o requisiti predefiniti;
- valutazione dell'uso della discrezionalità nelle pubbliche amministrazioni, con riferimento in particolare alla possibilità di identi-

ficare il grado di varianza delle decisioni/valutazioni sullo stesso caso.

Le chiavi interpretative individuate, seppur non esclusive, possono guidare processi di conoscenza utili per l'ente e favorire un processo di apprendimento attraverso il text mining. Tale apprendimento può avvenire a vari livelli e in varie fasi del ciclo amministrativo e di policy. Per *apprendimento* intendiamo la creazione di un sistema di segnali (cruscotto) che aiutano a migliorare le decisioni future, ovvero la *capacità* di svolgere determinate attività in modo diverso rispetto al passato. Il miglioramento della qualità dei processi, degli output e degli outcome è valutabile attraverso un maggiore grado di efficienza, efficacia e soddisfazione interna e soprattutto esterna di utenti e stakeholder.

Dati testuali per rafforzare la capacità di governance e amministrativa

Rafforzare la *policy capacity* significa rafforzare la capacità di mettere a fuoco problemi condivisi, disegnare programmi integrati, adottare strumenti coerenti ed efficaci di policy, garantire l'allocazione controllata delle risorse sulla base dei risultati attesi, garantire la soddisfazione dei clienti-utenti dei servizi e creare feedback con gli utilizzatori di reti, piattaforme, applicazioni.

L'apprendimento è reso possibile dalla domanda di partenza, ossia dalla capacità di mettere a fuoco sia il problema conoscitivo e operativo che si intende affrontare utilizzando l'analisi testuale e sia la tecnica che s'intende adottare per generare una conoscenza utile al decisore in vari ambiti. Sono esempi di domande: cosa cerchiamo nei testi? Come e perché lo cerchiamo? Quale può essere l'utilità dei risultati prodotti da tecniche di analisi automatica o semiautomatica? Vogliamo solo descrivere la serie temporale degli eventi e le linee di tendenza prevalenti? Vogliamo capire la loro distribuzione nel tempo in relazione ad eventi esterni specifici? Vogliamo capire dove si colloca un certo cambiamento decisionale, cosa lo determina e quali effetti produce?

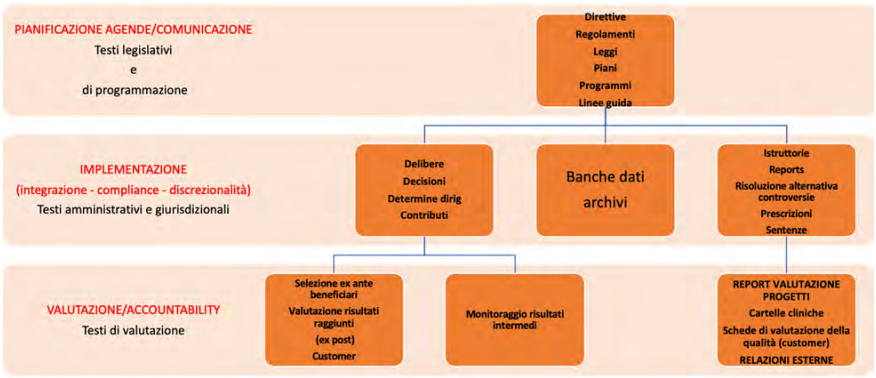
Tali domande possono sorgere in fasi diverse del ciclo di policy: design, implementazione e valutazione di una politica, programma o servizio. Le domande di conoscenza riguardano attività tra loro diverse ma interconnesse, quali:

- comunicazione interna/esterna;
- definizione del problema: contenuti, idee e valori, estrazione a

vari livelli politico, amministrativo, tecnico;
- decisione (ai vari livelli): legislativa, amministrativa, tecnica;
- partecipazione multilivello e multi-stakeholder;
- trasparenza: cosa si decide e con che grado di 'rumore' nelle organizzazioni;
- stili di governance e implementazione;
- feedback: reazione dei destinatari delle decisioni e delle politiche;
- valutazione esiti, outcome, efficacia delle politiche, valutazione performance, valutazione qualità dei servizi, valutazione quali/quantitativa di dati di *survey*, monitoraggio delle segnalazioni degli utenti.

Per semplificare, possiamo collocare le domande di conoscenza che possono sorgere all'interno di tre macroaree di attività che si collocano lungo il processo di policy: 1. Pianificazione, formazione dell'agenda istituzionale e comunicazione; 2. Implementazione e monitoraggio; 3. Valutazione e *accountability* (Fig. 3). All'interno di ciascuna area si possono individuare output dei processi decisionali che possono diventare oggetto di analisi testuale.

Fig. 3 Ciclo di policy e analisi testuale automatizzata.



Come vedremo nei prossimi capitoli, nell'ambito del ciclo di policy il text mining può essere utilizzato a vari scopi e con diverse funzionalità: esplorazione dei contenuti prevalenti, trend, *cluster* di argomenti nell'ambito del processo di formulazione delle politiche (*leadership*, valori, *topics*), stili di governance, implementazione amministrativa (customizzazio-

ne), trasparenza ed equità delle decisioni tecniche (aspetti tecnici delle decisioni), finanziamenti, valutazione (monitoraggio, valutazione ex ante ed ex post), comunicazione (interna/esterna).

Ogni giorno, una immensa quantità di dati e informazioni è prodotta dalle autorità di governo e dalle pubbliche amministrazioni. Tali decisioni transitano anche in rete, sia sotto forma di comunicazione istituzionale e sia di messaggi in Facebook, Twitter, blog, piattaforme, forum e gruppi di discussione, pagine web e *news*.

La *digitalizzazione* degli archivi (bibliografici, delle biblioteche e degli archivi storici a tutti i livelli, comprese le attività giudiziarie, di governo, delle pubbliche amministrazioni e delle autorità di regolazione indipendenti) sta inoltre lentamente contribuendo a generare nuovi fonti e canali di accesso ai dati. Ma da sola la digitalizzazione non basta.

Si tratta di una massa enorme di informazioni sui *comportamenti linguistici pubblici* che era impensabile fino a qualche tempo fa e rappresenta un'opportunità di analisi preziosa per la *lettura* e la *comprensione* dei processi decisionali, dei loro esiti e dei fenomeni sociali.

Le nostre capacità di gestione e calcolo non sono ancora in grado di sfruttare pienamente tutte le informazioni che potremmo estrarre da questa crescita esponenziale di testi digitalizzati, tuttavia alcuni sforzi, anche in chiave sperimentale, possono essere fatti. Tali sperimentazioni necessitano di essere più sistematicamente inquadrate nell'attività di *routine* delle pubbliche amministrazioni. Per capire meglio le implicazioni dell'analisi testuale sul funzionamento della macchina amministrativa e di policy dobbiamo rifarci all'idea stessa di capacità di policy e amministrativa alla luce del processo continuo di riforma che ha attraversato il settore pubblico negli ultimi due decenni dalla riforma Brunetta (D.lgs. 150/09). La capacità di governance ed amministrativa degli enti è supportata dalla capacità analitica e valutativa (in senso lato e tecnico) delle prestazioni pubbliche che interagisce costantemente con la capacità di pianificazione, la capacità di erogazione di servizi di qualità e la capacità di coordinamento (Melloni et al. 2019). L'analisi testuale si inserisce in questo quadro agendo su più fronti – valutazione, pianificazione, coordinamento ed erogazione di servizi – fornendo elementi di conoscenza e di giudizio altrimenti inaccessibili ai policy maker. In particolare, nelle condizioni di turbolenza ambientale causate dalla pandemia, e in previsione di nuove inaspettate turbolenze future, la conoscenza più approfondita

dei contenuti delle proprie decisioni passate e la capacità di orientare e correggere le decisioni future diventa cruciale per accrescere una reale capacità di mitigazione e adattamento dei governi a tutti i livelli.

Banche dati testuali pubbliche e private: costruire *usable knowledge*

Le banche dati utilizzabili per costruire corpora di testi da analizzare sono molteplici e possono essere utilizzate anche congiuntamente, incrociando cioè dati testuali prodotti a vari livelli di governo o decisionali secondo la domanda di conoscenza. Qui di seguito un elenco delle principali fonti a cui attingere per l'analisi dei testi:

- banche dati internazionali ed europee (leggi, programmi, rapporti, progetti finanziati);
- banche dati nazionali (legislazione, finanziamenti a progetti, delibere, determine dirigenziali);
- banche dati di Agenzie e Autorità indipendenti nazionali ed europee (relazioni annuali, decisioni, bilanci, consultazioni), (ANAC, Privacy);
- banche dati regionali legislative, amministrative e tecniche (determine dirigenziali, decisioni tecniche, valutazioni), consultazioni, comunicazione istituzionale e social, progetti finanziati;
- banche dati locali (decisioni, *customer*, comunicazione social e web, consultazioni);
- banche dati create da processi decisionali e partecipativi multilivello (anche online);
- banche dati aziendali (spesa, decisioni, consultazioni, curricula);
- banche dati del Terzo settore (progetti, decisioni, bandi);
- banche dati giurisprudenziali (sentenze, massime, leggi, contrattazione);
- banche dati di discorsi pubblici ed istituzionali, articoli di stampa, interviste.

Come si vedrà meglio nella sezione metodologica, per la quale si rimanda al capitolo introduttivo (Sbalchiero) del presente volume, una volta creato il corpus di dati testuali di interesse si può procedere seguendo due strade:

- l'individuazione e l'estrazione automatica dai testi di argomenti, inerenti concetti predefiniti (*topic detection*);
- la categorizzazione dei documenti, e di porzioni di testo, contenuti nel database secondo variabili di interesse.

Poiché secondo stime attendibili, circa l'80% delle informazioni in una pubblica amministrazione o in azienda si trovano in forma di testi e solo il 20% è costituita da dati numerici, si intuisce la sfida che deriva dall'interagire con enormi masse di materiali testuali, spesso già disponibili in rete o digitalizzati. Si capisce pertanto l'importanza dell'analisi testuale automatica, che deve essere condotta partendo da domande o ipotesi che a monte mettano in luce il problema che si vuole affrontare per estrarne *informazione capace di produrre valore per l'ente.* La capacità degli operatori di adottare tecniche di AI costituisce una nuova leva nella gestione della conoscenza e nella cosiddetta *intelligenza amministrativa collettiva* (*collective amministrative intelligence*), (Kolbjørnsrud et al. 2016).

La struttura del volume

Il volume, la cui articolazione interna è brevemente illustrata in questa sezione, rappresenta un contributo originale nel suo genere. I casi di analisi testuale automatizzata – realizzata prevalentemente tramite il software libero Iramuteq – presentati nei vari capitoli mostrano come l'approccio TasD possa essere alla portata di tutti, supportare lo sviluppo della capacità decisionale pubblica e favorire la comprensione di teorie e tendenze di natura locale, sistemica e comparata dei sistemi politici ed amministrativi (Gilardi et al. 2021). L'approccio TasD può diventare uno strumento a portata di mano degli operatori per favorire l'innovazione della governance delle politiche, dei processi decisionali, dei servizi e dei processi valutativi e dei loro esiti. I vari capitoli mostrano come un affinamento della *capacità di governance e di policy basata sui dati testuali* da parte dei funzionari che operano nelle organizzazioni costituisca un'opportunità di innovazione di processo e/o di prodotto.

Sono qui di seguito illustrati sinteticamente i contributi presenti nel volume e i principali risultati raggiunti dall'applicazione delle tecniche di text mining, che potranno essere approfondite con la lettura dei rispettivi capitoli.

Il capitolo introduttivo di Stefano Sbalchiero illustra, da un punto di vista metodologico e tecnico, l'evoluzione di approcci e procedure del text mining e le caratteristiche e potenzialità dei programmi più utilizzati nelle applicazioni che sono state fatte nell'ambito del Master PISIA. Sbalchiero spiega come il text mining sia oggi il risultato di una crescente ricerca scientifica e tecnica e del moltiplicarsi di studi, soprattutto in am-

bito linguistico, che hanno incrociato l'informatica e lo sviluppo tecnologico. L'analisi del contenuto è illustrata a partire dai primi studi sulla propaganda di guerra effettuati dal politologo americano Harold Lasswell all'inizio del secolo scorso, fino alle più recenti applicazioni che vedono l'introduzione di tecniche di analisi automatica. Il capitolo illustra i software utilizzati nei capitoli successivi dai vari autori e spiega le logiche applicative, rispondendo alla domanda: quale tipo di conoscenza è possibile ottenere attraverso ricerche che utilizzano metodi di analisi statistica dei dati testuali e tecniche di text mining? La risposta non può che essere: dipende dalla domanda. E qui Sbalchiero sottolinea l'utilità di connettere i molteplici sviluppi dell'analisi statistica dei testi e del text mining alla crescente accessibilità a grandi collezioni di materiale empirico digitale e alla policy analysis. Tutte le tecniche presentate nel capitolo – *topic detection, Latend Dirichelet Allocation* (LDA), analisi delle distanze e delle corrispondenze – rispondono a domande di policy che vanno costruite a partire dall'esperienza sul campo e dall'addestramento dei decisori a porsi domande di ricerca funzionali al miglioramento di policy.

Nel capitolo due Salvatore Pinello analizza la giurisprudenza in materia di tutela dei consumatori a supporto dell'innovazione delle politiche della Regione Veneto. L'obiettivo di questa applicazione è l'indagine sull'allineamento tra le politiche formative a tutela dei consumatori e i principali temi alla base delle pronunce giurisprudenziali, allo scopo di favorire una convergenza tra le politiche regionali e i principali temi e problemi che emergono dalla giurisprudenza.

Nel capitolo tre Irene Diaz Mina applica l'analisi testuale a una porzione della banca dati dell'Archivio Diaristico Nazionale degli immigrati in Italia e collocato presso Pieve Santo Stefano, in Regione Toscana. Lo scopo è quello di elaborare i contenuti di tale archivio in modo automatico per estrarre conoscenza utile dalle esperienze e dai racconti degli immigrati. Tale conoscenza mette a fuoco dimensioni rilevanti dei percorsi di vita e integrazione degli immigrati sul territorio italiano, e consente di riflettere su politiche e strumenti che possono migliorare i processi di accoglienza e integrazione dal punto di vista di chi ne è diretto beneficiario.

Il capitolo quattro costituisce uno spaccato delle attività di analisi testuale che Luca Cipriani coordina all'interno della Regione Toscana da qualche anno. La frequenza del Master ha permesso di valorizzare e

mettere a sistema alcune esperienze sporadiche e applicazioni di analisi testuale automatizzata in Regione Toscana e di proiettarle in una dimensione di policy dell'ente. Questo tipo di tecniche garantiscono una velocità e una sistematicità delle operazioni di ricerca, spoglio e sintesi delle informazioni, offrono a monte utili spunti per approfondire e orientare le analisi e a valle la possibilità di verificare e applicare su larga scala le intuizioni che possono emergere da un primo esame diretto di un insieme di documenti normalmente più ristretto. Cipriani presenta alcune sperimentazioni realizzate nell'ambito di due distinte collaborazioni che Regione Toscana ha attivato rispettivamente con il Team per la Trasformazione Digitale e con il Dipartimento di Scienze Politiche, Giuridiche e Internazionali (SPGI) dell'Università di Padova, allo scopo di applicare le tecniche e acquisire conoscenza, misurare i risultati sul campo e valutare la loro adozione in futuro, sia attraverso sviluppi condotti con personale interno, sia acquisendo soluzioni in riuso da altre amministrazioni oppure disponibili sul mercato. Il capitolo illustra molteplici applicazioni, dalla valorizzazione dei contenuti della programmazione energetica, all'applicazione della *topic detection*, dall'analisi delle determine dirigenziali secondo il framework di Agenda 2030, all'analisi *sentiment* dei social.

I capitoli cinque e sei sono il frutto della sperimentazione effettuata da due archivisti dell'Archivio di Stato di Venezia. Essi applicano il text mining a due corpora distinti collocati presso l'Archivio: i verbali della Commissione araldica Veneta, ad opera di Salvatore Alongi, e il Codice diplomatico Veneziano, ad opera di Andrea Erboso. Entrambi i contributi utilizzano la *topic detection* e si collocano nel solco della valorizzazione e dell'accrescimento della fruibilità del patrimonio archivistico storico attraverso l'elaborazione automatica dei contenuti delle fonti. I due contributi evidenziano aspetti complementari ma differenti.

Il capitolo di Alongi, che analizza il contenuto dei verbali delle riunioni della Commissione araldica veneta svolte nei due quadrienni 1889-1893 e 1938-1942, intende offrire agli storici uno strumento di orientamento e di studio dei temi, problemi e campi di interesse toccati dall'attività della Commissione. Infatti, l'analisi testuale ha consentito di verificare la convergenza dei risultati dell'analisi automatica con le ipotesi avanzate in sede storiografica sulla base della lettura delle norme generali che nel tempo hanno disciplinato il funzionamento delle diverse commissioni araldiche regionali. Inoltre, l'analisi automatica fornisce all'Archivio di Stato la possibilità di coniugare in maniera innovativa le funzioni di conservazione, fruizione e valorizzazione del patrimonio documentario con-

servato. Il progetto presentato da Alongi rientra in una più ampia politica culturale, avviata dall'Istituto fin dal 2003, di recupero e valorizzazione delle fonti per la storia biografico-famigliare e araldico-genealogica.

Il capitolo di Erboso, in linea con i contenuti del Codice dei beni culturali e del paesaggio, prende le mosse da una delle principali sfide nella gestione del patrimonio culturale italiano: cercare un punto di equilibrio tra le esigenze collegate alla conservazione dei beni culturali e, dall'altro, le esigenze legate alla valorizzazione del bene, cioè alla sua fruibilità da parte dei cittadini e all'estrinsecazione del suo potenziale culturale. Il servizio di consultazione rappresenta infatti il fulcro delle attività dell'Istituto, ma anche il momento più delicato per la fragile documentazione archivistica. L'amministrazione archivistica italiana ha individuato da tempo, nel processo sistematico di digitalizzazione, la via maestra per garantire la conservazione del patrimonio e la sua accessibilità, un indirizzo di policy sfociato nel 2017 nel Piano nazionale di digitalizzazione del patrimonio culturale. Il capitolo ambisce a dimostrare come le tecniche di analisi testuale nella gestione archivistica possano migliorare l'offerta di servizio degli Archivi, offrendo nuove modalità di studio e di approccio al patrimonio archivistico. L'analisi testuale automatica conferma la validità dello strumento nel fornire originali chiavi di ricerca e nel facilitare la fruizione dei contenuti mediante la produzione di output specifici ed originali. Tali output supportano il lavoro dell'archivista e del ricercatore, tutelando ad un tempo la fragilità delle fonti e garantendo una migliore conservazione dei documenti.

Nel capitolo sette Monica Ibba applica l'analisi testuale ai contenuti dei social network per valutare la performance organizzativa e la qualità dei servizi culturali digitali offerti dalle Gallerie degli Uffizi di Firenze durante la pandemia, nel periodo compreso tra il 10 marzo 2020 e il 30 giugno 2021. L'analisi mostra come la fruizione digitale abbia costituito non solo una valida strategia alternativa di fruizione culturale durante il difficile periodo di *lockdown* e gestione della crisi sanitaria, ma si possa configurare altresì come una strategia integrativa anche in periodi di 'normalità'. Ibba, prendendo le mosse dal D.lgs. 150/2009, presenta una strategia di analisi della performance attraverso la *digital customer satisfaction* degli utenti dei contenuti culturali digitali creati dalle Gallerie degli Uffizi – prevalentemente nella forma di foto, video o dirette *streaming* – nella loro pagina Facebook. La realizzazione di un'analisi di *customer satisfaction* ha lo scopo di valutare la performance della policy digitale

dell'istituto culturale delle Gallerie degli Uffizi relativamente alle percezioni degli utenti rispetto al servizio online ricevuto e all'invariabilità o mutevolezza nel tempo di tali percezioni. Il lavoro vuole in primo luogo mettere in luce il valore prodotto dall'esperienza digitale degli utenti del museo degli Uffizi. In secondo luogo, intende incentivare l'analisi dei contenuti di tale esperienza per migliorare accesso e fruizione dei servizi digitali tramite strategie innovative in grado di sfruttare l'immenso – ma scarsamente valorizzato – patrimonio di dati custoditi dai social network. Infine, l'implementazione di uno strumento permanente di d*igital customer satisfaction analysis* rivela potenzialità di diffusione a tutte le altre piattaforme online del museo. L'analisi automatica del contenuto dei commenti presentata nel capitolo ha infatti mostrato come, a costi bassi, sia possibile analizzare, a livelli variabili di dettaglio, i feedback degli utenti, raccogliendo evidenze circa le percezioni e l'efficacia delle politiche di digitalizzazione della fruizione del patrimonio culturale, ricavando da esse indicazioni per il miglioramento dell'efficacia delle policy.

Il capitolo otto di Antonio Aggio risponde a una serie di domande a cui le amministrazioni regionali si troveranno sempre più frequentemente a dover fornire una risposta: possiamo, attraverso l'uso del text mining, accelerare e rendere più efficaci le fasi istruttorie di procedure concorsuali che prevedono la presentazione di documenti testuali? È possibile affidarsi all'analisi automatica dei testi per conoscere e classificare i contenuti, ad esempio, di centinaia di progetti? La stessa valutazione dei progetti può essere supportata dall'analisi dei testi che può fornire una misura della salienza di alcune parole chiave nei progetti esaminati? Può l'ente finanziatore individuare (e scartare) i progetti 'fotocopia' e premiare quelli più originali? L'applicazione dell'analisi testuale presentata da Aggio è stata effettuata su un corpus di progetti presentati a valere su un bando della Formazione professionale della Regione del Veneto emanato nel 2018 e finanziati nell'ambito del Fondo Sociale Europeo. La condivisione dei risultati dell'analisi testuale dentro l'amministrazione regionale ha determinato, *in primis*, un "effetto sorpresa". La ricostruzione dei mondi lessicali effettuata dal software sui contenuti di migliaia di progetti ha rivelato come questi mondi lessicali fossero pienamente in linea con le parole chiave e le linee di intervento del bando preso in esame. L'analisi potrebbe pertanto estendersi ad altri bandi in altre materie. In secondo luogo, è emerso come la variabile tempo, ovvero la 'velocità' di elaborazione dei contenuti di migliaia di progetti – trentotto secondi – e la

rapida capacità di individuazione di parole chiave e mondi semantici, sia un elemento di grande interesse per capire come si collocano i contenuti dei progetti rispetto alle richieste del bando (quali i temi più frequenti e quelli più associati tra loro), la loro distribuzione territoriale e il loro grado di originalità. Poche ore di lavoro hanno permesso di condurre analisi approfondite sul corpus dei progetti presentati nel bando. In un'epoca in cui si moltiplicano i bandi e il numero di progetti da valutare ed esaminare, l'applicazione delle tecniche di text mining nell'ambito dell'attività amministrativa è sfidante.

Il capitolo nove di Stefano Acciaioli illustra un'applicazione del text mining alla valutazione della qualità delle decisioni istruttorie nell'ambito del Settore Antisismica della Regione Toscana. Acciaioli analizza i contenuti delle richieste di integrazione prodotte dai tecnici istruttori del Settore Sismica della Regione Toscana nel corso della fase istruttoria del procedimento di controllo dei progetti strutturali delle costruzioni edilizie che sono depositati presso il medesimo Settore e che per legge sono assoggettati all'attività di controllo, ricercando la discrezionalità tecnica degli Uffici decentrati e i suoi potenziali effetti sistemici. Nel settore antisismico, la capacità di intervento e di policy del governo regionale appare condizionata dalla qualità delle procedure istruttorie sui progetti edilizi. L'individuazione delle cause principali all'origine di divergenze/difformità istruttorie ritenute 'anomale' dovrebbe permettere di migliorare la qualità del processo e l'uniformità delle decisioni alla normativa tecnica, con la conseguente riduzione del contenzioso amministrativo e del malcontento nei cittadini verso l'operato degli uffici regionali. L'analisi automatica è stata dapprima effettuata testando i contenuti delle istruttorie su uno stesso caso compiute dai vari istruttori regionali. In seguito, l'analisi testuale è stata estesa a una banca dati (P.O.R.T.O.S) di istruttorie, ricomprendenti cinque anni di attività dell'ufficio (2015-2020). I risultati ottenuti hanno rimarcato la presenza di difformità nelle richieste istruttorie sia a livello del singolo istruttore, sia con riferimento al dato aggregato su base territoriale dell'ufficio di competenza e con riferimento alla specifica formazione delle competenze dei tecnici del settore. L'analisi automatica dei testi applicata ciclicamente permetterebbe di innescare un meccanismo di monitoraggio periodico dell'attività istruttoria del settore antisismica, anche alla luce della modificazione della normativa tecnica. Questo per consentire un maggiore allineamento delle decisioni e un maggiore coordinamento territoriale nel settore per accrescere l'efficacia dell'implementazione della normativa tecnica.

Nell'ultimo capitolo, Simone de Lellis propone l'elaborazione di una 'matrice di sostenibilità' per l'analisi e il confronto dei Bilanci regionali e locali (ma non solo), che attuano lo standard di classificazione internazionale di secondo livello adottato dall'Unione Europea, denominato *Classification Of the Functions Of Government* (COFOG). A partire dall'ipotesi di una sinergia tra gli standard di classificazione della spesa pubblica in Missioni/Programmi, e gli *SDGs* dell'Agenda 2030, oggetto del capitolo è la costruzione di una matrice di correlazione tra le modalità di distribuzione della spesa pubblica e gli Obiettivi di sviluppo sostenibile. La matrice della sostenibilità si propone come modello standard declinabile a tutti i livelli di amministrazione pubblica (locale, regionale, nazionale ed europeo), per supportare la programmazione, l'analisi e la verifica delle politiche e degli interventi, sulla base della distribuzione della spesa tra gli *SDGs* di Agenda 2030. Per costruire la matrice, sono stati preparati 17 corpora, uno per ciascun *SDGs* dell'Agenda 2030, contenenti la descrizione dei 17 Obiettivi, dei relativi 169 *Target* e dei 273 Indicatori associati. Successivamente, i medesimi testi sono stati utilizzati per l'individuazione di alcune parole o espressioni chiave. Infine, è stato preparato il corpus relativo alle categorie di spesa, contenente la descrizione delle 19 Missioni e dei corrispondenti Programmi e Gruppi COFOG. È stato quindi possibile individuare le correlazioni tra le categorie di spesa e gli *SDGs* dell'Agenda 2030. La 'matrice per la sostenibilità' vuole essere uno strumento per incrementare la base conoscitiva di un ente pubblico e valutare il proprio posizionamento rispetto agli *SDGs*. Dalla realizzazione della matrice è emerso come quasi tutte le Missioni di spesa siano correlate con gli *SDGs*. L'Obiettivo associato al maggior numero di Missioni è risultato essere l'11 'Rendere le città e gli insediamenti umani inclusivi, sicuri, duraturi e sostenibili', con ben 6 Missioni associate. Al contrario, la Missione associata al maggior numero di Obiettivi è la 9 'Sviluppo sostenibile e tutela del territorio e dell'ambiente', con ben 5 Obiettivi associati. Alla fine dell'anno 2021, la matrice per la sostenibilità illustrata nel capitolo è stata impiegata dalla Regione Toscana nell'ambito del Bilancio di previsione 2020-2022 e per l'analisi dei dati di spesa del Programma Attuativo Regionale del Fondo di Sviluppo e Coesione (PAR FSC) 2007-2013 e 2014-2020, analizzando come la spesa del Programma si sia distribuita nei due periodi di programmazione tra gli Obiettivi di Agenda 2030.

Quando i dati sono i testi. Approcci e procedure per l'analisi dei dati testuali

Stefano Sbalchiero[1]

Text as data, Text mining, Analisi statistica dei testi.

Introduzione

«Text as data» (Gilardi et al. 2018) è una formula, più che un semplice espediente linguistico, oggigiorno entrata nel linguaggio della ricerca sociale e rappresenta, utilizzando un'affermazione che potrebbe sembrare perfino troppo esplicita, solamente un punto di arrivo di un percorso la cui caratteristica intrinseca è certamente una irrefrenabile pluralità, sia rispetto ai metodi e le tecniche sia per quanto riguarda settori scientifico -disciplinari differenti. E questo per almeno tre ragioni fondamentali. La prima si riferisce all'ambito metodologico in quanto, nonostante siano presenti approcci ben consolidati nel tempo, non solo i contesti di ricerca risultano mutevoli, ma con loro pure le tecniche e gli strumenti che gli studiosi mettono in campo al fine di estrarre informazioni dai testi per restituirle sotto una nuova forma. La seconda ragione rimanda, invece, agli ambiti disciplinari: chi si dovrebbe occupare dei testi come dati? Quali sono le discipline interessate? La risposta non può che essere, a sua volta, provvisoria. Molte sono le discipline che storicamente si sono interessate

[1] Sociologo, Dipartimento di Filosofia, Sociologia, Pedagogia applicata (FISPPA), Università degli Studi di Padova.

all'analisi dei testi, come ad esempio la linguistica, l'informatica, la statistica, per nominarne solamente alcune, e molte altre sono oggi le discipline che nutrono un certo interesse per le possibilità offerte dall'analisi dei dati testuali. La terza questione pertiene invece all'oggetto stesso della discussione: i discorsi, i testi, le parole. Se da un lato quanto detto pocanzi permette di comprendere i vantaggi, riconosciuti da più parti, di un approccio interdisciplinare al variegato universo di testi e parole che quotidianamente permeano la nostra realtà sociale, dall'altro lato l'utilizzo del termine 'discorso' non è casuale e improprio: l'etimologia stessa rimanda a 'scorrere', con il senso figurato di muoversi da una cosa all'altra. I discorsi sono, quindi, comunicazioni in movimento che non solo caratterizzano le nostre attività quotidiane e sociali, ma partecipano a pieno titolo alla produzione e condivisione di senso e significato che gli attori sociali attribuiscono a pratiche, siano esse soggettive o collettive. *Les Mots et les Choses*, direbbe Foucault (1966), Le parole e le cose. Se a questo si aggiunge la gran mole di informazioni con le quali ognuno di noi si interfaccia ogni giorno, dai social media ai flussi informativi, dalla carta stampata al web, dai *post* ai siti istituzionali, l'effettiva quantità di dati prodotti, ma anche utilizzati, o utilizzabili, è abnorme. Nell'epoca dei *big data*, altra formula difficile da definire in modo univoco, ulteriore elemento di complessità deriva dal rapido e costante cambiamento che investe la natura stessa dei rapporti tra media, *new media*, processi di digitalizzazione e modalità di fruizione e gestione dei contenuti. Entro tali processi, la partecipazione degli utenti ha innescato, infatti, l'irreversibile modifica sia nella produzione, attraverso anche la digitalizzazione, sia nell'interpretazione di testi. Per fare soltanto un esempio, basti pensare agli esiti della cultura definita *grassroot*, generata dall'utente: una stessa idea può essere espressa in una forma differente, attraverso la molteplicità e la divergenza degli strumenti e dei supporti disponibili, pur mantenendo i medesimi contenuti (Jenkins 2007). Che dire, poi, dei processi di digitalizzazione e dell'importanza che rivestono oggigiorno in ambiti quali l'amministrazione digitale, o *e-government*, entro cui assume particolare rilevanza il fenomeno della data governance, com'è stata definita nell'introduzione. Al di là delle problematiche relative alla gestione digitalizzata della pubblica amministrazione, tra i fenomeni che più interessano in questa sede vi sono, senza ombra di dubbio, tutte quelle azioni che rientrano a pieno titolo entro tali processi di cambiamento organizzativo e che in misura differente richiedono di mettere in campo strumenti e competenze diversificate. Saper organizzare e trattare la documentazione resa disponi-

bile, al fine di estrarre informazioni e, così facendo, ottimizzare non solo il lavoro del funzionario e degli enti preposti, ma offrire supporto allo sviluppo di politiche e servizi. Come già anticipato, se i dati in possesso delle pubbliche amministrazioni possono essere suddivisi, seppure con uno sforzo classificatorio, tra dati numerici – ad esempio il numero di delibere, di piani strategici o di sentenze e così via – e dati testuali, allora questi ultimi possono essere considerati strategici nella misura in cui vi sia la possibilità di accedere ai contenuti per comprenderne le implicazioni rispetto allo sviluppo di politiche e di supporto alle decisioni. Si pensi all'analisi di programmi, piani e delibere che a vari livelli, locali o sovralocali, molto ci dicono rispetto alle traiettorie progettuali intraprese, oppure all'analisi della documentazione inerente alla valutazione o agli strumenti di *accountability* con cui gli enti preposti comunicano verso l'esterno la propria identità rispondendo alle domande 'chi sono', 'cosa hanno fatto', 'cosa si accingono a fare'.

Difronte a scenari di questo tipo, l'assunto incontrovertibile dal quale muovere per una considerazione che faccia eco alla valorizzazione del dato testuale, accanto a quello numerico, è il seguente: a fronte della complessità, disponibilità e accessibilità al rapido flusso di parole che scorrono, è crescente l'esigenza di riuscire a intercettare e comprendere dinamiche veicolate attraverso i testi tramite la sintesi e la restituzione di dati testuali, avendo la possibilità di trovare dei riscontri oggettivi. Senza ombra di dubbio, questa considerazione pone alcune questioni fondamentali sia rispetto agli approcci e ai frame teorici, sia per quanto riguarda gli strumenti da adottare. Da un punto di vista dello scienziato sociale, infatti, il tipo di approccio teorico potrebbe essere parzialmente differente da quello del linguista, così come l'interesse dello statistico nei confronti dei testi può prevedere l'utilizzo di strumenti diversi da quelli considerati maggiormente utili dal politologo. Tuttavia, questo ragionamento non deve essere fuorviante: va ribadito, in linea con quanto anticipato, che l'approccio interdisciplinare non è solo auspicabile, ma fondamentale nel panorama teorico dell'analisi dei testi. In questa ottica, l'analisi dei testi moderna e supportata da software e tecniche sempre più raffinate vede linguisti, *computer scientist*, statistici, psicologi, politologi e sociologi attivare proficue collaborazioni al fine di sviluppare nuovi strumenti e percorsi di indagine in grado di rispondere alle specifiche quanto diverse esigenze di ricerca. È persino inutile sottolineare che questo dipende fortemente dalle domande che vengono poste, vale a dire dagli obiettivi dell'indagine. In questo caso, i due concetti di affidabilità, vale a dire la

capacità di produrre misurazioni tra di loro coerenti, e quello di validità, cioè il grado in cui uno strumento misura effettivamente ciò che si vuole misurare, risultano centrali nella valutazione aprioristica delle scelte sia di approcci sia di strumenti. Se per esempio l'intento del ricercatore fosse quello di studiare l'evoluzione del linguaggio amministrativo in una prospettiva storica, non risponderebbe al criterio di validità l'applicazione ai testi di un algoritmo di *stemming*, che riduce alla sua forma radice la parola, dato che verrebbero perse preziose informazioni e peculiarità linguistiche. Dall'altro lato lo *stemming* potrebbe produrre risultati più consistenti e validi volendo analizzare i temi maggiormente affrontati dalle pubbliche amministrazioni, anche in ottica comparativa, attraverso documenti programmatici relativi all'economia circolare o al cambiamento climatico. Allo stesso modo, per lo studio delle rappresentazioni di determinate categorie di attori sociali nella documentazione prodotta da un ente locale, potrebbe essere fondamentale ricorrere a specifici strumenti di analisi linguistica, come il *tagging* grammaticale, per lo studio delle forme verbali ad esse associate. In questo caso, se l'utilizzo del verbo 'supportare' riferito a una categoria specifica di beneficiari risultasse sottoutilizzato rispetto a 'contrastare' evidentemente le conclusioni sarebbero differenti, così come 'riconoscere' e 'disapprovare' avrebbero un peso diverso in seno al dibattito riferito alle politiche di risparmio energetico. In quest'ottica, l'utilizzo del dato testuale, accanto al dato numerico, può costituire un terreno di incontro privilegiato tra differenti discipline e dar luogo ad ambiti di sperimentazione continua, in grado di stimolare gli strumenti che abbiamo a disposizione rispetto alle crescenti sollecitazioni provenienti dalla realtà sociale.

Al di là di questi esempi, molte sono le problematiche che caratterizzano la discussione sui metodi e le applicazioni per analizzare materiale testuale o estrarre informazioni dai testi, come si vedrà di seguito. Per il momento basterà dire che quando facciamo riferimento all'analisi dei dati testuali questa non può che costituirsi come interdisciplina o, detto in altri termini, più che rappresentare un'isola essa non può che configurarsi come una terra di mezzo tra saperi fortemente interrelati.

Nei paragrafi seguenti, a partire da questa riflessione, il tentativo sarà quello di collocare l'analisi dei testi, secondo diversi orientamenti, proponendo alcuni indirizzi di indagine che interessano le scienze umane e sociali e che hanno trovato una diretta applicazione nei singoli lavori che verranno presentati a seguire.

Contesti e dibattiti: Text Mining, *Digital Methods* e *Big Data*

Quello che oggi conosciamo con il nome di text mining è il risultato, senza ombra di dubbio, della crescente bibliografia e del moltiplicarsi di studi soprattutto in ambito linguistico che hanno incrociato l'informatica e lo sviluppo tecnologico (Bolasco 2005, Sbalchiero 2018). Fortemente connotata dalla collaborazione tra ambiti disciplinari differenti, tale espressione può essere considerata come un settore entro il quale la linguistica computazionale, la matematica e la statistica, così come le scienze sociali, informatiche e *computer science* hanno creato un terreno fertile alle collaborazioni e agli sviluppi di tecniche e metodi. Non può tra l'altro sfuggire l'assonanza tra text mining e *data mining*: si tratta, infatti, di quell'insieme di metodi e tecniche il cui fine è la produzione di conoscenza attraverso l'estrazione di informazioni da grandi quantità di dati tramite metodi automatici. Gli esiti delle ricerche sono sotto gli occhi di tutti e hanno trovato espressione sia nella ricerca di base sia nella ricerca applicata e industriale: il riconoscimento vocale, la traduzione automatica, i sistemi informativi così come la gestione della conoscenza per mezzo delle tecnologie dell'informazione da parte di imprese, il così detto *knowledge management* (Giuliano et al. 2008). Ed è proprio in questo contesto che le tecniche di text mining si sono moltiplicate nel tentativo, in primo luogo, di affrontare i problemi connessi all'analisi e alla gestione di elevate quantità di testi liberi, ovvero non strutturati, con l'obiettivo di estrarre e ricavare da queste collezioni di documenti, in secondo luogo, informazioni utili per organizzare e alimentare database di grandi dimensioni. Il successo di queste tecniche, oltre all'ambito meramente industriale e aziendale, è dovuto alla possibilità di applicarle a qualsiasi tipo di testo non strutturato, ovvero con una struttura che non può essere identificata, come ad esempio le pagine web, agenzie stampa, archivi online e così via, stimolando l'interesse di molte discipline (Sanger et al. 2007). Fra le tappe che hanno segnato lo sviluppo di tale ambito, possiamo annoverare gli studi pionieristici di tipo lessico-testuale (Luhn 1959) e focalizzati sull'indicizzazione. Da questo punto di vista, l'*information retrieval* e *l'information extraction*, rispettivamente il recupero di documenti tramite *query* di ricerca e l'assegnazione di categorie ai documenti per facilitarne la consultazione, hanno non solo permesso di gestire grandi corpora testuali, ma gettato le basi per lo sviluppo di tecniche sempre più sofisticate (Bolasco 2005). L'attenzione è stata anche posta maggiormente sulla possibilità di una vera e propria descrizione dei testi attraverso il linguaggio matematico-statistico, o ancora attraverso innovative strategie per

recuperare le informazioni attraverso, ad esempio, il *clustering* gerarchico automatico dei documenti (Jardine et al. 1971). Non solo, ma attraverso l'utilizzo di dizionari elettronici e lessici di frequenza, ricorrendo quindi anche a meta-informazioni linguistiche, alcuni studi pionieristici diedero vita alla linguistica computazionale (Busa 1974-1980), espressione che ha poi portato a riflessioni sulla disciplina e a coniare il termine informatica umanistica (Orlandi et al. 2003), utilizzato nella lingua italiana anche come traduzione di *digital humanities*:

> «l'informatica umanistica costituisce il punto di contatto tra scienze umane e scienze esatte: ragionando sui caratteri comuni delle diverse discipline umanistiche e formalizzando le procedure necessarie per condurre la ricerca nei diversi ambiti, propone l'integrazione dei due mondi superando la semplice applicazione di tecnologie avanzate a settori delle scienze» (Celentano et al. 2004, p. 44).

Ed è proprio a partire dai numerosi studi che si sono succeduti a partire dagli anni '70 e '80, assieme agli sviluppi dell'informatica, che si assiste al moltiplicarsi di tentativi e allo sviluppo di algoritmi nell'ambito, soprattutto, dell'intelligenza artificiale e del *machine learning* applicati ai dati testuali e alla ricerca di tipo linguistico (Porter 1980, Berger et al. 1996). Il resto è storia recente, e il passo dal *data mining* al text mining è risultato breve. In sostanza, non può certamente sorprendere che l'interrogativo rispetto a che cosa si intenda oggi con *digital humanities* ci dice molto rispetto all'evoluzione di una disciplina piuttosto recente, il cui sviluppo è stato accompagnato dalla rapida moltiplicazione di gruppi di ricerca, conferenze, pubblicazioni, riviste e così via, ma che contemporaneamente necessita ancora di una definizione largamente condivisa (Gold 2012). Questo per dire che la sua definizione, in modo simile ad altre discipline, non può rappresentare uno stadio dello sviluppo e maturazione della disciplina stessa (Arthur et al. 2014). Infatti, all'interno di tale dibattito, è emersa anche, sotto forma di critica, la necessità di una riflessione maggiormente profonda capace di cogliere i motivi per cui, a fronte della sempre maggior diffusione e utilizzo di strumenti digitali e risorse di text mining, la difficoltà di collegamento tra l'umanistica digitale e ricerca *mainstream* in ambito umanistico risulta tutt'oggi evidente. In questo senso, Patrik Juola (2008) parla di un vero e proprio abbandono, percepito dai ricercatori che si occupano di *digital humanities*, da parte della più generale comunità umanistica. Questo sarebbe dovuto, da un lato, alla constatazione che tale comunità nel suo complesso non è sempre a conoscenza degli strumenti sviluppati dai professionisti in metodi digitali e,

dall'altro lato, perché il *mainstream* umanista tenderebbe a non prendere sul serio molti dei risultati delle ricerche prodotte e ottenute con metodi e strumenti propri delle *digital humanities* (*ivi*, p.73). Senza voler entrare nei dettagli di tale dibattito, ma volendo fornire una possibile interpretazione, si potrebbe anche aggiungere, in continuità con quanto detto, che il problema della definizione di tale ambito di ricerca è assimilabile al problema della demarcazione dei confini tra le discipline scientifiche (Gieryn 1983, 1995). Una demarcazione non tanto verticale, vale a dire tra ciò che può essere definito scientifico o meno, quanto invece orizzontale tra ambiti disciplinari differenti e/o tra specializzazioni entro la stessa disciplina. Dovrebbe risultare chiaro, seguendo questo ragionamento, che nel caso delle *digital humanities* il paradosso tra la crescente diffusione di questo termine ombrello, il suo inconfutabile successo nel panorama odierno, fino alla denuncia di mancato riconoscimento, potrebbe in realtà essere visto come un punto di forza, trattandosi di un orientamento fortemente interdisciplinare, capace di generare nuove prospettive di analisi anche senza catturare l'attenzione della ricerca *mainstream* (Prescott 2012). Questo perché, nella ricerca sociale odierna, i ricercatori devono confrontarsi con quesiti epistemologici e metodologici da un lato rispetto ai fenomeni, a volte inediti, che necessitano ancora di essere definiti e, dall'altro lato, che riguardano i presupposti stessi che stanno alla base dello sviluppo di nuove pratiche di ricerca attraverso strumenti a loro volta nuovi o innovativi. A questo proposito, un'ulteriore nota a margine rispetto a quanto accennato può essere utile per collocare il text mining entro i complessi scenari della ricerca odierna. Il tema *Big Data*, che secondo il modello delle tre V (Laney 2001), sono caratterizzati da alta varietà (fortemente eterogenei), velocità (acquisiti ed elaborati in tempo reale) e volume (ordini di grandezza in termini di misurazione), necessita di un constante confronto interdisciplinare dal momento che i database raccolgono tipi di dati che possono essere strutturati o non strutturati e quindi offrire nuovi modi di fare scienza sociale (Roberts et al. 2016). In effetti, quando si parla di text mining tale concetto si accompagna di sovente a quello di *big data* (Delmastro et al. 2019). Questo perché, come già anticipato, le procedure di text mining riguardano principalmente l'analisi automatizzata di grandi quantità di testo, espressione questa che viene associata al concetto di *big data*, a volte in modo fuorviante: con esso ci si riferisce a vaste raccolte di dati che per le loro caratteristiche e dimensioni non possono essere archiviate e analizzate con strumenti convenzionali e tecniche informatiche standard. I *big data*, dunque, possono

certo fare riferimento a vari tipi di dati, che comprendono database di grandi dimensioni, ma anche *file* audio e video, così come grandi quantità di testo, per lo più non strutturato. Ma tale espressione può essere correttamente utilizzata quando l'insieme di questi dati e la loro definizione, siano essi testuali o di altra natura, supera una soglia che non solo non viene definita in modo unanime in letteratura, ma soprattutto risulta talmente complessa e di dimensioni tali da richiedere nuove tecniche e strumenti per estrarre, elaborare, gestire e restituire le informazioni in una nuova forma (De Mauro 2019). Tale constatazione porta sicuramente a una considerazione di questo tipo: nella ricerca sociale e nell'ambito della ricerca umanistica, ciò significa mettere a punto strategie di ricerca, nel nostro caso applicate ai testi, facendo ricorso a procedure e software in grado di estrarre informazioni e renderle disponibili all'interpretazione con approcci adeguati rispetto alla dimensione del materiale empirico oggetto di indagine. Di più, possiamo affermare che il nodo cruciale è relativo alla capacità dei ricercatori di promuovere una costante unione tra scienze umane e tecnologiche, e tra le discipline umanistiche stesse, e che il successo di tali approcci negli anni a venire dipenderà notevolmente sia dalla diffusione della pratica scientifica in esse svolta, sia nella capacità di rappresentare e interpretare la realtà sociale, vale a dire selezionare ed includere metodi, orientamenti e caratteristiche conoscitive utili (in quel momento) per raggiungere dei fini pragmatici, legittimando così l'autorità culturale e le pretese di conoscenza in esse rivendicate.

Approcci, metodi e tecniche

Tra i numerosi approcci per l'analisi dei testi possono essere distinti diversi orientamenti, i quali, per caratteristiche tecniche e per obiettivi conoscitivi, risultano utili ai fini della presente ricostruzione. Un primo orientamento è, a livello generale, collocabile entro la tradizione dell'analisi del contenuto nella versione moderna, l'analisi automatica del contenuto, supportata da software, in cui uno degli obiettivi è l'identificazione degli argomenti principali presenti in una collezione di testi. Altri approcci, invece, non si pongono tanto l'obiettivo di fare analisi del contenuto, come viene tradizionalmente intesa, quanto invece di analizzare i testi secondo procedure e tecniche utili, tra le altre cose, alla classificazione dei testi e all'analisi di documenti a seconda della 'distanza' esistente tra di loro. Tra tutte le possibilità offerte dall'analisi testuale e dal text mining, gli approcci e gli strumenti presentati, come si vedrà, sono risultati particolarmente utili per esplorare grandi quantità di documenti, estrarre le

informazioni principali, analizzare i principali contenuti o classificare in modo automatico i testi, anche in assenza di un sistema di classificazione preesistente.

Prima di addentraci nei principali orientamenti e dettagli tecnici relativi ai software e alle procedure, può risultare utile chiarire alcuni termini che verranno utilizzati in seguito. Quando parliamo di *corpora* facciamo riferimento a una collezione di testi:

> «il materiale testuale oggetto delle analisi prende il nome di corpus e si configura come una collezione di testi. Il corpus raccoglie testi coerenti con gli scopi perseguiti dalla ricerca e questa coerenza è valutabile solo discrezionalmente. Nello studio dell'intera opera di un autore i testi costituenti il corpus possono essere, per esempio, le singole opere inedite e/o inedite di cui si conosce l'esistenza; nello studio di un romanzo i singoli capitoli; nell'analisi dei risultati di un'indagine con intervista a domande aperte le trascrizioni dei colloqui […] nell'analisi di annate di stampa i quotidiani (o i settimanali o mensili ecc.) pubblicati» (Tuzzi 2003, p. 29).

Il *vocabolario*, uno dei principali strumenti prodotti attraverso i software che vedremo, si presenta come una lista ordinata, generalmente per frequenza decrescente, di *parole*, ovvero *forme grafiche*, alle quali è associata la frequenza con cui si presentano nel corpus. Una forma grafica altro non è che una sequenza di caratteri appartenenti all'alfabeto, delimitata da due separatori, come lo spazio (il *blank*) e i segni di interpunzione. La forma grafica, dunque, viene utilizzata per identificare quella che nel linguaggio comune viene definita parola. A sua volta la forma grafica può essere definita in due differenti modi: come *word type*, quando facciamo riferimento alla lista di parole diverse presenti nel vocabolario; come *word token*, invece, quando intendiamo le occorrenze presenti nel vocabolario.

L'analisi automatica del contenuto

Il primo approccio cui possiamo fare riferimento è la *topic detection* nella versione che possiamo annoverare nel panorama degli strumenti e dei metodi automatici per l'analisi del contenuto (Sbalchiero 2018). Anche se l'analisi del contenuto, da un punto di vista storico, non è certamente recente (Losito 1993, Tuzzi 2003), basti pensare ai lavori di Lasswell sulla *Propaganda Technique in the World War* (Lasswell 1927) e sull'analisi quantitativa del linguaggio politico (Lasswell 1949), o al lavoro di Thomas e Znaniecki riguardante l'analisi delle lettere nel celebre *Il contadino po-*

lacco in Europa e in America (Thomas et al. 1918-1920), per citarne soltanto alcuni, il percorso di affinamento dell'approccio ha conosciuto fasi alterne (Krippendorff 1983). Interessante, da questo punto di vista, sono le numerose considerazioni e le riflessioni metodologiche che si sono susseguite negli anni rispetto ai modi di fare ricerca, ben interpretate da Sorokin (1956), che utilizzò il termine 'quantofrenia' facendo emergere un problema rilevante anche dal punto di vista del metodo: l'eccessiva rigidità imposta da un percorso di ricerca che sotto l'egida dell'oggettività scientifica finiva per trascurare, se non perdere totalmente, la ricchezza qualitativa della ricerca. Accanto a questo, lo sviluppo tecnologico, dei computer e dei *software* portò ad accrescere l'interesse verso l'analisi dei contenuti: se da un lato si potevano gestire molti più dati, è soprattutto nel campo degli studi linguistici che emersero approcci non soltanto orientati al mero aspetto quantitativo, e quindi orientamenti di tipo lessicale (Tuzzi 2003). In particolare, in seno alla scuola francese (Beaudouin 2016), e grazie ai lavori e alle innovazioni metodologiche introdotte da Jean P. Benzécri (1982), lo sviluppo dell'approccio lessico-testuale superava da un lato i limiti dell'analisi delle sole frequenze e si orientava, dall'altro lato, all'analisi delle relazioni tra più variabili adottando una prospettiva multidimensionale, come nell'analisi delle corrispondenze lessicali (Benzécri 1992). Tutte queste premesse ci conducono al nocciolo della questione: la necessità di gestire grandi quantità di testi, attraverso analisi del contenuto maggiormente complesse, e quindi la necessità di una sintesi tra aspetti quantitativi e approfondimenti qualitativi al fine dello sfruttamento statistico dei dati testuali (Lebart et al. 1988, Bolasco 1999).

Ed è in questa cornice che si situa l'algoritmo utilizzato per i lavori che seguiranno. Si tratta del così detto metodo Reinert (1983, 1990), implementato nel software Alceste *(Analyse Lexicale par Contexte d'un Ensemble de Segments de Texte)* e, più di recente, disponibile nella versione *R-based* di Iramuteq *(Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires),* (Ratinaud 2014).

Per quanto riguarda i fini del presente contributo, è utile sottolineare che il metodo Reinert risulta di particolare interesse considerando l'analisi del contenuto che vede la collaborazione tra linguisti, *computer scientist*, statistici, psicologi, politologi e sociologi nello sviluppo e utilizzo di nuovi strumenti e percorsi di indagine in grado di rispondere alle specifiche quanto diverse esigenze di ricerca. In particolare, i concetti di affidabilità (la capacità di produrre misurazioni tra di loro coerenti e indipendenti dal ricercatore), e quello di validità (la capacità di uno stru-

mento di misurare effettivamente ciò che ci si propone di misurare), sono risultati centrali nel tentativo di coniugare, nello stesso strumento, rigore scientifico e approfondimenti qualitativi nell'analisi automatica dei contenuti. Detto in modo differente, l'algoritmo ideato da Reinert si pone l'obiettivo di integrare quantità e qualità attraverso un percorso di ricerca valido ed efficacie, facendo ricorso, da un lato, all'approccio moderno lessico-testuale, ma, dall'altro lato, perseguendo fini e obiettivi conoscitivi che valorizzano anche l'aspetto qualitativo.

A livello operativo, la procedura consta di diverse fasi successive così come vengono implementate dal software Iramuteq (Ratinaud 2014, Sbalchiero 2018). Il corpus testuale, composto da diversi testi, viene ridotto in porzioni di testo che coincidono con una frase, un enunciato o un paragrafo delimitato da punteggiatura, e che prendono il nome di *Elementary Context Units* (ECU). In questo modo, il corpus viene organizzato in modo automatico dal software in porzioni di testo di lunghezza simile. Ad un passo successivo, l'algoritmo identifica le co-occorrenze delle parole in ogni ECU attraverso la costruzione di una matrice di contingenza *parole x ECU* (Tab. 1)

Tab. 1 Esempio di una tabella *(parole x ECU)* e relativo dendrogramma.

	Parola 1	*Parola 2*	*Parola 3*	*…*	*Parola N*				
ECU 2	1	0	1	0	0	classe 1			
ECU 4	1	0	1	0	1				
ECU 1	0	1	0	1	1	classe 2			
ECU 3	1	1	0	1	1				
ECU 5	0	1	0	1	1	classe 3			
ECU 6	0	1	0	1	1				
ECU 7	0	1	0	1	1				

Tale matrice, organizzata come tabella presenza *vs* assenza, dove "0" corrisponde all'assenza mentre "1" alla presenza della parola nella porzione di testo, viene costruita come base per analizzare la similarità delle ECU, che viene identificata tramite una procedura di *clustering*. Tale procedura rileva, in modo gerarchico e attraverso la distanza del χ^2 tra le classi, quelli che vengono definiti mondi lessicali (Reinert 1983) o classi semantiche (Ratinaud et al. 2012, 2015, Smyrnaios et al. 2017). Il dendrogramma, come risultato principale, viene costruito a passi successivi: all'inizio dell'analisi vengono create due classi, ognuna delle quali raggruppa le ECU che rispecchiano un contenuto lessicale simile, che con-

dividono cioè parole e che, di converso, si differenziano maggiormente l'una dall'altra, cercando in questo modo di ridurre al minimo le parole in comune. Si procede quindi con ulteriori ripartizioni delle ECU per classi fino a quando risultano sufficientemente omogenee da non poter essere ulteriormente disaggregate. I risultati ottenuti sono interessanti per diverse ragioni. Da un lato, la classe semantica (Reinert 1993), interpretabile come una variabile latente (Reinert 1998, pp. 292-293), è caratterizzata da uno specifico vocabolario di parole tra di loro associate, che co-occorrono nelle stesse porzioni di testo e che, dall'altro lato, rendono particolarmente efficacie e intuitivo il processo di interpretazione delle classi, ovvero dei principali argomenti – *topics* – presenti nel corpus analizzato. Nella fase di interpretazione, inoltre, per facilitare il compito dei ricercatori, tramite l'utilizzo dei valori di associazione del χ^2 delle parole con la classe, il software permette di estrarre le porzioni di testo (ECU) maggiormente significative per dar conto di quella classe semantica. In questo senso, l'algoritmo riporta, per ognuna delle porzioni di testo classificate dal metodo, un valore che corrisponde alla media dei valori del χ^2 delle parole significative e che sono associate a quella classe. Di conseguenza, la significatività di una porzione di testo per dar conto di uno specifico argomento è tanto maggiore quanto più contiene un numero di parole che risultano significative per quella classe e quindi in grado di rappresentarla in termini di contenuto. Infatti, l'identificazione dei principali argomenti e quindi dei contenuti è l'obiettivo principale dell'algoritmo, che identificando i mondi lessicali partecipa all'individuazione e alla costruzione di vocabolari di parole co-occorrenti che li caratterizzano. Detto in modo differente, l'esito della classificazione delle porzioni di testo (ECU) in un insieme di classi che includono parole rilevanti per la classe indica che tanto più le ECU sono simili, quanto più condividono parole. Va altresì specificato che l'elenco delle parole più significative e che meglio rappresentano un mondo lessicale – il vocabolario dei *topics* – viene identificato tramite l'associazione del χ^2 tra le parole e le classi. In questo senso, la soglia di significatività del χ^2 viene identificata tramite il *p-value*: le parole che presentano un *p-value* < 0.0001 saranno quindi considerate più significative rispetto a quelle che hanno un valore superiore alla soglia *p-value* ≥ 0,05.

Come si può notare nell'esempio (Tab. 2), la procedura di *clustering* evidenzia gruppi tematici omogenei e che presentano contenuti simili interpretabili osservando le forme significativamente associate ad ogni classe.

Tab. 2 Esempio di vocabolari delle classi semantiche. Le parole vengono ordinate per valore di associazione decrescente (χ^2) e con una significatività del *p-value* < 0.0001.

Classe 1	χ^2	**Classe 2**	χ^2	**Classe 3**	χ^2
ambiente	253,609	energia	206,603	Mobilità	113,497
territorio	134,013	sostenibilità	149,65	Urbana	105,728
Turismo	117,431	rinnovabili	124,805	Impatto	93,7
promozione	96,304	risorse	114,543	Sociale	75,603
:	:	:	:	:	:
politiche	34,247	alternative	46,763	Scelte	45,272
ambientali	30,867	futuro	41,353	Policy	42,827

Infine, i risultati dell'analisi del contenuto prodotti dall'algoritmo possono essere utilizzati per valutare il grado di associazione delle classi semantiche con le modalità di altre variabili per rispondere, ad esempio, alla domanda 'chi dice che cosa' (incrociando i *topics* ottenuti con le testate giornalistiche, in caso di analisi di articoli, oppure con i paesi, in caso di analisi comparativa di documenti istituzionali), oppure ancora con la variabile anno, qualora sia necessario verificare la presenza di determinati argomenti in una prospettiva longitudinale. Anche in questo caso, le relazioni delle classi semantiche individuate con le altre variabili sono basate sui valori di associazione del χ^2 che misurano in che modo tali variabili sono associate con le classi semantiche individuate.

Infine, tra le varie opportunità offerte dal software, vi è anche l'analisi delle corrispondenze, uno strumento classico per l'analisi dei dati testuali, utilizzata in diversi lavori che seguiranno. L'analisi delle corrispondenze (Greenacre 2007, Lebart et al. 1984, Tuzzi 2018) è particolare caso di *Principal Component Analysis* (PCA) applicata a una tabella di contingenza che incrocia le parole (righe) per le modalità della variabile (nelle colonne). Tale tecnica ha l'obiettivo di trasformare le frequenze delle parole in coordinate su un sistema di assi cartesiani multidimensionali: in questo modo, è possibile visualizzare modalità di variabili e parole traducendo il concetto di similarità in una specifica distanza euclidea (Murtagh 2005) e quindi in piani cartesiani (Sbalchiero et al. 2016). L'analisi delle corrispondenze applicata ai testi, dunque, si basa sui profili lessicali e mira a mettere in evidenza le relazioni tra modalità delle variabili (come, ad esempio, gli anni, oppure i documenti prodotti da diverse istituzioni e così via) e le parole. Le posizioni delle modalità delle variabili sui piani sono quindi determinate dal grado di somiglianza dei profili lessicali: ne consegue che quelle presenti nello stesso quadrante del piano rispecchia-

no contenuti simili in termini di parole ed è quindi possibile interpretare le posizioni reciproche (modalità delle variabili e parole) e ricostruire i principali contenuti entro una mappa.

Text mining: *topic detection*

Se l'estrazione dei *topics* con il metodo Reinert rientra propriamente nel panorama dei metodi per l'analisi automatica dei contenuti, strumento molto utile, come si è visto, al fine dell'analisi statistica dei testi, molti sono i recenti progressi nella tecnologia hardware e software che hanno incentivato lo sviluppo e la diffusione di algoritmi di text mining (Aggarwal et al. 2012). In particolare, ai fini del presente contributo, tra i numerosi approcci e algoritmi di *topic modelling* (Gilardi et al. 2021), sviluppati nell'ambito dell'informatica e del *Machine Learning* e che godono di un grande successo nelle applicazioni di text mining, troviamo la *Latent Dirichelet Allocation* (LDA), utilizzata anche in alcuni lavori a seguire. Il successo di modelli probabilistici è basato sul fatto che non solo possono essere potenzialmente applicati a qualsiasi collezione di testi, anche di vasta dimensione, ma che possono estrarre argomenti latenti secondo una logica diversa da quella che abbiamo descritto pocanzi. Nello specifico, l'LDA è un modello generativo probabilistico, presentato in uno studio pubblicato da David Blei e colleghi (2003), a cui sono seguite altre varianti basate su di esso, come il *dynamic topic model* (Blei et al. 2006), il *correlated topic models* (Blei et al. 2007) o la *structural topic model* (Roberts et al. 2014). Tali sviluppi rappresentano modelli adattati a esigenze specifiche, come ad esempio la necessità di mettere in relazione i *topics* rilevati con altre variabili, ma in ogni caso l'LDA rimane una pietra miliare nel contesto dei *topic models*. L'assunto dell'LDA è che ogni testo che compone un corpus può essere rappresentato da un insieme di *topics* latenti, onde per cui un documento viene considerato come una combinazione probabilistica di questi argomenti, ognuno dei quali è caratterizzato da una specifica distribuzione di parole (Griffiths et al. 2004). Si tratta, in questo caso, di un modello generativo in quanto, non potendo osservare direttamente i *topics*, vengono utilizzati i dati a disposizione, ovvero le parole, per ricostruire, a posteriori, la struttura latente del corpus analizzando i testi e le parole che lo compongono. In altre parole, questo processo, probabilistico e generativo, evidenzia l'interazione tra i documenti osservati e i *topics* latenti al fine di analizzare la probabilità che un documento contenga informazioni su un argomento sulla base della distribuzione delle parole contenute nel testo. Se, quindi, ogni testo è costituito da un numero T di

topics (j = 1..T), che a loro volta sono caratterizzati da parole specifiche, ognuno di questi *topics* può essere rappresentato come una distribuzione di probabilità sul vocabolario. Di conseguenza,

> «se abbiamo T topics, possiamo calcolare la probabilità della parola i-esima in un dato documento come:
>
> $$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$
>
> dove z_i è la variabile latente che indica il topic da cui è stata tratta la i-esima parola e $P(w_i|z_i=j)$ è la probabilità della parola rispetto al j-esimo topic. $P(z_i=j)$ fornisce la probabilità di scegliere una parola dai topics j nel documento, che varierà tra i diversi documenti. Intuitivamente, $P(w|z)$ indica quali parole sono importanti per un topic, mentre $P(z)$ è la prevalenza di quei topics all'interno di un documento» (Griffiths et al. 2004, p. 5228, trad. nostra).

Possiamo dire che i risultati principali ottenuti con questo approccio sono da un lato liste di parole associate ai *topics* e, dall'altro lato, quelli che risultano essere i documenti più significativi per i *topics* rilevati. L'LDA può essere applicata a un corpus di testi attraverso l'utilizzo del '*topicmodels' package* (Grün et al. 2011) disponibile in R (R development core team 2016), che permette di implementare l'algoritmo proposto da Blei in ambiente *open source*. Per ulteriori approfondimenti si rimanda alla bibliografia (Blei et al. 2009), anche se un ulteriore aspetto vale la pena di essere sottolineato. Come si è visto, trattandosi di un modello generativo, l'LDA ricostruisce (ovvero genera) i documenti del corpus assegnando il peso probabilistico di ogni *topic* ai documenti e, per ogni *topic*, la distribuzione delle parole, con la possibilità di evidenziare quelle con il livello di probabilità più elevato, vale a dire maggiormente rilevanti, per lo stesso *topic*. Ora, rispetto all'applicazione dell'algoritmo, è stato dimostrato che la *topic detection* funziona abbastanza bene con testi brevi, come ad esempio *abstract* di articoli, *post* sui *social media*, articoli di giornale, articoli scientifici, articoli di Wikipedia, per citarne alcuni, e il motivo è che intuitivamente forniscono informazioni concise sui contenuti principali. Cosa accade quando applichiamo la *topic detection* a testi lunghi? Di fatto, quando l'analisi viene applicata a testi lunghi, sorgono alcuni problemi (Michel et al. 2011). In particolare, è difficile dedurre la prevalenza di un *topic* che sia coerente con un lungo testo, ad esempio un libro, perché di solito contiene una gamma di argomenti diversi. Non solo, ma il problema dell'adattamento del modello all'analisi è cruciale

perché l'algoritmo LDA richiede che il numero dei *topics* sia specificato a priori: inutile dire che tale parametro influisce sui risultati dell'analisi. A questo proposito, un recente contributo (Sbalchiero et al. 2020) si inserisce in questo dibattito attraverso una serie di esperimenti che applicano l'LDA ai testi lunghi, cercando di gettare nuova luce sulla complessa relazione tra la lunghezza dei testi e la determinazione del parametro rispetto al miglior numero di *topics* da rilevare in un corpus. L'esito del lavoro sottolinea come la lunghezza delle porzioni di testo è una variabile per spiegare il cambiamento nella determinazione del miglior numero di *topics* da rilevare in un corpus. Di conseguenza, seguendo i risultati proposti, viene formulata la regola Sbalchiero-Eder, sotto forma di modello matematico (*ivi*, p. 1103), che mette in relazione il numero ottimale di *topics* da estrarre in un corpus e la determinazione della dimensione delle porzioni di testo utile all'organizzazione del corpus: dato un corpus, il numero migliore di *topics*, come parametro necessario all'implementazione dell'LDA, è inversamente proporzionale alla lunghezza delle porzioni di testo, ovvero più grandi sono le porzioni di testo, minore è il numero di *topics* da rilevare.

La classificazione e il concetto di distanza

Particolarmente interessante risulta il concetto di distanza quando applicato allo studio di collezioni di testi. Utilizzata nell'ambito della stilometria, ovvero l'analisi statistica dello stile linguistico e letterario, tra le varie applicazioni possibili molto diffuso è il suo utilizzo nell'ambito dell'attribuzione d'autore (Tuzzi et al. 2018). Il presupposto è che le caratteristiche di ogni autore, ovvero il suo stile, sarebbero rintracciabili e quantificabili attraverso lo studio dei testi al fine di distinguere, ad esempio, un autore da un altro (Holmes 1998, Joula 2006). Se, ad esempio, l'obiettivo principale di un'analisi fosse quello della classificazione dei testi tramite procedure di *clustering* (Berry 2004), ovvero un tipo specifico di classificazione di documenti finalizzata a raggruppare in *cluster* testi simili, separati da quelli dissimili a formare gruppi distinti, occorre una distanza (Sbalchiero et al. 2016).

Il concetto di distanza può essere inteso, quindi, nei termini di una misura di somiglianza in grado di valutare fino a che punto ogni testo può essere considerato simile o diverso da un altro.

A livello intuitivo, la distanza intertestuale, ovvero tra due testi, si basa sulla seguente assunzione: se due testi sono identici, tutte le parole compaiono nei due testi con la stessa frequenza e la distanza risulta pari a zero. Si considerino, ad esempio (Tab. 3), i seguenti due testi: testo A *(cambiare le carte in tavola)*, testo B *(la classe non è acqua*).

Tab. 3 Esempio di distanza tra due testi A e B.

Forma	testo A	testo B	$f_{i,A} - f_{i,B}$
Acqua	0	1	1
cambiare	1	0	1
Carte	1	0	1
classe	0	1	1
è	0	1	1
in	1	0	1
La	0	1	1
Le	1	0	1
non	0	1	1
Tavola	1	0	1
	5	5	10

$$d(A,B) = \frac{10}{5+5} = \frac{10}{10} = 1$$

Ne consegue che la distanza raggiunge il massimo teorico 1 quando due testi non hanno nessuna parola in comune, ovvero presentano una distanza massima. Viceversa, se due testi sono identici, tutte le parole compaiono nei due testi con la stessa frequenza e la distanza, in questo caso, risulta pari a 0. Per chiarire quanto detto, può risultare utile un richiamo alla distanza intertestuale di Labbé (Labbé et al. 2001, Labbé 2007, Tuzzi 2010, Cortelazzo et al. 2013), che tra le numerose misure di distanza disponibili (Rudman 1998, Trevisani et al. 2020), risulta di assoluto interesse.

Nello specifico, l'elenco delle parole, e il numero di occorrenze corrispondenti, rispecchia quello che può essere definito il profilo lessicale di ciascun documento. La distanza di Labbé è basata su una somma di differenze tra le frequenze delle parole presenti nei testi. In altri termini,

> «data una coppia di testi A e B di dimensione N_A e N_B con $N_A \leq N_B$, la loro distanza d è:

$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f^*_{i,B} \right|}{2N_A}$$

> dove V_{AuB} rappresenta il vocabolario di A e B e la frequenza $f_{i,B}$ di ogni parola i nel testo più grande B viene ridotta in base alla dimensione del documento più breve A per mezzo di una semplice proporzione
>
> $$f^*_{i,B} = f_{i,B} \frac{N_A}{N_B}$$
>
> La distanza tra A e B è uguale alla distanza tra B e A, cioè la distanza è simmetrica, e la distanza tra ogni testo e se stesso è pari a zero e, più in generale, se due testi contengono le stesse parole con la stessa frequenza, la loro distanza è zero. Se due testi non hanno parole in comune, sono separate da una distanza pari a uno» (Sbalchiero et al. 2016, p. 1339).

La traduzione del concetto di distanza entro la ricerca empirica può risultare di estremo di interesse, come mostrano alcuni lavori che seguiranno, per esempio, per l'analisi dei profili lessicali su corpora di documenti di programmazione o testi istituzionali, ottenendo così la distanza intertestuale tra i documenti analizzati al fine di verificare direttrici comuni tra i testi e il grado di recepimento di direttive più generali. Tra i software utilizzati per applicare questo tipo di analisi va segnalato il pacchetto 'stylo', *stylometry* (Eder et al. 2016), che risulta interessante perché combina analisi diverse e numerose misure di distanza in ambiente R (R development core team 2016).

Considerazioni a margine

Per ricomporre gli innumerevoli argomenti affrontati entro un quadro unitario è opportuno chiarire alcune questioni a margine che, per quanto provvisorie, sono necessarie nel tentativo di rispondere al seguente quesito: quale tipo di conoscenza è possibile ottenere attraverso ricerche che utilizzano metodi di analisi statistica dei dati testuali e tecniche di text mining? La risposta non può che essere di questo tipo: dipende dalla domanda. Si tratta di una delle questioni che attraversa l'insieme dell'esperienza del lavoro scientifico, dalla raccolta del materiale empirico alla restituzione dei risultati, passando dalla formulazione del percorso di indagine. Quanto detto è di fondamentale importanza per riuscire a comprendere i lavori che seguiranno e, soprattutto, per distinguere un percorso di ricerca finalizzato alla *descrizione* da un orientamento votato alla *comprensione* dei fenomeni (Sbalchiero 2021). La risposta a questa domanda, in effetti, segna la differenza tra descrivere, nel senso di riassumere un insieme di testi o, di converso, operare una rigorosa analisi di questi testi

attraverso la costruzione di evidenze empiriche ed elementi interpretativi in grado di cogliere quel fenomeno. Un'affermazione di questo tipo potrebbe sembrare banale, ma non lo è affatto: implica il costante misurarsi con i significati che orientano la ricerca e costituisce uno spartiacque tra il mero gioco intellettuale e l'attività scientifica. È chiaro che formulare delle strategie entro un percorso di ricerca significa operare delle scelte e rappresentarsi, fin dall'inizio, il percorso stesso che, a sua volta, dovrà essere coerente agli obiettivi conoscitivi che ci si prefigge di raggiungere. In tal senso,

> «un progetto di ricerca che possa definirsi rigoroso, e la relativa riflessione metodologica, devono saper coniugare un orientamento strategico alle tecniche, alle analisi e alle eventuali verifiche empiriche che il ricercatore ritiene più adatti. In tal senso non esiste il metodo migliore o la tecnica migliore, ma esistono scelte più o meno pertinenti, sulla base della loro appropriatezza e coerenza rispetto al problema selezionato» (Sbalchiero 2021, p. 108).

Un ulteriore aspetto è degno di nota: i lavori che seguiranno si collocano entro esperienze di ricerca attraverso l'uso di strumenti che cambiano a seconda delle finalità perseguite. Questo punto è estremamente rilevante perché, va ribadito con forza, l'utilizzo di metodi e tecniche non può prescindere da una riflessione sul disegno della ricerca e, quindi, sulla domanda di ricerca che coinvolge specifiche scelte pratiche e operative. Formulare una domanda di ricerca rispetto a un fenomeno significa chiedersi non soltanto come possa essere affrontato e definito, ma anche dove ci debba portare la ricerca che stiamo conducendo. Questo significa, allo stesso tempo, mantenere aperta la possibilità di apportare delle modifiche qualora determinate scelte presuppongano una riconfigurazione stessa del processo di ricerca oppure, come di sovente capita, quando si intravede la possibilità di operare ulteriori approfondimenti coniugando diversi approcci. Ad esempio, l'esito della *topic detection*, ottenuta con l'LDA, potrebbe essere utilizzata per approfondimenti tematici di determinati *topics* tramite il metodo Reinert, avendo a disposizione una classificazione di testi organizzati per argomenti. In questo caso, l'esito sarà un focus specifico su un argomento che verrà ulteriormente approfondito attraverso l'estrazione e l'analisi dei mondi lessicali caratterizzanti quel *topic*. Seguendo questo ragionamento, le possibilità che si aprono sono innumerevoli. Una buona regola rimane, comunque, quella di descrivere le scelte fatte durante il percorso di ricerca in modo tale che gli strumenti e gli approcci utilizzati risultino complementari, e non alternativi, anche al lettore. In

questo senso, di nuovo, è sempre necessario valutare caso per caso, sulla base di quali sono gli obiettivi perseguiti, qual è l'approccio più adeguato, nella consapevolezza che l'analisi automatica di testi non dovrebbe essere vista come un'alternativa ai più tradizionali approcci di tipo qualitativo, piuttosto li integra, mettendo a disposizione una vasta gamma di strumenti software e statistici che continuamente offrono nuove possibilità di ricerca. A fronte degli sviluppi nell'analisi statistica dei testi e del text mining, assieme alla crescente accessibilità a grandi collezioni di materiale empirico digitale, che ampliano notevolmente le opportunità di ricerca, si può affermare che i diversi percorsi costituiscono un proficuo terreno di incontro tra ricercatori di diversa estrazione. Infatti, anche quando il materiale impiegato nella ricerca ben si adatta all'elaborazione statistica, occorre tenere a mente che necessita di conoscenze, competenze e intuizioni che nella fase di interpretazione dei risultati rimangono, per ora, di stretta competenza umana.

Analisi della giurisprudenza per supportare le strategie regionali a tutela dei consumatori

Salvatore Pinello[1]

Tutela dei consumatori, Giurisprudenza, Decisioni di policy, Regione Veneto.

Introduzione

Le politiche regionali in materia di tutela dei consumatori si sostanziano principalmente in iniziative in tema di informazione, formazione e assistenza dei consumatori.

Snodo cruciale nel disegno di queste politiche è l'individuazione delle aree in cui sussiste un particolare bisogno di tutela, cioè delle aree in cui l'azione regionale a tutela del consumatore possa essere utilmente rafforzata.

In questo senso, il corredo informativo attualmente a disposizione degli attori della decisione nel contesto regionale veneto è riferito principalmente alla platea di coloro che si rivolgono alle associazioni dei consumatori per la tutela dei propri interessi, in virtù dell'interlocuzione privilegiata tra Regione e tali associazioni.

Le politiche regionali, tuttavia, intendono rivolgersi al novero ben più ampio costituito dalla generalità dei consumatori presenti nel territorio. In molti casi, infatti, il consumatore in difficoltà non sceglie di appoggiar-

[1] Funzionario con profilo giuridico-amministrativo della Regione del Veneto, tra il 2018 e il 2022 è stato responsabile della Posizione Organizzativa Tutela dei consumatori.

si ad un'associazione, ma si affida direttamente ad un'assistenza di tipo legale.

Appare auspicabile, pertanto, per la realizzazione di strategie regionali efficaci, poter accedere anche a informazioni riferite a gruppi differenti rispetto a coloro che si rivolgono alle associazioni dei consumatori, in modo da poter disporre di un quadro il più esteso e comprensivo di informazioni e dati sui bisogni dei consumatori cui prioritariamente dare risposta.

In questo lavoro volgeremo nello specifico lo sguardo verso coloro che si trovano a difendere i propri interessi di consumatori di fronte a un giudice, tenendo quindi in considerazione i bisogni espressi da un numero ben più ampio di consumatori rispetto a quelli inclusi nell'attuale disegno di policy regionale veneto.

Quali circostanze portano un consumatore a rivolgersi a un giudice per la tutela dei propri interessi, o a essere chiamato in giudizio? E quali questioni si trova ad affrontare?

Per tentare di rispondere a queste domande, nel corso del capitolo sarà presentata l'analisi testuale automatizzata di un insieme di oltre mille pronunce giurisdizionali in materia di tutela dei consumatori tratte da una banca dati giuridica già in uso all'Amministrazione.

La semplice lettura delle pronunce giurisdizionali, soprattutto se molto numerose, non riuscirebbe a disvelare infatti molte delle informazioni racchiuse in tali testi. Con l'ausilio della tecnologia e di specifici software è stato possibile estrarre dai testi informazioni che altrimenti sarebbero rimaste nascoste, e che possono invece rivelarsi utili per supportare le decisioni di policy in materia di tutela dei consumatori.

Come accennato, le pronunce oggetto di analisi sono state tratte da una banca dati giuridica già in uso all'Amministrazione. Nell'impostare lo studio, la previsione dell'utilizzo in forma innovativa di strumenti già in uso all'Amministrazione, con lo scopo di trarne nuova conoscenza, è parso elemento di ulteriore interesse, utile ad agevolare un'eventuale messa a regime dello strumento innovativo qui proposto.

La tecnica utilizzata è inoltre agevolmente trasferibile anche ad altre banche dati giuridiche o amministrative e ad altri ambiti di conoscenza, potendo essere utilizzata, pertanto, a supporto delle decisioni di policy della Regione anche in altre materie. Si pensi, ad esempio, alla materia tributaria, e alla possibilità di estrarre in modo innovativo da raccolte di testi – ricorsi dei contribuenti o sentenze delle Commissioni tributarie – già nella disponibilità dell'Amministrazione, informazioni utili a inqua-

drare e analizzare le questioni più frequentemente foriere di controversie in tale ambito. Ciò consentirebbe di intervenire in modo puntuale sulla normativa tributaria o sulle relative prassi applicative, allo scopo di prevenire dette controversie.

In ultima analisi, tornando all'ambito di specifico interesse del presente lavoro, obiettivo pratico è fornire agli attori delle diverse fasi della decisione un pilastro informativo ulteriore e non sovrapponibile rispetto a quelli già utilizzati a supporto delle strategie regionali a tutela dei consumatori, con il duplice vantaggio di un arricchimento delle informazioni 'sul tavolo' e di un possibile stimolo affinché tutte le ipotesi siano accompagnate da un corredo di informazioni e dati a supporto.

Il capitolo è organizzato in tre parti. Nella prima parte sarà illustrato il disegno della ricerca, presentando metodi e strumenti e ripercorrendo i passaggi fondamentali che hanno portato alla selezione delle pronunce da includere nel corpus oggetto di analisi. Nella seconda parte sarà illustrata l'analisi automatica del contenuto delle pronunce eseguita con l'utilizzo del software Iramuteq, che tramite l'identificazione di *cluster* semantici ha permesso di addentrarci in misura significativa nelle tematiche, talune delle quali trattate con maggiore frequenza e/o rivelatrici di caratteristiche interessanti e inattese delle controversie in esame, suscettibili di suggerire altrettanti spunti utili a calibrare le politiche regionali in materia. Infine, saranno svolte alcune considerazioni conclusive sui risultati raggiunti e sulle prospettive aperte dall'analisi.

Le domande di conoscenza relative alla banca dati giuridica e alla giurisprudenza in materia di tutela dei consumatori

Le politiche di tutela dei consumatori

Nell'ambito delle politiche regionali in materia di tutela dei consumatori hanno particolare rilevanza le azioni volte a promuovere iniziative in tema di informazione, formazione e assistenza dei consumatori. Periodicamente, anche in funzione della disponibilità di risorse, le regioni adottano programmi in materia di tutela dei consumatori che prevedono simili iniziative. Tra le questioni che si pongono nella fase di elaborazione di tali programmi, e quindi tra le domande che si possono formulare nella fase di applicazione dell'analisi testuale, vi è quella riguardante la scelta sui temi di fondo verso cui orientare le iniziative a tutela dei consumatori.

È preferibile promuovere iniziative in tema di sicurezza dei prodotti

o di educazione finanziaria? Informare in tema di *e-commerce* o di utenze energetiche?

Sebbene le indicazioni programmatiche in materia di tutela dei consumatori periodicamente elaborate a livello europeo e a livello nazionale possano fornire un'utile cornice, tuttavia esse non esauriscono di per sé, pure ove prese a riferimento, la questione delle scelte da operare in merito ai temi da porre al centro delle iniziative di formazione, informazione e assistenza.

Tali scelte restano quindi integralmente in capo alla molteplicità di attori chiamati a partecipare, a livello regionale, ai processi decisionali in materia di tutela dei consumatori, che nel caso della Regione Veneto[2] (analogamente a quanto avviene, nella sostanza, nelle altre regioni), comprende:

- la Giunta Regionale, che adotta le decisioni;
- il Comitato regionale dei consumatori e degli utenti, che esprime pareri e formula proposte, tra l'altro, sugli atti di programmazione, i progetti di legge e gli interventi in materia, ed è a sua volta composto da:
    - Assessore competente, che lo presiede;
    - rappresentanti delle associazioni dei consumatori;
    - rappresentanti di comuni, province e camere di commercio;
    - dirigenti delle strutture amministrative della Regione;
- associazioni dei consumatori iscritte nel registro regionale istituito con legge, che possono presentare progetti di iniziative.

Le ipotesi e le proposte provenienti da ognuno di questi attori, pur basate sulla rappresentazione delle necessità dei consumatori che emerge dalle esperienze e conoscenze di ciascuno di essi, non sempre sono accompagnate da un corredo di informazioni e dati sui bisogni reali dei consumatori e, anche quando questo corredo è presente, deriva principalmente dalle richieste dirette alle associazioni dei consumatori, escludendo di fatto i bisogni di tutti coloro che non operano la scelta di

[2] Viene qui esposto il quadro risultante, per semplificazione, dalla Legge regionale 23 ottobre 2009, n. 27 "Norme per la tutela dei consumatori, degli utenti e per il contenimento dei prezzi al consumo". Il novero dei procedimenti decisionali e dei possibili attori coinvolti nella decisione previsti da tale legge e dalle discipline che di volta in volta possono ad essa sovrapporsi (ad es. bandi relativi a finanziamenti sovraregionali) prevede una pluralità di schemi che possono coinvolgere anche soggetti ulteriori.

rivolgersi ad esse. Così come testimoniano recenti rilevazioni[3], infatti, solo una minoranza dei consumatori che si imbattono in circostanze che li inducono a cercare assistenza e tutela si rivolge alle associazioni dei consumatori. Di contro, le politiche regionali intendono coinvolgere il novero ben più ampio costituito dalla generalità dei consumatori presenti nel territorio, rendendo quindi auspicabile l'accesso ad informazioni che ricomprendano l'intera (o quasi) popolazione di riferimento.

Per rispondere a tale necessità, il presente studio propone un approccio innovativo al problema, tramite l'impiego di dati derivanti dalla giurisprudenza in materia di tutela dei consumatori e l'applicazione di tecniche di analisi testuale automatizzata, alla ricerca delle tematiche sottoposte all'attenzione dei giudici nelle controversie che vedono come parte un consumatore.

Le pronunce giurisdizionali oggetto di esame sono state tratte da una banca dati giuridica già in uso alle strutture regionali, la cui consultazione è un'attività che avviene di consueto e in modo puntuale al livello dell'Amministrazione regionale, per trarne informazioni di carattere squisitamente giuridico e con il fine di meglio comprendere singole questioni di diritto. In questa sede, la medesima banca dati è stata utilizzata con un differente scopo, ricavando un nucleo di oltre mille pronunce da esaminare non singolarmente ma con uno sguardo d'insieme, non solo al fine di individuare gli elementi di carattere giuridico, ma le circostanze che portano un consumatore di fronte a un giudice e le tematiche che in tali giudizi vengono trattate, per meglio orientare le policy regionali in materia. Infine, l'applicazione di tecniche di analisi testuale automatica consente in tempi relativamente brevi di aprire, su un insieme consistente di pronunce, una prospettiva d'insieme che sarebbe impossibile raggiungere con un'analisi di tipo 'tradizionale' – cioè attraverso lettura e massimazione di ciascuna delle pronunce selezionate ed estrazione e classificazione 'manuale' di informazioni riferite a circostanze di fatto e tematiche affrontate – a parità di tempo e di ricchezza di informazioni ricavate.

In conclusione, come già ribadito, l'ipotesi è che da un simile studio si possano trarre informazioni e dati utili a supportare le decisioni di policy della Regione in materia di tutela dei consumatori, in particolare

[3] Si veda il sondaggio "L'evoluzione delle conciliazioni paritetiche tra consumatori e aziende – Consumers Survey", Ipsos, 2019 (Committente Consumer's Forum, diffuso in https://www.helpconsumatori.it/secondo-piano/consumers-forum-conciliazione-paritetica-una-sconosciuta/180021).

per l'individuazione dei bisogni dei consumatori cui prioritariamente dare risposta e quindi dei temi di fondo oggetto delle iniziative in tema di informazione, formazione e assistenza dei consumatori di volta in volta promossi dalla Regione stessa, sulla base del presupposto che alle tematiche oggetto dei giudizi che vedono come parte un consumatore corrispondano aree in cui possa sussistere un particolare bisogno di tutela, e quindi aree in cui l'azione a tutela del consumatore possa essere utilmente rafforzata. Un consumatore meglio informato, formato, assistito *prima* su una determinata tematica, potrà *dopo*, con maggiore probabilità, meglio tutelare le proprie ragioni in sede giurisdizionale o, magari, evitare del tutto il ricorso a un giudice o il richiamo in giudizio, con il dispendio di tempo, energie e risorse che ciò comporta. Si veda la Fig. 1 di modellizzazione grafica del disegno di ricerca.

Fig. 1 Modellizzazione del disegno di ricerca.

La banca dati e la selezione delle pronunce giurisdizionali

La fonte dalla quale sono state tratte le pronunce oggetto di analisi è la banca dati giuridica 'Leggi d'Italia' dell'editore Wolters Kluwer Italia[4], che include raccolte relative a normativa, prassi, pronunce giurisprudenziali, riviste specializzate e altro (Fig. 2).

Come già ricordato in precedenza, si tratta di una banca dati attualmente già in uso alla Regione del Veneto, all'interno della quale, la sezione 'Banche dati di giurisprudenza', di interesse ai fini dell'analisi, presenta un'ulteriore suddivisione nelle seguenti sezioni:

- Repertorio giurisprudenza;
- Corte Costituzionale;

[4] La banca dati, fornita a pagamento, può essere raggiunta all'indirizzo http://online.leggiditalia.it/.

- Cassazione Civile;
- Cassazione Penale;
- Consiglio di Stato e TAR;
- Corte dei Conti;
- Corti di Merito.

Fig. 2 Banca dati 'Leggi d'Italia': schermate di consultazione.

La suddivisione in sezioni ricalca il criterio dell'organo di emanazione delle diverse pronunce, e non la distinzione tra le singole materie giuridiche o campi di interesse (ad esempio diritto tributario, diritto societario, diritto di famiglia, ecc.).

Se la materia "tutela dei consumatori" potrebbe pertanto essere rinvenuta in pronunce catalogate in differenti sezioni tra quelle presenti nella banca dati in esame, ai fini della presente ricerca l'interesse è caduto in prima battuta sulla sezione 'Corti di Merito', nell'ipotesi che dal testo di una pronuncia di merito possano emergere, in misura pari e forse superiore a quanto non sia possibile ad esempio in sede di legittimità (sezioni 'Cassazione Civile' e 'Cassazione Penale'), riferimenti alle circostanze di fatto che hanno portato il consumatore di fronte a un giudice.

Secondo la descrizione fornita nel sito internet ufficiale di 'Leggi d'Italia'[5], la banca dati 'Corti di Merito' consiste in «un'ampia selezio-

[5] https://www.leggiditaliaprofessionale.it.

ne delle principali pronunce di merito (Corti d'Appello, Corti d'Assise, Tribunali, Tribunali dei minori, Giudici di pace) emanate dai maggiori Fori italiani e depositati a partire da gennaio 2000». Una rapida consultazione della sezione, senza l'inserimento di alcuno specifico criterio di ricerca, indica un numero complessivo di 322.892 'risultati' o pronunce[6]. Si tratta di un numero certamente rilevante di testi, tale da rappresentare un campione sufficientemente rappresentativo degli orientamenti giurisprudenziali italiani. Tuttavia, va subito chiarito che esso non rappresenta una raccolta completa di tutte le pronunce delle corti di merito italiane, considerato che si calcola nell'ordine dei milioni il numero di pronunce emesse ogni anno in Italia (Celotto 2015). Pur con la consapevolezza di questo limite, il lavoro eseguito sulle pronunce ha, in ogni caso, consentito di esplorare il tema oggetto di analisi, acquisendo una maggiore conoscenza e formulando ipotesi e riflessioni.

Ciascuna pronuncia è contrassegnata, all'interno della sezione, da una o più 'voci' ed eventuali 'sottovoci', ossia 'etichette' applicate a ciascuna pronuncia, che indicano la tematica trattata o interessata dalla stessa e consentono all'utente di selezionare ed 'estrarre' le sole pronunce di interesse. In una lista di 230 voci in totale, è presente anche la voce 'consumatore (tutela del)', certamente confacente rispetto agli scopi della ricerca.

Anche altre voci, tra le 230 presenti, avrebbero potuto in astratto riferirsi a controversie in materia di tutela dei consumatori (ad esempio 'obbligazioni e contratti', 'danni in materia civile e penale', '*leasing*'). Tuttavia, ai fini della presente ricerca è apparso di maggiore utilità fare riferimento alla sola voce generale 'consumatore (tutela del)'. Infatti, pur se in tal modo non sono probabilmente state selezionate, tra le pronunce presenti nella banca dati, tutte quelle che vedono come parte un consumatore, si è disposto ugualmente di una selezione numericamente significativa di decisioni, coprendo con tutta probabilità uno spettro ampio, e potenzialmente esaustivo, di circostanze e tematiche tra quelle di interesse.

Dopo aver selezionato nel campo di ricerca 'Voci selezionate', la voce 'consumatore (tutela del)', è stata estratta una lista di tutte le pronunce contenute nella banca dati 'Corti di Merito' contrassegnate da tale voce-etichetta, per un totale di 1.124 'risultati' o pronunce, depositate fra il 30 gennaio 2001 e il 31 marzo 2021.

[6] Questo dato, così come gli altri relativi alla banca dati in esame presentati in questo lavoro, è riferito a una consultazione effettuata in data 7 giugno 2021.

Nella lista è presente, per ogni pronuncia, l'informazione relativa alla voce o alle voci ad essa associate, ad esempio 'consumatore (tutela del)' o "consumatore (tutela del) – ingiunzione (procedimento per) – obbligazioni e contratti", ed un breve scritto contenente il collegamento ipertestuale alla pronuncia che specifica le informazioni di base relative a organo giudicante, località, tipo di pronuncia, data (ad esempio "Tribunale Cuneo, Sentenza, 31-03-2021"), informazioni che sono state utilizzate per effettuare una classificazione delle pronunce disponibili utile anche all'impostazione dell'analisi.

Nella Fig. 3 è riportato il numero di pronunce estratte suddivise per anno e organo giudicante.

Fig. 3 Numero di pronunce estratte suddivise per anno e organo giudicante.

PRONUNCE ANNO/ORGANO GIUDICANTE						
	Autorità Garante della Concorrenza e del Mercato	Commissione Tributaria	Giudice di Pace	Tribunale	Corte d'Appello	TOTALE
2001				1	2	3
2002				1	1	2
2003				1	0	1
2004				1	0	1
2005			14	9	2	25
2006			5	17	3	25
2007		1	1	23	4	29
2008			2	34	5	41
2009	3		4	31	2	40
2010	1		7	29	1	38
2011			3	32	4	39
2012			2	31	1	34
2013		1	3	35	3	42
2014			0	32	1	33
2015			2	40	2	44
2016			2	43	3	48
2017			6	40	7	53
2018		2	4	86	14	106
2019			9	115	27	151
2020		3	4	207	58	272
2021		3	2	70	22	97
TOTALE	4	10	70	878	162	1124

Dalla figura appare evidente, a partire dall'esiguità dei numeri, la peculiarità rappresentata dalla presenza, nella lista di pronunce, di quattro provvedimenti dell'Autorità Garante della Concorrenza e del Mercato (AGCM) e di dieci sentenze di Commissioni tributarie.

La presenza, nella banca dati 'Corti di Merito' alla quale abbiamo attinto, di provvedimenti sanzionatori dell'AGCM, che non hanno natura giurisdizionale e sono pronunciati da un'autorità che non fa parte dell'ordine giudiziario, appare frutto, se non di un errore, di un metodo di classificazione che non coincide con i fini del presente studio. Per tale ragione, tali pronunce non sono state incluse nel corpus.

Quanto alle dieci sentenze delle Commissioni tributarie, anch'esse non sono state incluse nel corpus. Da un'analisi di tali dieci sentenze risulta, infatti, che quattro trattano argomenti che nulla hanno realmente a che vedere con il tema della tutela del consumatore, mentre le rimanenti sei attengono alla questione dell'assoggettabilità al c.d. 'contributo unificato'[7] dei giudizi, promossi dalle associazioni dei consumatori, tema molto specifico che esula dall'intento della presente ricerca.

All'esito dello stralcio dei quattordici provvedimenti di AGCM e Commissioni tributarie, residua dunque una selezione di 1.110 pronunce.

Le pronunce giurisdizionali: analisi dei bisogni dei consumatori per favorire migliori decisioni di policy

Per la preparazione del corpus da sottoporre ad analisi testuale, le pronunce sono state classificate in base a un gruppo di 14 variabili, riferite principalmente al tempo e al luogo di emissione, alla tipologia e all'organo emanante di ogni singola pronuncia.

L'utilizzo di tali variabili nelle elaborazioni proposte di seguito ha consentito, ad esempio, di effettuare analisi sull'evoluzione storica delle informazioni desumibili dalle pronunce in esame.

Nel seguito del capitolo sarà utilizzata in particolare la variabile 'Periodo2', con la quale ciascuna pronuncia è stata classificata in base alla data di emissione e collocata in periodi temporali corrispondenti di massima a quinquenni (anni 2001-2006 e quinquenni successivi, l'ultimo dei quali è il quinquennio 2017-2021). I testi delle pronunce non sono stati sottoposti ad ulteriori azioni di *pre-processing*, allo scopo di favorire la rapidità delle analisi, e quindi la proponibilità dell'approccio qui utilizzato quale strumento dell'attività amministrativa che consenta di svolgere con rapidità analisi su corpora di pronunce giurisdizionali.

[7] Il "contributo unificato di iscrizione a ruolo" è la forma di tassazione correlata alle spese degli atti giudiziari dovuta dalla parte che dà avvio a un processo civile, amministrativo o tributario.

I temi nelle controversie che vedono come parte un consumatore

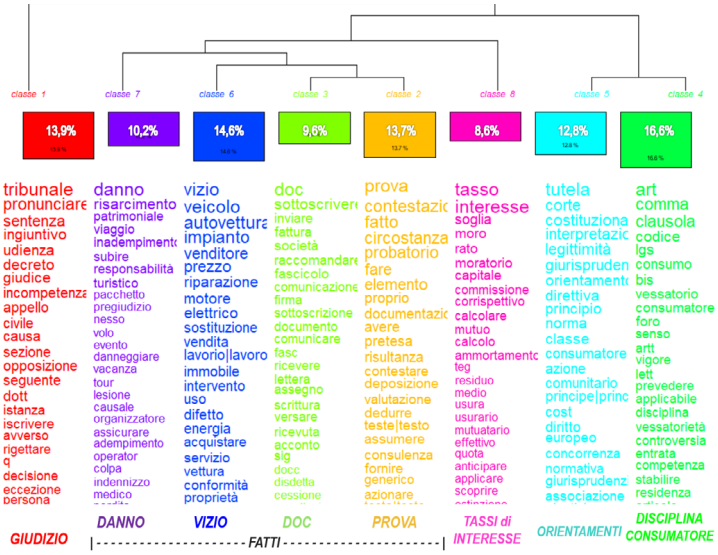
Nel paragrafo è presentata la *topic detection* delle pronunce giurisdizionali, eseguita con il metodo Reinert – volto all'estrazione dei *topics* alla base delle pronunce – per la quale è stato utilizzato il software Iramuteq.

In questo modo è stata ottenuta una classificazione per *topics* o nuclei tematici dell'intero corpus, individuando i 'mondi lessicali' alla base delle singole classi (*cluster*), dotate di specifici vocabolari.

Un'elaborazione che ha richiesto 12 minuti e 59 secondi.

Come illustrato dal dendrogramma in figura 4, il processo di *clusterizzazione* automatica ha individuato la presenza nel corpus di 8 classi.

Fig. 4 I temi più frequenti presenti nelle pronunce.



Il grafico illustra i passaggi che hanno portato all'individuazione delle classi, la loro consistenza quantitativa e le parole che maggiormente le caratterizzano, permettendo di individuare il tema sotteso ad ognuna di esse. Per ciascuna classe è stato aggiunto dall'autore, nella parte bassa della figura e in caratteri maiuscoli, un termine chiave scelto per definire sinteticamente i vari temi.

Il grafico in figura 4 riporta un estratto del vocabolario che caratterizza ciascuna classe, dove il massimo livello di significatività dell'associazione tra parola e classe è rappresentato dalla maggiore dimensione del carattere di ciascuna parola.

A ciascuna classe è stata manualmente assegnata un'etichetta (in caratteri maiuscoli in corrispondenza di ogni classe nella parte bassa del grafico in figura 4), recante uno o più termini chiave tra quelli presenti nel relativo vocabolario, allo scopo di condensare il risultato dell'operazione interpretativa dei diversi nuclei tematici.

Posto che il fulcro del presente lavoro consiste nella ricerca di dati relativi alle tematiche sottoposte all'attenzione dei giudici nelle controversie che vedono come parte un consumatore, ci soffermeremo ora con maggiore dettaglio sul contenuto delle singole classi.

Nell'esposizione che segue, le classi verranno esaminate non in ordine numerico, ma secondo l'ordine di *clusterizzazione* operato dal software, con l'utilizzo delle parole più significative ad esse associate.

Classe 1 "giudizio". La classe 1 (di colore rosso) raccoglie i riferimenti procedurali presenti nella struttura di ogni giudizio, e quindi nel testo di ognuna delle pronunce ("tribunale", "pronunciare", "sentenza", "udienza", "decreto", "giudice", "appello", "istanza", "rigettare", "decisione", ecc.), ma presenta anche un chiaro riferimento al peso assunto nelle pronunce dal particolare meccanismo dell'opposizione a decreto ingiuntivo. Entrambi questi aspetti rivestono un'importanza particolare ai fini della presente ricerca. Da una parte, il software ha riconosciuto e separato dal resto del corpus la parte procedurale presente in ogni pronuncia, evitando che questi contenuti, quantitativamente rilevanti, potessero divenire un 'rumore' tale da rendere meno chiara l'analisi sul resto dei contenuti. Dall'altra, il riferimento così forte al tema dell'opposizione a decreto ingiuntivo fornisce un elemento, in parte inatteso, molto rilevante in relazione agli obiettivi della ricerca. Il decreto ingiuntivo, infatti, è un atto con il quale viene ingiunto a un soggetto – nella tipologia di controversie in esame, di solito, il consumatore – di effettuare un pagamento. Per evitare di essere sottoposto a escussione, entro quaranta giorni costui deve proporre a un giudice la sua opposizione, instaurando così la controversia vera e propria, all'interno della quale potrà tentare di far valere le proprie ragioni, opponendo, ad esempio, che il pagamento non è dovuto perché il contratto si è per qualche ragione risolto. Talvolta però, il consumatore non sarà in grado di ottenere ragione per non avere attivato correttamente i diritti che pure le norme gli attribuiscono. Anzi, se avesse attivato correttamente i propri diritti, probabilmente avrebbe potuto evitare del tutto di ritrovarsi in giudizio. Focalizzare e approfondire questi meccanismi può senz'altro rivelarsi utile nel fornire spunti utili a calibrare le politiche regionali in materia di tutela dei consumatori.

Classe 4 “disciplina del consumatore”. La classe 4 (in verde, a destra) corrisponde al tema della disciplina (a tutela del) consumatore. Il riferimento forte è al “codice” del “consumo”, ovvero al decreto legislativo (“lgs”) n. 206 del 2005, alla disciplina a “tutela” del “consumatore”, anche con interessanti riferimenti a questioni specifiche come quella delle clausole vessatorie (“clausola”, “vessatorietà”) o della “competenza” territoriale del giudice del luogo di “residenza” del “consumatore” (c.d. “foro” del consumatore).

Classe 5 “orientamenti”. La classe 5 (di colore azzurro), affiancata nella Fig. 4 alla classe 4, attiene agli orientamenti interpretativi (“interpretazione”) sulla “normativa” a “tutela” del “consumatore”, rinvenibili in particolare nella “giurisprudenza” “costituzionale”, di “legittimità”, o anche della “corte” comunitaria (“comunitario”), posto che la normativa in materia è in larga parte di derivazione europea (“direttiva”, “diritto”, “europeo”). Una classe che può aprire una specifica finestra sul c.d. “diritto vivente”, permettendo di risalire a specifici orientamenti interpretativi e concretamente applicativi che potrebbero non emergere dalla semplice analisi delle normative.

Classe 8 “tassi di interesse”. La classe 8 (di colore ciclamino) corrisponde al tema dei tassi di interesse e più in generale alla tematica finanziaria, come emerge inequivocabilmente dal relativo vocabolario e dai riferimenti a “tasso” di “interesse”, interesse “moratorio”, interesse da considerare “usurario” se supera una determinata “soglia”, “usura”, “teg”, ecc., con riferimenti ai c.d. contratti bancari (commissione di massimo scoperto - “scoprire” - “mutuo”, “ammortamento”, “mutuatario”, ecc.).

Da ultimo, le rimanenti classi 7, 6, 2 e 3 possono essere trattate come una macrocategoria, presentando le stesse, in modo prevalente, riferimenti ai fatti, in un’accezione ampia, che include diversi profili: la contrattazione e il bene o servizio che ne è oggetto (classi 7 e 6), le circostanze successive relative a lamentati inadempimenti, danni (classe 7) e vizi (classe 6), le dimostrazioni da fornire in giudizio attraverso prova (classe 2) e documentazione (classe 3). Di seguito un'analisi delle quattro classi.

Classe 7 “danno”. La classe 7 (in viola) è fortemente caratterizzata dal tema del “danno”, “patrimoniale” o meno, conseguente a un lamentato “inadempimento” o inesatto “adempimento”, e del relativo “risarcimento” richiesto da colui che l’ha “subito”. La classe evidenzia un forte riferimento alla materia dei viaggi e dei pacchetti turistici (“pacchetto turistico”, “volo”, “vacanza”, “organizzatore”, “*tour operator*”).

Classe 6 "vizio". La classe 6 (di colore blu) si caratterizza per il riferimento all'emergere, successivamente alla contrattazione, di un "vizio", con ciò che consegue anche in termini di responsabilità del "venditore" e di richieste di riduzione o restituzione del "prezzo", o anche di "riparazione" o "sostituzione" del bene che presenta un "difetto" di "conformità". Sono rinvenibili nella classe riferimenti a vicende di vario genere, tra le quali caratterizzanti quelle relative a "veicoli" ("autovettura", "motore", "vettura"), a "riparazioni", a "lavori", a utenze elettriche ("elettrico") ed energetiche ("energia"), alla compravendita di immobili ("immobile").

Classe 2 "prova". La classe 2 (in giallo-arancio) corrisponde al tema della "prova" che in un giudizio è necessario "fornire" a supporto di ogni "contestazione" o "pretesa" circa ogni "elemento" di "fatto", anche attraverso "documentazione", testimonianze ("deposizione", "teste"), "consulenze" tecniche.

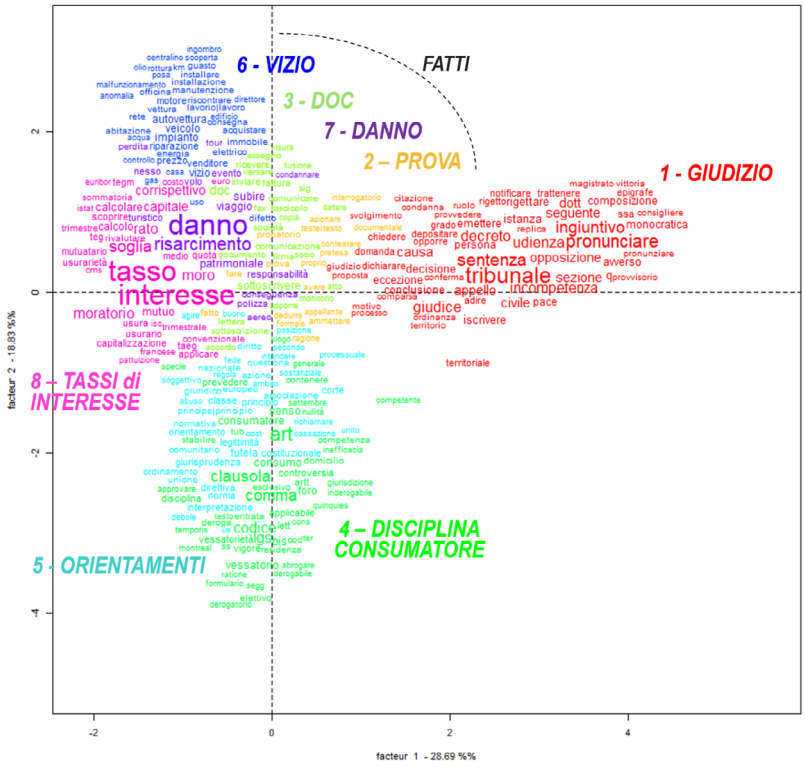
Classe 3 "documentazione". La classe 3 (in verde, al centro del grafico) corrisponde al tema della "documentazione" ("doc", "documento") presente a "fascicolo" perché il giudice vi faccia riferimento. Documentazione *in primis* relativa alla "sottoscrizione" del contratto, ma anche relativa a documenti di diverso tipo ("fattura", "ricevuta", "lettera", "scrittura") e a intervenute comunicazioni rilevanti (raccomandata "raccomandare", "comunicazione", "comunicare", "ricevere", "disdetta").

Per approfondire il rapporto tra le classi appena illustrate, il grafico in Fig. 5 illustra il risultato di un'analisi delle corrispondenze applicata ai *cluster* eseguita con il software Iramuteq. Le otto classi e le parole che le compongono sono rappresentate su un piano cartesiano in cui risulta visibile la distanza o la vicinanza, e in qualche misura la compenetrazione, tra le classi, e dunque fra i temi. Per facilitare la lettura, al grafico proposto da Iramuteq sono stati aggiunti i numeri assegnati alle diverse classi e i termini chiave utilizzati per individuare i corrispondenti temi di cui alla Fig. 4.

In particolare risulta evidente la separazione tra la classe 1 "giudizio" e la restante parte del corpus, che può essere interpretata come una prima suddivisione tra segmenti di testo a tema prevalentemente processuale/procedurale e segmenti a tema prevalentemente "sostanziale". Risulta netta anche la distinzione tra la nuvola formata dalle classi 4 e 5 ("disciplina consumatore" e relativi "orientamenti"), correlate alle questioni giuridiche di rilievo in materia di tutela dei consumatori, e la residua parte del corpus, nella quale si individua facilmente il tema finanziario (classe 8 "tassi di interesse"), e i riferimenti ad elementi relativi ai fatti (restanti 4 classi). In particolare, queste ultime classi si presentano in forte com-

mistione tra loro, pur con una più marcata distinzione, rispetto alle altre, della classe 6 “vizio”.

Fig. 5 Posizionamento delle tematiche su piano cartesiano.



Chiaramente distinti dagli altri, sulla destra il tema processuale/procedurale (classe 1 “giudizio”) e verso il basso i temi relativi a disciplina a tutela del consumatore e relativi orientamenti (classi 4 e 5, “disciplina consumatore” e relativi “orientamenti”). Nel quadrante in alto a sinistra e nelle aree limitrofe, il tema finanziario (classe 8 “tassi di interesse”) e, in forte commistione tra loro, i temi ascrivibili alla macrocategoria “fatti”.

Infine, può essere ulteriormente rinvenuta in Fig. 5 una sorta di bipartizione tra temi a carattere prettamente giuridico, nei quadranti a destra e in basso (da una parte di contenuto processuale/procedurale – classe 1 – dall’altra di contenuto sostanziale – classi 4 e 5), e temi che maggiormente raccontano delle vicende concrete che hanno portato al giudizio, nella parte centrale del grafico e nel quadrante in alto a sinistra. Proprio in queste ultime aree del grafico, e quindi nella classe 8 “tassi di interesse”

e nelle quattro classi ascrivibili alla macrocategoria "fatti", si dovrebbe concentrare maggiormente la ricerca di riferimenti alle circostanze di fatto che portano un consumatore di fronte a un giudice.

Un approfondimento: i *network* del tema "vizio"

Le analisi mostrate hanno permesso, mediante un approccio globale al corpus, di addentrarci in misura significativa nel contenuto delle pronunce, individuando 8 aree tematiche, ciascuna delle quali portatrice di elementi di forte interesse ai fini dello studio.

Allo scopo di meglio cogliere risvolti eventualmente utili ad immaginare in modo più compiuto e dettagliato il disegno delle politiche da promuovere, è stato operato un approfondimento tematico dei contenuti della classe 6 "vizio". La classe 6 appartiene infatti alla macrocategoria "fatti", che, come abbiamo visto, dovrebbe fornire informazioni sulle circostanze di fatto che portano un consumatore di fronte a un giudice.

Tramite il software Iramuteq, è stata dunque realizzata una *network analysis* delle parole maggiormente rappresentative della classe 6, che consente di identificare, a partire da occorrenze e co-occorrenze, la connessione tra le parole, ossia le "reti di parole" che emergono nella classe considerata (Fig. 6).

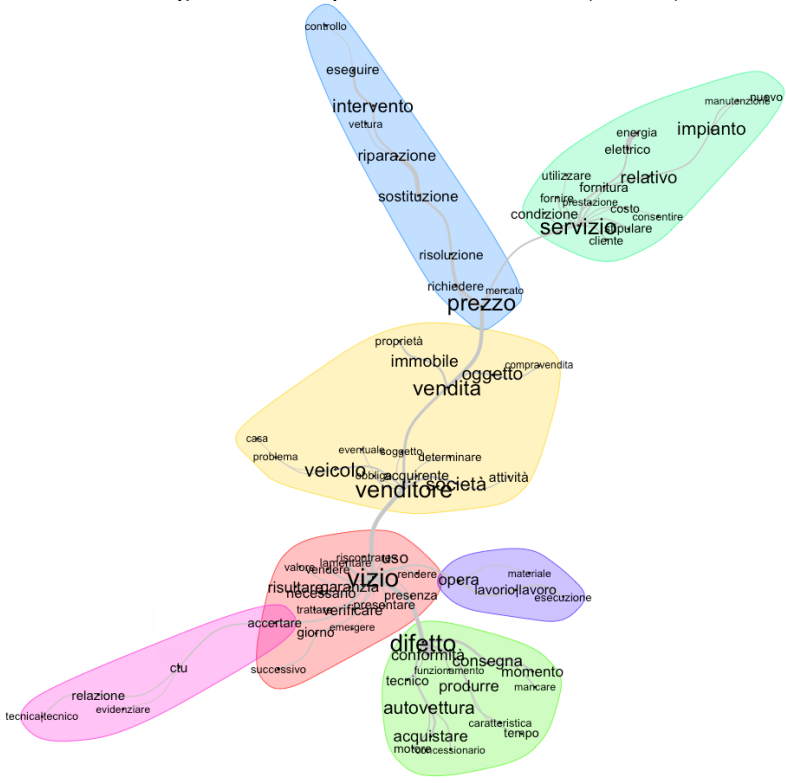
Nel grafico, le parole sono nodi riprodotti con caratteri di dimensione proporzionale alla loro significatività nella classe, e lo spessore della linea che congiunge due parole è proporzionale al numero delle loro co-occorrenze. Per facilitare la lettura del grafico, le parole tra loro maggiormente correlate sono presentate in nuvole di diversi colori.

Parallelamente a quanto fatto nel precedente paragrafo per l'insieme delle classi, nel seguito la classe 6 "vizio" sarà illustrata con l'utilizzo delle parole più significative ad essa associate.

La rete di parole si dirama a partire dal tema del "vizio" (nuvola rossa), che emerge ("risultare", "emergere") in un momento "successivo" a quello della contrattazione, e che porta chi ne lamenta ("lamentare") la "presenza" ad attivare la correlata "garanzia". Un vizio che in giudizio può rendersi necessario "accertare" (nuvola ciclamino) attraverso una "ctu", una consulenza "tecnica" d'ufficio che si sostanzia in una "relazione", che potrà "evidenziare" aspetti rilevanti ai fini della decisione. Immediato il nesso tra vizio e "difetto" (nuvola verde in basso), soprattutto nell'accezione tecnica di "difetto di conformità", che può rivelarsi al "momento" della "consegna", compromettere il "funzionamento" del bene, essere riferito a un aspetto "tecnico", magari al "motore" di una "autovettura" ac-

quistata (“acquistare”) da un “concessionario”. Ancora, dal termine vizio si dipana una connessione diretta con l’area riferita all’“esecuzione” di un’“opera” (nuvola viola), di “lavori”, alla posa di materiali (“materiale”), e una connessione ancor più marcata con l’area (nuvola gialla centrale) della “compravendita” (“venditore”, “acquirente”, “vendita”, “compravendita”), in forte correlazione con i più rilevanti beni che ne sono oggetto: veicoli (“veicolo”) e immobili (“immobile”, “casa”). Dalla vendita al “prezzo” (nuvola azzurra), con il verbo “richiedere” che fornisce, nella rete di questo grafico, un *trait d’union* con i termini “risoluzione”, “sostituzione”, “riparazione”, “intervento”, pregnanti nella classe. Infine, dal prezzo si passa anche al “servizio” (nuvola verde in alto), quindi a temi che vanno oltre la compravendita di beni e l’esecuzione di opere e lavori incontrati sin qui, con riferimenti alle forniture elettriche ed energetiche (“fornitura”, “elettrico”, “energia”), ma anche all’installazione ed esecuzione di impianti nuovi (“impianto”, “nuovo”) o alla loro “manutenzione”.

Fig. 6 La rete di parole del tema “vizio” (classe 6).



Guardando alle connessioni tra le parole è possibile scandagliare ulteriormente il tema.

In conclusione, l'analisi del *network* delle parole ci ha permesso di scandagliare ulteriormente la classe 6, ed evidenziando le connessioni tra le parole che ne compongono il vocabolario ha consentito l'affiorare in modo più chiaro dei profili di una classe composita, tra i quali quello, non centrale nella classe forse, ma significativo a livello sostanziale, dell'accertamento giudiziale dei fatti controversi, anche tramite consulenze tecniche d'ufficio: un'ulteriore conferma della particolare rilevanza e della stretta interdipendenza, all'interno della macro-categoria "fatti", delle questioni, "trasversali" a circostanze di fatto di diversa natura, relative a vizio, danno, prova e documentazione.

L'evoluzione nel tempo dei temi presenti nelle pronunce

Messo a fuoco il contenuto delle classi, il corpus è stato sottoposto ad un'analisi del χ^2, per individuare il livello di associazione tra le singole classi alla base del corpus e i periodi temporali in cui si collocano le pronunce[8], raggruppati in quinquenni (individuati dalla variabile 'Periodo2')[9]. Si veda a tal proposito la Fig. 7, che mostra i risultati dell'analisi eseguita con il software Iramuteq.

Non sorprende l'assenza di associazioni temporali significative per la classe 1 "giudizio", trasversale a tutte le pronunce. Anche le classi 2, 3 e 7 ("prova", "documentazione", "danno"), ascrivibili alla macrocategoria "fatti", non sembrano presentare significative variabilità nel tempo.

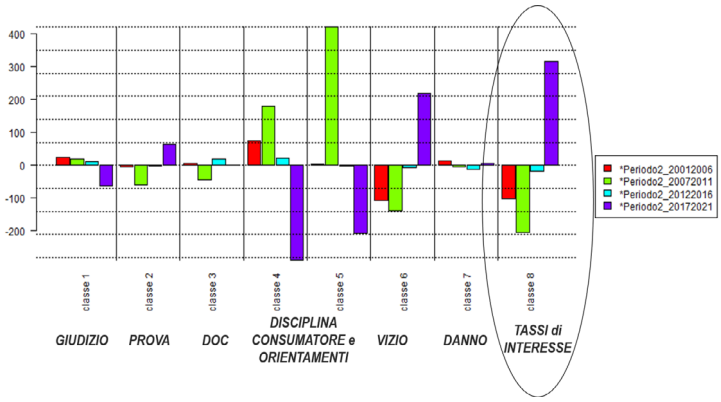
Quanto alle classi 4, 5, 6 e 8, caratterizzate dai picchi più evidenti, in generale non sembra possibile formulare un'ipotesi che li ponga in rela-

[8] Chi-quadrato (χ^2 o chi^2, appunto) è una misura statistica dell'associazione tra due o più caratteri statistici: valori positivi indicano correlazioni positive (o dirette), valori negativi indicano correlazioni negative (o inverse), valori pari o vicini allo zero non supportano un'associazione significativa. Se la presenza di quella classe in quel periodo è in linea con le attese, ovvero in linea con la presenza media della classe nel corpus, il grafico restituirà, per quello stesso periodo, una colonna di altezza prossima allo zero; in caso di classe sovrarappresentata o sottorappresentata nel periodo, avremo nel grafico una colonna corrispondente a un valore, positivo o negativo, indicativo dello scarto rispetto alle attese.

[9] La variabile 'Periodo2' fa riferimento al periodo, corrispondente di massima a un quinquennio, in cui si colloca la data di pubblicazione di ciascuna pronuncia; nello specifico sono stati individuati il periodo 2001-2006 e i quinquenni successivi, l'ultimo dei quali è il quinquennio 2017-2021, pur se 'incompleto' dal momento che le pronunce più recenti sono datate 31 marzo 2021.

zione con specifici eventi caratterizzanti i quinquenni considerati, quantomeno allo stadio attuale dell'analisi, anche visto il carattere composito ed eterogeneo delle classi individuate, che non permette di fare riferimento, per ciascuna di esse, a singole questioni puntualmente individuate.

Fig. 7 I temi nel tempo.



La parte del grafico relativa alla classe 8 "tassi di interesse" sembra supportare l'ipotesi di un aumento, negli anni più recenti, della rilevanza delle questioni in tema di tassi di interesse, e più in generale della tematica finanziaria, nelle controversie del tipo di quelle in esame. Tale andamento potrebbe essere correlato al susseguirsi, con cadenze sempre più ravvicinate negli anni più recenti, delle crisi finanziarie.

Unica eccezione, in questo senso, potrebbe essere rappresentata dalla classe 8 "tassi di interesse", più specificamente riferita ad un tema ben definito come quello, appunto, dei tassi di interesse e della materia finanziaria. I picchi riferiti a tale classe, dunque, sono gli unici rispetto ai quali si potrebbe ipotizzare, seppure con grande cautela, il segnale di una più generale variazione nel tempo della rilevanza di una certa tematica. In particolare, l'ipotesi è che dai picchi presenti in questa classe, che mostrano un andamento in crescita nel tempo, si possa dedurre un aumento, negli anni più recenti, della rilevanza delle questioni in tema di tassi di interesse, e più in generale della tematica finanziaria, nelle controversie del tipo di quelle in esame, magari in seguito alle crisi finanziarie che si susseguono con cadenze sempre più ravvicinate.

Per provare a formulare qualche ulteriore ipotesi relativamente alle altre classi, occorrerebbe inoltrarsi più nello specifico nella composizione di ognuna di esse, con un'opera di approfondimento che trascende l'eco-

nomia del presente lavoro. Ad ogni modo, l'analisi presentata ha mostrato, assieme ad un primo parziale risultato relativo alla tematica finanziaria, il rilevante interesse della possibilità offerta dal software di effettuare analisi di tipo comparativo all'interno di un corpus.

In prospettiva, poi, tale possibilità potrebbe essere sfruttata ad ampio spettro, ad esempio guardando in modo ancor più mirato all'evoluzione nel tempo della rilevanza di varie questioni, analogamente a quanto qui fatto per temi generali, o anche tentando comparazioni su base geografica, ad esempio tra regioni.

Ancora, analisi di questo tipo potrebbero accrescere via via il proprio interesse, in corrispondenza di eventuali arricchimenti e/o aggiornamenti nel tempo del corpus con pronunce di nuova emissione. L'aggiornamento periodico del corpus potrebbe consentire di disporre di analisi sempre più affinate, meglio fondate e aggiornate, rafforzando l'auspicio che lo strumento proposto possa realmente arrivare a costituire, nel tempo, un pilastro informativo ulteriore a supporto delle scelte regionali in tema di tutela dei consumatori.

Considerazioni conclusive

Nel disegno di questo lavoro si è partiti dalla constatazione della necessità di ampliare la base informativa a supporto delle strategie regionali a tutela dei consumatori, per cui, a fronte di politiche regionali in materia di tutela dei consumatori che intendono rivolgersi alla generalità dei consumatori, il corredo informativo che attualmente le supporta è riferito principalmente alla platea, significativa ma pur sempre parziale, di coloro che si rivolgono alle associazioni dei consumatori.

Di qui la scelta di volgere lo sguardo verso coloro che si trovano a difendere i propri interessi di consumatori di fronte a un giudice, a partire dalla ricerca delle circostanze che portano i consumatori ad intraprendere una simile azione per la tutela dei propri interessi o a essere chiamati in giudizio, e di quali questioni si trovano ad affrontare.

Partendo da un nucleo di oltre mille pronunce, tratte da una banca dati già in uso all'Amministrazione, e senza preventivamente conoscerne il contenuto, è stato possibile, attraverso tecniche di analisi testuale automatica, addentrarsi in misura significativa nel loro contenuto, individuando numerose tematiche, talune delle quali trattate con maggiore frequenza e/o rivelatrici di caratteristiche interessanti e inattese delle controversie in esame, suscettibili di suggerire altrettanti spunti che po-

trebbero rivelarsi utili a supportare le decisioni di policy della Regione in materia.

Tra queste, è emersa con chiarezza la tematica finanziaria. L'analisi sembra suggerire che questa sia una delle aree in cui sussiste un particolare bisogno di tutela dei consumatori, cioè delle aree in cui l'azione regionale a tutela del consumatore possa essere utilmente rafforzata, magari attraverso specifiche azioni in tema di educazione finanziaria, al consumatore ma anche all'operatore dello sportello dell'associazione dei consumatori chiamato ad assisterlo. Valgono le medesime considerazioni relativamente alla classe semantica dei tassi di interesse e dei c.d. contratti bancari, e i segnali nel senso di un aumento, negli anni più recenti, della rilevanza di questa tematica, così come al forte peso del tema del decreto ingiuntivo – atto che spesso segue al cattivo esito di contratti di tipo bancario/finanziario – all'interno della classe riferita ai temi processuali/procedurali. L'emergere del tema del decreto ingiuntivo fornisce spunti anche quanto all'importanza dell'informazione e formazione sulla disciplina specifica a tutela del consumatore, posto che talvolta il meccanismo del decreto ingiuntivo si avvia perché il consumatore non ha correttamente attivato la disciplina che lo tutela. Allo stesso modo, la determinazione di classi relative a vizio, danno, prova e documentazione all'interno della macrocategoria riferita alle circostanze di fatto suggerisce la presenza di determinate vicende e circostanze che hanno portato al giudizio, e, di conseguenza, la presenza di specifiche questioni sulle quali è possibile concentrare gli sforzi regionali affinché i cittadini possano essere maggiormente consapevoli dei propri diritti di consumatori e delle corrette modalità di attivazione di tali diritti, quindi più tutelati. Infine, l'analisi ha rilevato la scarsa presenza, ma non l'assenza, nel corpus esaminato, di quelle materie che più sono interessate da meccanismi di risoluzione alternativa delle controversie (c.d. *Alternative Dispute Resolution,* ADR). Ciò sembrerebbe suggerire che questi meccanismi in generale funzionano, ma che, allo stesso tempo, c'è forse un margine di miglioramento, ad esempio nel senso di una più diffusa conoscenza sull'esistenza e sul funzionamento di tali procedure, di cui si potrebbe tenere conto nell'elaborazione delle politiche regionali.

Ulteriori spunti di policy potrebbero emergere procedendo ancor più in profondità nell'esplorazione e interpretazione del corpus e delle singole tematiche, per esempio tramite ulteriori *network analysis* dei contenuti specifici delle classi, o approfondendo la possibilità di effettuare analisi di tipo comparativo all'interno del corpus, guardando ad esempio in modo

ancor più mirato all'evoluzione nel tempo della rilevanza di specifiche questioni, o ancora tentando comparazioni su base geografica, ad esempio tra regioni. Infine, il corpus potrebbe essere arricchito e aggiornato nel tempo con pronunce di nuova emissione, consentendo di disporre di analisi sempre più affinate, meglio fondate e aggiornate, costituendo di fatto un pilastro informativo ulteriore a supporto delle scelte regionali in tema di tutela dei consumatori, e, al contempo, consentendo l'elaborazione di politiche sempre più accompagnate da un corredo di informazioni e dati a supporto.

Più in generale, punto di interesse delle tecniche innovative utilizzate e presentate nel capitolo consiste nella loro agevole trasferibilità anche ad altre banche dati giuridiche o amministrative e ad altri ambiti di conoscenza, consentendo il supporto delle decisioni di policy della Regione anche in altre materie.

I diari del cambiamento. Un'analisi dei diari degli immigrati dell'Archivio Diaristico Nazionale per migliorare le politiche di integrazione regionali

Irene Diaz Mina[1]

Archivio Diaristico Nazionale, Diari degli immigrati, Analisi bisogni, Politiche di integrazione.

Introduzione

Lo studio che sarà presentato nel capitolo prende le mosse a partire dal materiale messo a disposizione nella forma dei diari scritti dagli immigrati dall'Archivio Diaristico Nazionale, che ha sede a Pieve Santo Stefano, in provincia di Arezzo, con l'obiettivo di analizzare e comprendere, a partire dalle parole degli immigrati stessi, le ragioni che conducono le persone a decidere di migrare in Italia, le aspettative che esse nutrono nei confronti di questa scelta e i percorsi di inclusione intrapresi all'arrivo in Italia. I diari utilizzati sono stati redatti nell'ambito del progetto DiMMi (Diari Multimediali Migranti), nato con l'obiettivo di sensibilizzare i cittadini sui temi della pace, della memoria e del dialogo interculturale e con il fine di creare un fondo speciale dei diari migranti. Lo studio esplorativo

[1] Nell'ambito del Master in Innovazione, Progettazione e Valutazione delle Politiche e dei Servizi. Agenda 2030 - PISIA ha svolto un tirocinio presso il Settore Tutela dei Consumatori, Politiche di Genere, Promozione Cultura di Pace della Regione Toscana, per il quale ha realizzato l'analisi del contenuto di alcuni diari multimediali custoditi presso l'Archivio Diaristico Nazionale.

intende fornire elementi conoscitivi per indirizzare le politiche regionali in materia di sensibilizzazione e integrazione degli immigrati e riguarda 36 diari e racconti scritti in prima persona da uomini e donne che hanno maturato un'esperienza di migrazione negli ultimi vent'anni.

Fra gli obiettivi di questo lavoro vi è dunque quello di valorizzare una banca dati costituita dall'Archivio Diaristico Nazionale, che rappresenta una fonte informativa assai importante e che ci ha consentito di effettuare un'indagine quali-quantitativa, per andare oltre i numeri ufficiali della migrazione, scoprendo le storie che si nascondono dietro questi numeri e fornendo una lettura di testimonianze dirette degli immigrati del tutto inedite e fino ad oggi poco utilizzate a fini analitici. Attraverso la lettura e l'analisi testuale dei documenti messi a disposizione, si è tentato dunque di analizzare più a fondo le dinamiche e i meccanismi che hanno caratterizzato il fenomeno migratorio, di prima accoglienza e di inclusione in Italia.

Raccontare il fenomeno migratorio, utilizzando le memorie e le parole delle persone che ne sono state protagoniste, significa anche tentare di superare gli stereotipi che spesso caratterizzano il dibattito pubblico intorno a questo tema. Significa capire quali sono le esperienze che favoriscono o inibiscono l'inclusione e l'accoglienza dal punto di vista di chi l'ha vissuta e la vive quotidianamente.

I diari degli immigrati: fonti di dati per supportare il decision making

Assicurare il raggiungimento dei *target* che riguardano la dignità di ogni migrante, le loro necessità primarie, i diritti familiari, sociali, economici e, soprattutto, il diritto di accesso all'istruzione di qualità, sono probabilmente alcuni tra gli obiettivi di sviluppo sostenibile più ambiziosi dell'Agenda 2030, adottata durante l'Assemblea Generale delle Nazioni Unite del settembre 2015. Questa prospettiva determina, tuttavia, la necessità di realizzare un'analisi approfondita di un fenomeno complesso quale quello dell'immigrazione.

Quest'ultimo, infatti, è il frutto di molteplici fattori di spinta e di attrazione – di natura sociopolitica, demografica, economica ed ambientale – determinando, conseguentemente, la necessità di modelli di intervento politico che considerino la varietà dei bisogni che lo caratterizzano e che sono alla base delle scelte e delle aspettative di migrazione dei singoli immigrati e dei fattori che favoriscono l'integrazione e l'accoglienza.

A partire da tali valutazioni, tramite il presente lavoro abbiamo provato a raggiungere una migliore comprensione del fenomeno migratorio in Italia e a identificare alcuni temi trasversali alle diverse esperienze individuali, che possano supportare l'attività dei policy maker volta all'attuazione di politiche che integrino bisogni e aspettative degli individui immigrati.

Lo studio vuole contribuire alla costruzione di una nuova narrazione che superi gli stereotipi e le semplificazioni che riguardano tale complesso fenomeno e che finiscono per limitare fortemente le opportunità di inclusione offerte alle persone immigrate. Il processo migratorio, infatti, se analizzato dal punto di vista dell'esperienza umana e personale, non termina con l'arrivo in Italia, ma è un processo che va ben oltre e comprende fasi della vita importanti, ma, come si vedrà nel seguito del lavoro, spesso trascurate ed eccessivamente stereotipate dalle politiche, a partire dal tema della formazione (linguistica, scolastica e lavorativa).

La sfida per i policy maker si configura nella capacità di poter intercettare informazioni e dettagli particolari che aiutino a definire il quadro della situazione reale e l'ampiezza del fenomeno, al fine di realizzare politiche sociali maggiormente rispondenti ai bisogni effettivi della popolazione immigrata.

A questo scopo, l'analisi dei testi attraverso la tecnica dell'analisi automatica del contenuto, che consente di gestire un gran numero di dati e informazioni di tipo qualitativo e quantitativo, permette di agire attraverso un nuovo approccio nel disegnare e implementare le politiche pubbliche e garantire una maggiore efficacia dell'operato delle amministrazioni.

L'Archivio Diaristico Nazionale

Nel 1984 a Pieve Santo Stefano, in provincia di Arezzo, nacque nella sede del municipio un archivio pubblico nazionale, che da allora raccoglie e custodisce memorie scritte di vita e di storia di gente comune. L'Archivio è riconosciuto dalla Regione Toscana come una delle grandi istituzioni di eccellenza del territorio toscano. Nel 1991 nacque la Fondazione Archivio Diaristico Nazionale, in seguito divenuta Onlus riconosciuta con Decreto Ministeriale, che nel 2009 ricevette il Codice dei Beni Culturali dello Stato.

L'attività dell'Archivio Diaristico Nazionale è riconosciuta e finanziata da diversi istituti, ditte e benefattori privati. L'istituzione si prefigge non solo di raccogliere documenti, diari e racconti di scrittura popolare,

ma anche di conservare tale banca dati come una fonte preziosa di informazioni. Oggi l'Archivio conserva più di 8.000 testi.

Dal 1998, l'Archivio promuove numerose iniziative, tra le quali il progetto DiMMI di Storie migranti, nato con l'obiettivo di raccogliere e far conoscere le storie di persone che provengono da altri paesi e che risiedono o hanno vissuto in Italia o nella Repubblica di San Marino.

Il presente lavoro prende le mosse a partire dall'analisi di 36 documenti che raccolgono le memorie degli immigrati partecipanti al concorso DiMMi di Storie Migranti dell'anno 2019.

L'immigrazione in Italia: l'evoluzione degli ultimi trent'anni

Una breve analisi dell'evoluzione negli ultimi trent'anni del fenomeno migratorio in Italia consente di focalizzare l'attenzione su certi aspetti che risultano determinanti per comprendere meglio tale fenomeno e i bisogni di tante persone che cercano di migliorare le proprie condizioni di vita in un territorio con cultura, usi e costumi diversi da quelli di appartenenza.

In Italia, l'immigrazione è un fenomeno abbastanza recente. Il flusso migratorio nel territorio italiano ha infatti iniziato ad aumentare a partire dagli anni Settanta, diventando tuttavia oggi una realtà consolidata, come dimostrato dai dati Istat. I censimenti[2] mostrano, infatti, un continuo aumento della presenza di stranieri nel territorio italiano, passando da 298.749 stranieri residenti nel 1980 a 649.000 nel 1991, e una tendenza di crescita continua negli ultimi vent'anni (si veda la Fig. 1).

La crescita costante dei flussi immigratori nel territorio nazionale caratterizzante il primo decennio degli anni duemila ha successivamente iniziato a rallentare.[2] Il 2018 ha registrato 5.255.503 di stranieri residenti, pari all'8,7% della popolazione totale residente. Nel 2020, l'inasprirsi delle misure di contenimento dei flussi in entrata a causa dello scoppiare della pandemia Covid-19[3] ha confermato la flessione del tasso di crescita immigratorio in Italia. Gli immigrati residenti al 1° gennaio 2020 sono circa 5.306.548 e rappresentano l'8,8% del totale dei residenti nel territorio nazionale, di cui il 58,7% risiede nel Nord, il 25,3% nel Centro e il 16,9% nel Mezzogiorno.

Infine, un dato di interesse ai fini del presente studio riguarda il grado di istruzione della popolazione straniera residente.[4] La percentuale di

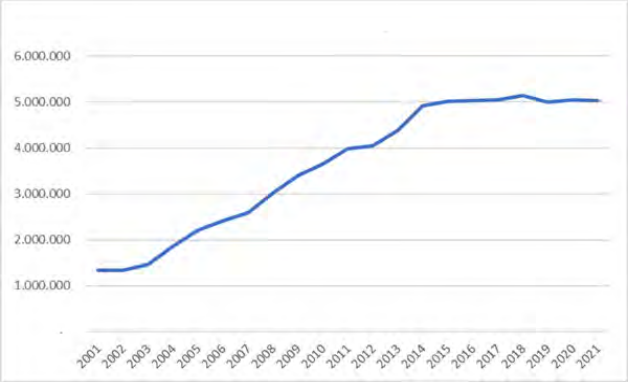
[2] https://www.istat.it/it/files/2019/07/Statistica-report-Bilancio-demografico-2018.pdf.

[3] https://temi.camera.it/leg18/temi/iniziative-per-prevenire-e-contrastare-la-diffusione-del-nuovo-coronavirus.html#iniziative-per-prevenire-e-contrastare-la-diffusione-del-nuovo-coronavirus-.

[4] https://noi-italia.istat.it/pagina.php?L=0&categoria=4&dove=ITALIA.

stranieri residenti in Italia nel 2020 con un titolo di studio è inferiore a quella degli italiani. Il 55% degli stranieri possiede solo la licenza media, contro il 37,5% degli italiani. Gli immigranti residenti con un diploma di scuola superiore rappresentano il 34,5% a fronte del 43,7% di italiani diplomati e il 10,3% ha conseguito una laurea a fronte del 18,7% di italiani laureati.

Fig. 1 Crescita continua degli stranieri residenti in Italia negli ultimi vent'anni.



Indagare una realtà sconosciuta: storie non raccontate, voci poco ascoltate

L'indagine sulle testimonianze e sulle memorie delle persone immigrate raccontate nei diari raccolti dall'Archivio Diaristico Nazionale consente, tramite la lettura di storie spesso non raccontate dai media e l'ascolto di voci scarsamente ascoltate nei processi decisionali ai vari livelli di governo, una migliore conoscenza del fenomeno migratorio e dei bisogni degli immigrati nel loro percorso di integrazione.

Tramite il presente studio si è tentato di rispondere alla domanda di ricerca dando rilievo alle seguenti tematiche:

- identificare alcuni temi di rilievo e portarli in evidenza in modo da favorire una maggiore attenzione verso le problematiche che riguardano direttamente gli immigrati in Italia;
- offrire informazioni utili ai decisori politici per valutare e rivedere alcune politiche e programmi condotti a livello nazionale o locale;
- far emergere una nuova narrazione riguardante il fenomeno migratorio, superare gli stereotipi e contribuire alla maturazione di una nuova consapevolezza intorno al tema delle comunità multiculturali;

- utilizzare i risultati della ricerca per pensare a nuovi piani di formazione che mirino a favorire un'istruzione multiculturale.

Il metodo d'analisi

Lo studio del contenuto gioca un ruolo centrale nel nostro progetto in quanto i diari degli immigrati non sono mai stati analizzati. Nello specifico, l'analisi automatica del contenuto consente un più efficace accesso, rispetto ad un'analisi di tipo solo qualitativo, alle informazioni nascoste o meno visibili contenute nei documenti oggetto di analisi, permettendo di rappresentare tendenze generali comuni alla maggioranza e allo stesso tempo specificità dei dati testuali, che con l'approccio classico manuale non è possibile realizzare senza una forte mediazione interpretativa, che può risultare distorsiva, del ricercatore.

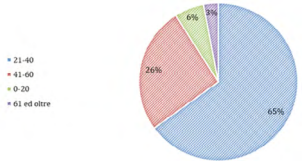
L'analisi del corpus costituito dai diari dei migranti consente infine di ampliare la visione d'insieme sul problema dell'integrazione degli immigrati e di analizzarla dal loro specifico punto di vista.

Elementi socio anagrafici dei soggetti che hanno scritto i diari

I 36 diari utilizzati nello studio sono stati scritti da 17 donne e 19 uomini di età compresa tra i 17 e i 73 anni, distribuiti in quattro fasce di età (Fig. 2).

I testi contengono un numero di pagine che va da 2 a 36, per un totale di 854 pagine, e sono stati scritti o tradotti in lingua italiana. I racconti hanno per lo più mantenuto la loro forma originale, con poche correzioni sotto forma di parentesi quadre o minimi aggiustamenti redazionali. Questo ha comportato, ai fini della presente analisi, la necessità di leggere integralmente i documenti e di apportarvi ulteriori correzioni per facilitare l'analisi attraverso il software, sottoponendo i diari ad una serie di operazioni preliminari di normalizzazione dei dati testuali.

Fig. 2 Distribuzione in fasce d'età degli autori dei diari.



Tra gli autori dei diari scelti per l'analisi (quelli vincitori del concorso DiMMI nell'anno 2019), la maggior parte provengono dall'Africa (21/36 autori), seguita dall'Europa (7/36 autori), dall'America latina (4/36 autori) e dall'Asia (4/36 autori).

Il corpus testuale: i diari degli immigrati

I 36 diari utilizzati ai fini del presente studio sono racconti scritti in prima persona dai vincitori del concorso DiMMI di Storie Migranti dell'anno 2019, distribuiti in tre fascicoli:

- fascicolo 1: *Il Confine tra di Noi*, comprendente i documenti dall'1 al 16;
- fascicolo 2: *Se il mare finisce*, contenente i diari dal 17 al 26;
- fascicolo 3: *Parole oltre le frontiere*, formato dai documenti dal 27 al 36.

Sulla base degli obiettivi che guidano la ricerca, sono state identificate alcune variabili e relative modalità (Tab. 1) al fine di operare un'ulteriore classificazione dei testi, che, come si vedrà nei prossimi paragrafi, è stata utilizzata per eseguire alcune analisi specifiche.

Tab. 1 Variabili e modalità socio-anagrafiche per la classificazione dei diari.

	Variabili						
	Genere	**Continente di provenienza**	**Età**	**Motivo del Viaggio**	**Status giuridico in Italia**	**Titolo di studio**	**Aspettative**
Modalità	Maschio Femmina	Africa America Asia Europa	17-29 anni 30-37 anni 38-73 anni	Albina Altro Amore Giustizia Guerre Lavoro Libertà Omofobia Salute Studio Turismo	Altro Attesa Residente Rifugiato Cittadinanza	Altro Scuola media Scuola superiore Laurea Master Dottorato	Altro Diritti Studio Sensibilizzare

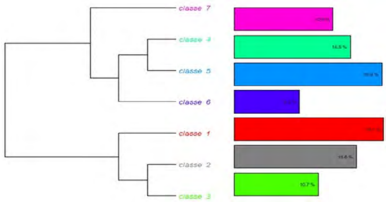
Analisi testuale del contenuto dei diari

Di fronte alla necessità di estrarre conoscenza da dati testuali destrutturati, quali sono i diari degli immigrati, per l'analisi automatica del contenuto dei diari è stata innanzitutto operata una *topic detection* del corpus tramite il metodo Reinert, realizzata utilizzando il software Iramuteq, che ha consentito di delineare 'mondi lessicali', ovvero classi semantiche che

vedono al loro interno espressioni tipiche che ricorrono con maggior frequenza nelle porzioni di testo ricomprese nei singoli *cluster*.

Il software ha individuato la presenza nel corpus di 7 classi semantiche, rappresentate nella Fig. 3.

Fig. 3 I contenuti prevalenti nei diari.



Argomenti dei diari: il viaggio, il *background* e l'inserimento

Attraverso la lettura delle parole maggiormente caratterizzanti le singole classi, di cui si riporta un estratto in Fig. 4 (risultati della *topic detection* rappresentata tramite dendrogramma, dove la grandezza delle parole è proporzionale alla loro significatività nelle singole classi), le 7 classi semantiche individuate dal software possono così riassumersi:

classe 1: l'argomento rilevante è il viaggio e la parola più ricorrente è "autista", con una frequenza nel *cluster* pari a 109 contro 133 nell'intero corpus;

classe 2: l'argomento rilevante è la sofferenza durante il viaggio e la parola chiave più ricorrente è "piangere", con una frequenza di 68 nel *cluster* e 104 nel corpus.

classe 3: l'esperienza del viaggio è l'argomento rilevante e la parola "mangiare" è la più frequente tra le parole, con una frequenza nel *cluster* pari a 88 contro 190 nel corpus.

classe 4: la famiglia è l'argomento che identifica il *cluster* e la parola maggiormente presente è "padre", con una frequenza di 139 nel *cluster* e 226 nel corpus.

classe 5: sono presenti due argomenti prevalenti, le motivazioni del viaggio e le aspettative degli immigrati, e "vita" è una parola chiave di forte rilevanza con 221 frequenze nel *cluster* e 457 nel corpus.

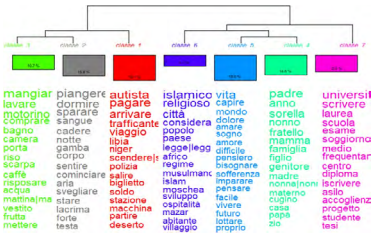
classe 6: anche in questo *cluster* sono presenti due argomenti prevalenti, continente di provenienza e religione, dove la parola "islamico" presenta una frequenza di 22 nel *cluster* e 23 nel corpus.

classe 7: l'argomento prevalente riguarda le aspettative degli immigrati. La parola maggiormente ricorrente è "università" con 48 frequenze nel *cluster* e 60 nel corpus.

Le 7 classi appena illustrate, e quindi gli argomenti in esse prevalenti, possono essere sintetizzati in tre macro-tematiche rilevanti, quali:

- il viaggio. Comprende le classi 1, 2 e 3, che fanno riferimento ad un grande nucleo semantico: viaggio/rotta, sofferenze ed esperienze durante il viaggio d'arrivo;
- *background* personale. Include le classi 4, 5 e 6, che riguardano un nucleo semantico diviso in quattro sotto-argomenti afferenti ai motivi e alle aspettative alla base della decisione di immigrare, alla famiglia, al proprio Paese e alla propria religione;
- l'inserimento e la formazione. Comprende la classe 7, che riguarda in prevalenza le aspettative degli immigrati e incide sulla fase di inserimento e in generale sul rapporto del migrante con i luoghi di integrazione dell'Università, della scuola e della formazione.

Fig. 4 I contenuti dei diari rappresentati tramite dendrogramma.



Analisi dei contenuti dei diari in relazione alle variabili socio-anagrafiche

Al fine di realizzare un'indagine più approfondita dei testi, si è cercato, tramite il software Iramuteq, di identificare la relazione fra le sette classi semantiche individuate e le variabili socio-anagrafiche degli autori dei diari scelte per classificare il corpus testuale (cfr. Tab. 1). In questo paragrafo sono pertanto riportate le analisi del rapporto tra le classi e le variabili, tramite la misura del valore χ^2 di associazione tra le stesse. Si ricorda che le variabili prese in considerazione al fine del presente studio sono: genere, età, continente di provenienza, motivo del viaggio, *status* giuridico in Italia, aspettative delle persone immigrate e titolo di studio. Di seguito sono riportati i risultati dell'analisi.

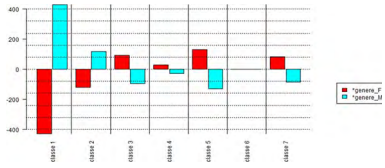
Contenuti dei diari prevalenti secondo la variabile "genere"

Le immigrate (donne) parlano delle motivazioni che hanno determinato l'emigrazione, dell'esperienza del viaggio verso l'Italia, delle aspetta-

tive future, della famiglia e della fase di inserimento nel Paese ospitante.

Gli immigrati (uomini) parlano per lo più del viaggio verso l'Italia, innanzitutto della rotta (l'attraversamento del deserto del Sahara) e la sofferenza affrontata durante il viaggio verso l'Italia (si veda la Fig. 5).

Fig. 5 Distribuzione dei contenuti prevalenti in base al "genere" degli immigrati.



Distribuzione dei contenuti dei diari secondo la variabile "continente di provenienza"

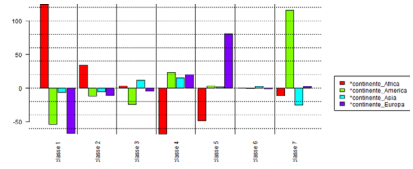
Le persone provenienti dall'Africa parlano prevalentemente della rotta scelta e delle sofferenze vissute durante il viaggio affrontato per arrivare in Italia.

Le persone che provengono dall'America Latina parlano prevalentemente della loro famiglia, delle motivazioni che hanno determinato la decisione di emigrare e le aspettative per il futuro.

Le persone provenienti dall'Asia parlano prevalentemente dell'esperienza del viaggio, della famiglia, del luogo di provenienza e della loro religione.

Per ultimo, le persone provenienti dall'Europa parlano prevalentemente della loro famiglia, delle motivazioni che hanno determinato l'emigrazione e delle aspettative per il futuro (si veda Fig. 6).

Fig. 6 Distribuzione dei contenuti dei diari secondo la variabile "continente di provenienza".



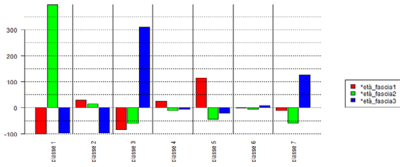
Contenuti dei diari prevalenti secondo la variabile "età"

Gli immigrati nella fascia di età 1 (da 17 a 29 anni) parlano delle sofferenze durante il viaggio verso l'Italia, della famiglia e delle motivazioni che hanno determinato l'emigrazione.

Nella fascia di età 2 (da 30 a 37 anni) si parla per lo più della rotta del viaggio verso l'Italia e delle sofferenze del viaggio.

Nella fascia di età 3 (da 38 a 73 anni) si parla innanzitutto dell'esperienza del viaggio verso l'Italia e delle aspettative verso il futuro (si veda la Fig. 7).

Fig. 7 Distribuzione dei contenuti dei diari secondo la variabile "età".



Contenuti dei diari secondo la variabile "motivazioni dell'emigrazione"

Gli argomenti presenti nella classe 1 (rotta del viaggio scelta) sono trattati da chi ha avuto come motivazione il desiderio di fuggire dalle guerre, di vivere in un contesto dove viene garantita la giustizia e dove sia possibile avere maggiori opportunità di vita e di lavoro.

Gli argomenti della classe 2 (sofferenze durante il viaggio verso l'Italia) sono presenti nei racconti delle persone che hanno sofferto discriminazioni per avere caratteristiche diverse, come il fatto di nascere albina, o per un determinato orientamento sessuale.

Gli argomenti della classe 3 (esperienza del viaggio) sono affrontati dalle persone che hanno avuto come motivazione il desiderio di studiare.

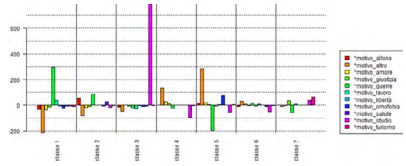
La classe 4 (famiglia) riguarda argomenti trattati da persone motivate ad emigrare da ragioni amorose o per scappare dalle costanti guerre che si vivono nel proprio Paese.

Gli argomenti della classe 5 (motivazioni che hanno determinato l'emigrazione) sono maggiormente frequenti fra le persone che hanno sofferto la discriminazione per orientamento sessuale o caratteristiche fisiche e persone che hanno deciso di trasferirsi in Italia per amore.

La classe 6 (la religione e situazione nel luogo di provenienza) riguarda argomenti trattati da persone motivate dal desiderio di libertà e di fuga dalle guerre.

I discorsi della classe 7 (aspettative per il futuro) riguardano persone che si aspettano di poter studiare, e di vivere in un luogo dove sia garantita la libertà e la giustizia (si veda la Fig. 8).

Fig. 8 Distribuzione dei contenuti dei diari secondo la variabile "motivazione dell'emigrazione".



Distribuzione dei contenuti dei diari in base alla variabile "aspettative"

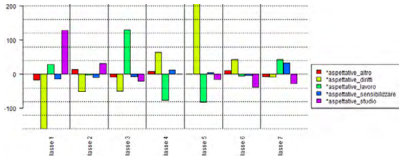
Gli immigrati che hanno come aspettativa il desiderio di avere maggiori diritti parlano della loro famiglia, delle motivazioni dell'emigrazione, della religione che professano e della situazione nel luogo di provenienza.

Le persone immigrate con l'aspettativa di lavorare parlano della rotta scelta per il viaggio d'arrivo in Italia, dell'esperienza del viaggio e delle aspettative future.

Chi ha come aspettativa la sensibilizzazione su certi argomenti che riguardano la propria realtà parla della propria famiglia, delle motivazioni che hanno determinato l'emigrazione e delle aspettative future che spera di poter realizzare nel Paese ospitante.

Gli immigrati che hanno come aspettativa la possibilità di studiare parlano della rotta scelta per il viaggio d'arrivo in Italia e delle sofferenze vissute durante il viaggio (si veda la Fig. 9).

Fig. 9 Distribuzione dei contenuti dei diari in base alla variabile "aspettative".



Distribuzione dei contenuti dei diari per la variabile "titolo di studio"

Gli argomenti che riguardano la rotta scelta per il viaggio sono trattati dagli immigrati che possiedono un diploma di scuola superiore.

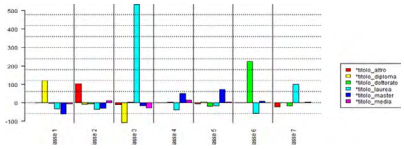
Le persone immigrate con un titolo di scuola media parlano della loro famiglia e delle sofferenze vissute durante il viaggio.

Gli argomenti che riguardano il luogo di provenienza e la religione sono presenti nei racconti degli immigrati che possiedono un dottorato.

Gli immigrati che possiedono un diploma di master parlano della

loro famiglia e delle motivazioni che hanno determinato l'emigrazione (si veda Fig. 10).

Fig. 10 Distribuzione dei contenuti dei diari per la variabile "titolo di studio".



Distribuzione dei contenuti dei diari secondo la variabile "*status* giuridico" in Italia

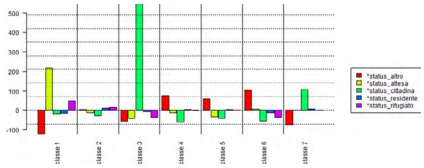
Gli immigrati a cui è stato concesso lo *status* di rifugiato parlano della rotta scelta per il viaggio verso l'Italia e delle sofferenze/difficoltà vissute durante il viaggio.

Gli immigrati che si trovano in attesa di risposta alla domanda d'asilo o di rifugiato trattano gli argomenti presenti nella classe 1 (rotta del viaggio).

Gli immigrati che hanno acquistato la cittadinanza italiana parlano dell'esperienza del viaggio per l'arrivo in Italia e delle aspettative future.

Gli immigrati che risultano regolarmente residenti nello Stato italiano parlano della sofferenza e difficoltà affrontata durante il viaggio verso l'Italia (si veda la Fig. 11).

Fig. 11 Distribuzione dei contenuti dei diari secondo la variabile "*status* giuridico" in Italia.



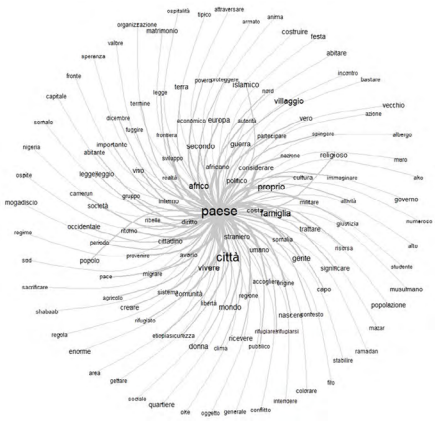
Contenuti dei diari tra passato, presente e futuro

A partire dai termini più ricorrenti, che hanno costituito la base per l'interpretazione dei singoli *cluster*, sono state successivamente realizzate, tramite il software Iramuteq, *network analysis* delle parole del corpus, intendendo in questo modo esplorare i discorsi degli immigrati sulle ragioni dell'immigrazione, sulle aspettative nei confronti di questa scelta e sui percorsi di inclusione intrapresi all'arrivo in Italia.

Il passato. Le ragioni dell'immigrazione

I grafici 12 e 13 permettono di dare uno sguardo al passato degli autori dei testi, consentendo di scoprire e di comprendere gli elementi determinanti che hanno obbligato o influenzato gli autori dei diari a lasciare il proprio Paese. Il *network* rappresentato in Fig. 12 è stato costruito a partire dalla parola chiave "Paese", con il fine di identificare l'interconnessione con le altre parole e rendere più chiaro il quadro della situazione del luogo di provenienza e la rilevanza del tema della famiglia. Si può notare, infatti, che le parole maggiormente interconnesse sono: "Africa", "guerra", "politico", "vivere", "militare", "giustizia", "libertà", "regime", "religione", "famiglia", "musulmano", "islamico" (Fig. 12).

Fig. 12 Uno sguardo al passato: interconnessioni a partire dalla parola "Paese".

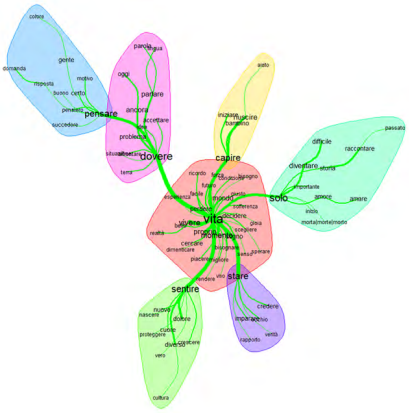


Per la rappresentazione della Fig. 13 si è partiti dalla parola "vita", in quanto l'analisi dei profili ha dimostrato che la stessa ha una frequenza molto significativa nel testo, pari a 457 frequenze nell'intero corpus.

Attraverso l'analisi delle parole maggiormente connesse con la parola vita, si può ipotizzare che le motivazioni del viaggio siano legate alla ricerca di una nuova vita e alla volontà di dimenticare le sofferenze della vita passata.

Infatti, le parole maggiormente ricorrenti sono "vivere", "decidere", "cercare", "gioia", "sofferenza", "cercare", "dimenticare" (Fig. 13).

Fig. 13 Uno sguardo al passato: interconnessioni a partire dalla parola “vita”.



Le aspettative degli immigrati nei confronti della scelta migratoria: indagare il presente

I grafici 14 e 15 forniscono informazioni circa il presente degli autori dei testi. La Fig. 14 è stata elaborata a partire dalla parola “accoglienza”, con lo scopo di identificare l’interconnessione con le altre parole che permettesse di approfondire il tema dell’integrazione. La Fig. 14 mostra, in particolare, la centralità della scuola come risorsa e come vettore di dialogo già dalla fase dell’accoglienza, centralità ulteriormente confermata dalla Fig. 15, realizzata a partire dalla parola “studiare”. Per entrambi i grafici le parole maggiormente ricorrenti sono: “scuola”, “università”, “italiano”, “conoscere”, “lavorare”, “scrivere”, “storia”, “lingua”, “parlare”, “aiutare”, “Italia”, “centro”, “iniziare”, “permettere”.

Fig. 14 Il Presente: la fase di inserimento in Italia. Interconnessioni a partire dalla parola “accoglienza”.

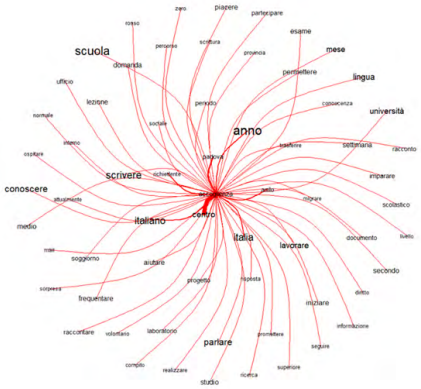
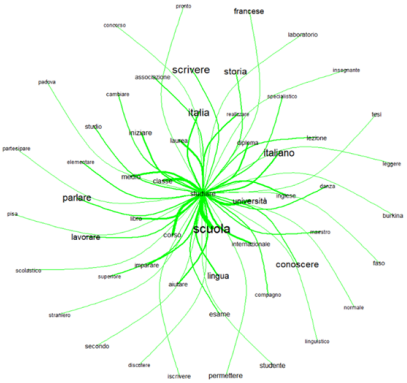


Fig. 15 Il presente: la fase di inserimento in Italia. Interconnessioni a partire dalla parola “studiare”.



I percorsi di inclusione intrapresi all’arrivo in Italia e le prospettive delle persone immigrate: uno sguardo al futuro

I grafici 16 e 17 sono stati elaborati nel tentativo di individuare le interconnessioni con le parole maggiormente ricorrenti al fine di ipotizzare le aspettative, i progetti di vita, le aspirazioni per il futuro degli autori che hanno scritto i diari una volta stabiliti in Italia. A partire, per entrambi i grafici, dalla parola “scuola”, è possibile osservare l’importanza, per gli autori dei diari, delle tematiche della scuola e della formazione. Scuola e formazione che sono legate alla conoscenza della “lingua” e alla fase di “accoglienza”, ma fondamentali anche per assicurare sia l’integrazione sociale (tramite il percorso di studi dell’obbligo), sia lavorativa (tramite il percorso di studi universitario).

Fig. 16 Il Futuro: la scuola come vettore di dialogo. Interconnessioni a partire dalla parola “scuola”.

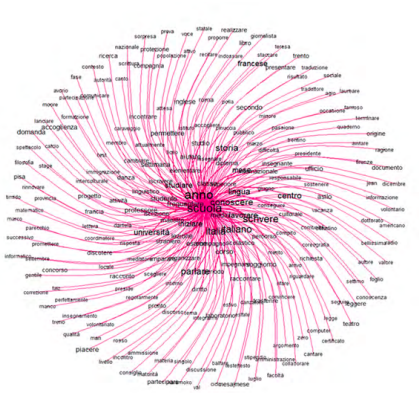
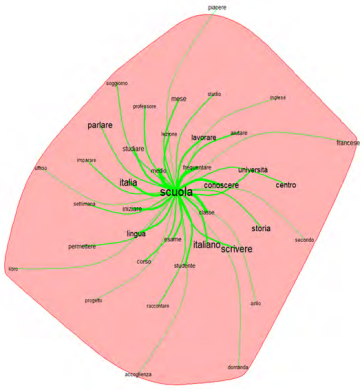


Fig. 17 Il futuro: la scuola come riscatto sociale. Interconnessioni a partire dalla parola “scuola”.



Interviste in profondità per confermare la validità delle evidenze emerse dall’analisi testuale

Ai fini dell’analisi condotta in questo capitolo e a partire dalle evidenze emerse dall’analisi testuale presentata nei precedenti paragrafi, si è ritenuto importante compiere un’indagine tra i diversi attori che quotidianamente affrontano la sfida di mettere in campo e gestire azioni in funzione dell’inclusione degli immigrati in Italia.

Questo ha comportato la necessità di individuare attori che, in vari ambiti territoriali e a diversi livelli istituzionali, siano portatori di idee, progetti e azioni, cercando di capire in quale modo essi operino attraverso politiche e servizi per realizzare una vera inclusione degli immigrati nel territorio italiano.

Sulla base di queste considerazioni, si è cercato di indagare:

- quali sono le principali difficoltà che può incontrare un migrante nell’accesso all’istruzione;
- quali sono i progetti e le politiche pubbliche che cercano di aiutare le persone migranti nel migliorare la loro istruzione, quali sono le caratteristiche di questi progetti, se si tratta di interventi di successo e quali sono le principali criticità che ne caratterizzano l’attuazione;
- quali nuove politiche potrebbero essere introdotte per aiutare le persone migranti a migliorare la loro istruzione;
- in che modo le politiche di riconoscimento della cittadinanza

possono aiutare o rendere più difficile l'inserimento nel sistema sociale di una persona che proviene da un altro Paese.

Sono stati quindi individuati diversi attori che si distinguono per il loro ruolo, relazioni, conoscenza ed esperienza nell'ambito delle politiche per l'immigrazione in Italia, come meglio specificato nell'Appendice 2 al presente volume.

Per la realizzazione dell'indagine è stato scelto il metodo delle interviste a domande aperte, così da poter adattare la griglia di domande in base all'interlocutore e alla sua volontà di fornire risposte più o meno sintetiche.

Sono principalmente tre le informazioni emerse dall'indagine, che confermano i dati emersi dall'analisi testuale:

- ripensare i piani formativi e l'educazione interculturale, che si configurano come principali *driver* di integrazione;
- il problema dell'efficacia dei corsi formativi proposti per l'inserimento sociale e lavorativo;
- la formazione (linguistica, scolastica e lavorativa) alla cittadinanza per una società multiculturale.

Dall'analisi testuale e dall'indagine condotta sul campo tramite le interviste emerge dunque la necessità di realizzare politiche e progetti che possano andare incontro all'esigenza di garantire un percorso di inclusione degli immigrati più efficace e sostenibile nel tempo. A tal proposito, un'analisi dei diari come quella eseguita in questo studio potrebbe divenire uno strumento di policy volto a realizzare politiche che rappresentino al meglio i bisogni e le aspettative della popolazione immigrata, e le esperienze di vita raccontate nei diari potrebbero essere uno strumento educativo da utilizzare nelle scuole di ogni ordine e grado, per favorire una migliore sensibilizzazione al fenomeno migratorio.

Conclusioni: lezioni apprese dai diari per le policy d'integrazione

Gli obiettivi conoscitivi del presente studio hanno riguardato la comprensione delle ragioni che conducono le persone a decidere di immigrare in Italia, delle aspettative che esse nutrono nei confronti di questa scelta e dei percorsi di inclusione intrapresi all'arrivo nel territorio nazionale. Il fine è duplice: da un lato raggiungere una migliore conoscenza del fenomeno migratorio, dall'altro indirizzare le politiche in materia di integrazione degli immigrati.

I 36 diari analizzati nel capitolo, vincitori del concorso DiMMI di Storie Migranti nell'anno 2019, si sono rivelati una fonte di dati e informazioni molto articolata, in grado di fornire conoscenza delle ragioni e delle aspettative alla base delle scelte migratorie e dei percorsi di inclusione. L'analisi automatica del contenuto dei diari ha restituito un'immagine di persone immigrate che, in un certo momento della loro vita, sono state obbligate a compiere una scelta difficile, talvolta obbligata, che in molti casi si è rivelata tragica. Una scelta tra un'esistenza caratterizzata da situazioni di sofferenza estrema, prodotte da ingiustizie, discriminazioni, violenze e abusi, e una remota prospettiva di rinascita e riscatto sociale che passa principalmente attraverso la formazione e la scuola. Per questo, si può ipotizzare che le persone che hanno affrontato questa esperienza siano consapevoli che, per garantirsi una migliore integrazione e concretizzare il proprio progetto di felicità e realizzazione, debbano innanzitutto investire nella loro formazione (linguistica, scolastica e lavorativa) a vari livelli.

Relativamente al tema dell'inserimento, le parole emerse dall'analisi automatica in quanto più spesso ricorrenti riguardano infatti la scuola, la formazione e la necessità di saper usare correttamente la lingua del Paese ospitante. Gli immigrati di lungo periodo non inseguono un sogno di benessere materiale, di predazione di territori e risorse che appartengono ad altri, ma perseguono un progetto di emancipazione dai bisogni essenziali e di inclusione in una società che li accetti e ne valorizzi le esistenze.

Come prescritto anche dall'Obiettivo 4 'Istruzione di qualità' dell'Agenda 2030, emerge dunque l'esigenza per gli immigrati di poter accedere ad un'istruzione di qualità, che permetta loro sia di difendere i propri diritti, sia di contribuire alla costruzione di una società più equa e più sostenibile. La scuola si configura come una risorsa e un vettore di dialogo già dalla prima fase dell'accoglienza, che costituisce l'inizio del processo di integrazione. Un investimento che nasce anche dalla voglia di contribuire alla costruzione di una società più giusta nel proprio Paese d'origine. Ma alcuni aspetti legati all'insegnamento della lingua e al rapporto lingua/società/lavoro possono essere certamente migliorati tenendo conto dei bisogni.

Come emerso dalle interviste realizzate, i percorsi formativi sono infatti costruiti in un'ottica unilaterale e in numerosi casi in ottica emergenziale, non rispondendo quindi ai bisogni reali degli immigrati, ma anzi favorendo la nascita di un disinteresse a frequentare i corsi formativi. Si evince pertanto la necessità di offrire agli immigrati occasioni di appren-

dimento e socializzazione già dalle primissime fasi di inserimento, a partire dallo studio della lingua del Paese di arrivo, differenziando le modalità di insegnamento a seconda delle differenti classi di età.

Lo studio suggerisce che una buona strategia potrebbe essere quella di offrire percorsi di inclusione e formazione in un'ottica integrata, vale a dire promuovendo corsi di insegnamento della lingua italiana differenziati per vari livelli di integrazione e che siano più agganciati a percorsi specifici di inserimento sociale e lavorativo.

In conclusione, *è necessario* pensare agli immigrati non come un gruppo omogeneo ma composito, nei bisogni e nelle risposte, un gruppo attivo e interattivo con società e mondo del lavoro, in una parola una risorsa o meglio un'opportunità da cogliere, che richiede politiche di inclusione sempre più mirate e organizzate. Le politiche nazionali, ma anche quelle regionali e locali, al fine di promuovere il cambiamento dovrebbero pertanto ascoltare le voci dei cittadini immigrati. Il metodo di studio presentato nel capitolo ha dimostrato che questo è possibile a costi molto bassi, favorendo l'emergere dei bisogni, delle aspettative, e la ricchezza di professionalità che sta dietro questo mondo.

Valorizzazione del patrimonio informativo digitalizzato in Regione Toscana: dalla pianificazione energetica regionale all'analisi delle determine dirigenziali e della comunicazione social

Luca Cipriani[1]

Piani energetici, Decreti dirigenziali, Social mining, Topic modeling, Tassonomie.

Introduzione

In Regione Toscana c'è da tempo attenzione alla sperimentazione e alla valorizzazione di tecniche e strumenti utili ad elaborare documenti digitali costituiti da file testuali, tenuto conto sia della vastità e della potenzialità informativa di questo patrimonio (si pensi alla quantità di documenti, messaggi, *abstract*, ecc., che quotidianamente si generano all'interno di un'amministrazione pubblica), sia della progressiva disponibilità di tecnologie *big data* e della possibilità offerta da queste ultime di immagazzinare dati e documenti con strutture e contenuti anche molto variabili, estraendo conoscenza attraverso l'applicazione di metodi statistico-computazionali.

[1] Si occupa di elaborazione ed analisi dati presso Regione Toscana, dove è responsabile dello sviluppo di sistemi finalizzati alla condivisione e accesso dell'informazione. Ha collaborato e ringrazia per la stesura di questo capitolo molte persone da cui è stato positivamente contaminato, ma in particolare i colleghi Francesca Dodero e Luca Bonuccelli, con i quali ha condiviso una parte significativa delle esperienze riportate di seguito.

Dalle esperienze fino ad oggi condotte, questo tipo di tecniche sembra garantire una velocità e una sistematicità delle operazioni di ricerca, spoglio e sintesi delle informazioni, offrendo a monte utili spunti per approfondire e orientare le analisi e a valle uno strumento per verificare e applicare su larga scala le intuizioni che possono emergere da un primo esame diretto di un insieme di documenti normalmente più ristretto.

Il presupposto che spiega l'adozione di questi metodi è che i significati e le strutture latenti che emergono dai testi siano (almeno in parte) correlati a ranghi, frequenze, probabilità con le quali le singole parole, porzioni di parole (c.d. n-grammi) o insiemi di parole che sono tra loro in stretta relazione (si pensi ad esempio all'espressione 'risparmio energetico') si distribuiscono all'interno di un testo.

Si è anche compreso che non esiste in questo ambito una tecnica che sia perfetta o applicabile in ogni circostanza, e che la sua utilità nel sostenere il processo decisionale e predisporre servizi innovativi sia fortemente condizionata dai seguenti presupposti:

- che si identifichino e si rendano espliciti i limiti dei metodi e degli algoritmi di volta in volta impiegati, con riferimento al dominio e all'obiettivo di interesse;
- che si adottino, per ciascun contesto, i metodi più adatti alle proprie finalità informative e meglio calibrati rispetto ai dati disponibili;
- che si combinino i risultati ottenuti da metodi diversi e idealmente complementari, per derivare conclusioni quanto più possibile complete e robuste;
- che si accompagni l'elaborazione automatica e massiva dei testi con un processo di valutazione critica e di validazione diretta dei risultati ottenuti da parte del ricercatore.

Nel presente capitolo sono presentate alcune sperimentazioni realizzate nell'ambito di due distinte collaborazioni che Regione Toscana ha attivato rispettivamente con il Team per la Trasformazione Digitale e con il Dipartimento di Scienze Politiche, Giuridiche e Internazionali (SPGI) dell'Università di Padova, allo scopo di applicare le tecniche e acquisire conoscenza, misurare i risultati sul campo e valutare la loro adozione in futuro, sia attraverso sviluppi condotti con personale interno, sia acquisendo soluzioni in riuso da altre amministrazioni oppure disponibili sul mercato.

Nell'ambito della prima collaborazione, avviata tra Regione Toscana e il Team per la Trasformazione Digitale e prevista dalla delibera n. 1302 del

27 novembre 2017 e dal successivo protocollo di intesa tra Regione Toscana e Commissario Straordinario del Governo per l'Attuazione dell'Agenda Digitale, sono stati sperimentati i primi modelli di *machine learning*, con i seguenti obiettivi:

- classificare un certo insieme di documenti in categorie predefinite o identificare le strutture organizzative competenti per la loro ricezione (*classification*);
- analizzare il sentimento desumibile dai canali social in un dato intervallo temporale o attorno ad una certa tematica di interesse (*sentiment analysis*).

La collaborazione scientifica e metodologica tra Regione Toscana e il Dipartimento SPGI dell'Università di Padova (avviata nell'ambito del Master in Innovazione, Progettazione e Valutazione delle Politiche e dei Servizi. Agenda 2030 – PISIA) ha avuto invece ad oggetto la sperimentazione – mediante uso di tecniche di *text mining* – di analisi comparative tra documenti di programmazione e tra documenti amministrativi, con l'obiettivo di fornire una rappresentazione complessiva, sia attuale che storicizzata, della produzione regionale (anche in chiave comparata) di indirizzi, politiche e azioni direttamente o indirettamente riconducibili alle finalità di Agenda 2030. La sperimentazione ha ricompreso in particolare:

- la costruzione di una metodologia di *clustering* a supporto o in sostituzione dei sistemi più tradizionali di classificazione, questi ultimi finalizzati ad un monitoraggio di natura deterministica e sovente contraddistinti da un alto impatto organizzativo;
- l'individuazione di una metodologia di ricostruzione di griglie concettuali e ontologie a partire dai fatti e dalle loro modalità di manifestarsi, proponibili anche in termini di confronto e validazione rispetto a modalità di classificazione degli stessi fatti più tipicamente di natura *top-down*.

I corpora dei Decreti Dirigenziali, dei Piani energetici e della comunicazione social

I casi di studio che saranno riportati nel seguito del capitolo riguardano tipologie di corpora e metodologie di elaborazione differenti, poiché l'obiettivo non è stato quello di rispondere ad un caso d'uso specifico e concreto, ma piuttosto di sperimentare queste tecniche ad ampio raggio,

su tipologie e volumi di dati diversi. I principali risultati perseguiti e le principali tecniche sperimentate riguardano, in sintesi:

- l'individuazione delle strutture organizzative competenti, a livello di Direzione regionale, cui ricondurre i Decreti Dirigenziali che sono stati emessi da Regione Toscana nel corso degli anni, attraverso l'applicazione di tecniche di *machine learning* supervisionato facenti uso di reti neurali;
- l'analisi comparata dei Piani energetici delle regioni italiane e di quello nazionale rispetto al lessico utilizzato, attraverso il calcolo e il confronto di appositi indicatori statistici che hanno consentito di raggruppare i documenti disponibili in base a una misura di 'distanza' tra i testi analizzati;
- la classificazione dei documenti contenuti nei due corpora suddetti – quello costituito dai Decreti emessi da Regione Toscana e quello dei Piani energetici regionali e nazionale – rispetto alla trattazione di determinati argomenti, ricercati attraverso vocabolari controllati costruiti a priori sul tema della lotta al cambiamento climatico;
- la ricerca, nei medesimi due corpora e secondo un approccio inverso al precedente, dei principali argomenti trattati, facendo emergere questi ultimi direttamente dai contenuti dei testi attraverso l'uso di un metodo non supervisionato per il *topic modeling* chiamato *latent dirichlet allocation* (LDA);
- infine, l'analisi di un ulteriore e distinto corpus, costituito da messaggi acquisiti dal social network Twitter, sia con l'obiettivo di raggrupparli secondo tematiche omogenee, sia con lo scopo di desumere il livello di polarizzazione e di '*sentiment*'.

Le suddette applicazioni hanno quindi interessato complessivamente tre corpora (anche se, come si vedrà, per l'analisi dei Decreti sono stati utilizzati nelle diverse elaborazioni due corpora non pienamente corrispondenti), costituiti da tre tipologie di documenti ben distinte in termini di volumi (quantità di documenti ricompresi nel corpus), contenuti (argomenti trattati nel corpus e livello di omogeneità tra i vari documenti presenti) e struttura (lunghezza del testo e complessità linguistica).

Il primo corpus è costituito dall'insieme dei Decreti Dirigenziali consultabili nella banca dati degli atti regionali presente sul sito *web* di Regione Toscana (Fig. 1).

Il secondo corpus è invece costituito da 22 documenti di programma-

zione, nella fattispecie i Piani energetici regionali, nazionale e delle due province autonome, acquisiti dagli appositi siti istituzionali. Nella Fig. 2 sono elencati, per ciascun ente, gli indirizzi web di riferimento dai quali sono stati scaricati i documenti al momento delle elaborazioni. Ciascuna riga della tabella riporta inoltre l'anno di riferimento del Piano energetico.

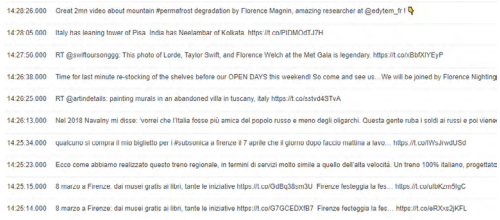
Fig. 1 Database degli atti regionali disponibile all'indirizzo http://www.regione.toscana.it/bancadati/atti/.

Fig. 2 Elenco dei Piani energetici elaborati con il relativo anno di riferimento.

REGIONE	INDIRIZZO	ANNO DI RIFERIMENTO
Abruzzo	https://www.regione.abruzzo.it/content/pianificazione-energetica	2009
Basilicata	https://www.regione.basilicata.it/giunta/site/giunta/department.jsp?dep=100050&area=602412&level=0	2010
Calabria	http://enerweb.casaccia.enea.it/enearegioni/UserFiles/Calabria/calabria.htm	2005
Campania	http://regione.campania.it/regione/it/tematiche/energia-da-fonti-rinnovabili/piano-energetico-ambientale-regionale-pear	2019
Emilia Romagna	http://energia.regione.emilia-romagna.it/piani-programmi-progetti/programmazione-regionale/piano-energetico-regionale	2017
Friuli-Venezia Giulia	http://www.regione.fvg.it/rafvg/cms/RAFVG/ambiente-territorio/energia/FOGLIA111/	2015
Lazio	http://enerweb.casaccia.enea.it/enearegioni/UserFiles/Lazio/PEAR_Lazio_proposta2017.pdf	2015
Liguria	http://www.ambienteinliguria.it/lirgw/sco3/sp/linkPagina.do?canale=Home/010/sviluppoosostenibile/020/energiasost020/pianoenergeticoambientale	2017
Lombardia	https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioRedazionale/istituzione/direzioni-generali/ direzione-generale-ambiente-e-clima/programma-energetico-ambientale-regionale	2015
Marche	https://www.regione.marche.it/Regione-Utile/Energia/Piano-Energetico-Ambientale-Regionale	2005
Molise	http://www3.regione.molise.it/flex/cm/pages/ServeBLOB.php/L/IT/IDPagina/15779	2017
Piemonte	https://www.regione.piemonte.it/web/temi/sviluppo/proposta-definitiva-nuovo-piano-energetico-ambientale-regionale-pear	2018
Puglia	http://www.regione.puglia.it/web/ambiente/documenti	2007
Sardegna	http://enerweb.casaccia.enea.it/enearegioni/UserFiles/Sardegna/PEARS_2015-2030.zip	2016
Sicilia	http://pti.regione.sicilia.it/portal/page/portal/PIR_PORTALE/PIR_LaStrutturaRegionale/PIR_AssEnergia/PIR_DipEnergia/PIR_AreeItematiche/PIR_Attricontinuti/PIR_PianoEnergeticoAmbientaledellaRegioneSiciliana/PEARS	2019
Toscana	http://enerweb.casaccia.enea.it/enearegioni/UserFiles/Toscana/PAER_Disciplinare%20feb2015.pdf	2015
Provincia Bolzano	https://www.gse.it/normativa/scheda-documento?ID_Documento=141	1997
Provincia Trento	http://enerweb.casaccia.enea.it/enearegioni/UserFiles/Trento/PEAP_Trento_2013_2020.pdf	2013
Umbria	https://www.gse.it/normativa/scheda-documento?ID_Documento=901	2004
Valle d'Aosta	https://www.regione.vda.it/energia/pianificazione_energetica_regionale/piano_energetico_ambientale_regionale-pear_i.aspx	2014
Veneto	https://bur.regione.veneto.it/BurvServices/pubblica/SommarioSingoloBur.aspx?num=20&date=21/02/2017	2017
ITALIA	http://www.rinnovabili.it/energia/piano-nazionale-energia-clima/	2018

L'ultimo corpus oggetto di sperimentazione è costituito da 4.479 *tweet* (testi di lunghezza breve), che sono stati acquisiti direttamente dal social network Twitter il 6 marzo 2018, tramite l'utilizzo, come filtri di recupero, degli *hashtag* contenenti il nome delle province toscane e dell'*hashtag* #toscana (Fig. 3).

Fig. 3 Estratto esemplificativo dei *post* recuperati da Twitter.



In seguito all'importazione dei documenti e *post* costitutivi dei tre corpora – un'operazione eseguita manualmente per i Piani energetici e con modalità massive e automatiche negli altri due casi – i testi sono stati estrapolati da ciascun documento in formato *raw* e sottoposti ad una fase preliminare di trattamento e ripulitura, allo scopo di eliminare contenuti ridondanti o difformità sintattiche che, per un verso, avrebbero ridotto la significatività dei risultati e, per altro, appesantito inutilmente le successive elaborazioni. Si ritiene di omettere in questa sede la descrizione di tali operazioni di *pre-processing* e delle specifiche modalità di acquisizione (*ingestion*) dei corpora, in quanto fasi di natura prettamente tecnica e preparatoria a quelle successive di elaborazione, analisi e restituzione dei risultati ottenuti.

Attribuzione dei Decreti Dirigenziali alle strutture organizzative competenti attraverso *machine learning* supervisionato con rete neurale

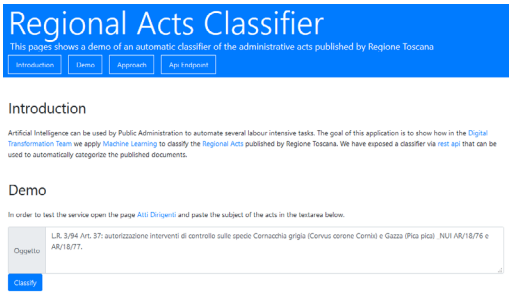
Il primo caso di studio che presentiamo riguarda il corpus dei Decreti Dirigenziali (atti appartenenti alla produzione amministrativa di Regione Toscana), costituito da 152.455 testi. Tale insieme è stato elaborato allo scopo di 'addestrare' un algoritmo che, a fronte dell'oggetto presente nel decreto, fosse in grado di predire la struttura organizzativa (o Direzione regionale) emittente.

Al di là dello specifico contesto sperimentale qui presentato e quindi della sua diretta e immediata utilità, lo studio effettuato ha avuto lo scopo di lavorare su un corpus di documenti pubblici facilmente ottenibili, offrendo la possibilità di trasferire successivamente la metodologia ad altri ambiti, quali ad esempio quello della protocollazione interna dei documenti, suggerendo e agevolando il loro indirizzamento verso le strutture destinatarie, oppure quello della metadatazione di archivi.

Uno dei benefici che è possibile intuire immediatamente è costituito dalla possibilità di automatizzare almeno in parte queste operazioni di indirizzamento e classificazione, limitando l'errore umano e impiegando le risorse per attività più specialistiche, meno burocratiche e che richiedono maggiore creatività e competenza.

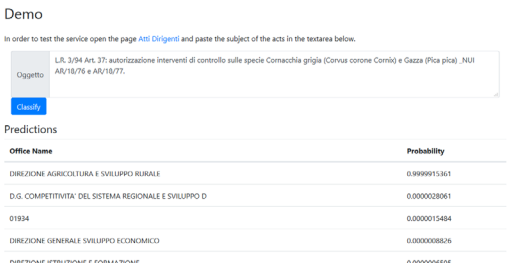
A conclusione della sperimentazione condotta, il risultato dell'attività è consistito nella realizzazione di un 'classificatore' che è stato reso disponibile ad un indirizzo web pubblico, a scopo dimostrativo e per consentirne il riuso in contesti e per necessità analoghe o assimilabili (Fig. 4).

Fig. 4 Classificatore pubblicato all'indirizzo http://ml-api.daf.teamdigitale.it/ (link non più attivo).



L'applicazione, messa a disposizione come demo nel corso dell'anno 2018, ha consentito di testare il servizio di classificazione per struttura organizzativa competente di un qualunque atto dirigenziale recuperabile dalla maschera di ricerca disponibile sul sito web di Regione Toscana (cfr. Fig. 1), semplicemente copiando e incollando il testo contenuto nell'oggetto dell'atto di interesse (Fig. 5).

Fig. 5 Esecuzione del classificatore per uno specifico atto.



Come mostrato nella figura precedente, il risultato della classificazione è un elenco di strutture organizzative (classi) in ordine decrescente rispetto alla probabilità che ciascuna struttura abbia effettivamente emesso l'atto: un valore di probabilità elevato (come si evince in Fig. 5 per la Direzione Agricoltura e Sviluppo Rurale) esprime una sostanziale affidabilità della classificazione, mentre le strutture cui corrispondono valori di probabilità molto bassi sono da escludere come candidate all'emissione dell'atto esaminato.

Oltre all'interfaccia rappresentata in Fig. 5, il servizio è stato reso disponibile anche attraverso un'apposita API[2], utile a trasferire questo

[2] API è l'acronimo di *Application Program Interface*. Si tratta di un servizio che rende

processo di classificazione supervisionata in una componente proceduralizzata e automatizzata all'interno di un sistema informativo (Fig. 6).

Fig. 6 Uso del classificatore via API.

Il caso di studio appena presentato è stato realizzato tramite una serie di passaggi intermedi e tramite l'impiego di molteplici tecniche, fino alla selezione di quella che si è dimostrata maggiormente adeguata al contesto specifico[3].

Il livello di accuratezza del risultato ottenuto può essere rappresentato da una matrice di confusione (Fig. 7), nella quale ogni colonna (j) è riferita ai valori predittivi dati dall'algoritmo di classificazione ed ogni riga (i) rappresenta invece i valori reali. Ciascuna cella (i, j) della matrice contiene quindi il numero di casi in cui l'algoritmo ha classificato la classe reale i come classe prevista j.

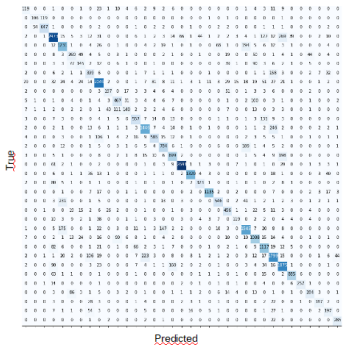
A fronte di un livello complessivo di accuratezza del modello pari all'80%, ottenuto al termine dei diversi passaggi di ottimizzazione dell'algoritmo, sulla diagonale della matrice di confusione si riscontra come

possibile l'interazione tra due applicazioni mediante apposite chiamate.

[3] Si riportano brevemente tali passaggi: il primo step è costituito dall'importazione massiva dei testi dei decreti messi a disposizione da Regione Toscana sul proprio sito istituzionale; successivamente, sono stati creati gli opportuni dataset, necessari alle fasi di addestramento (*training*) e test dell'algoritmo di classificazione; è stato poi condotto l'addestramento di una rete neurale di tipo *fully connected layers*, adottando *rectified linear unit* come funzione di attivazione, con l'obiettivo di valutare in prima approssimazione la possibilità di classificare i documenti basandosi sul potere discriminativo rispetto alle classi dei termini presenti nell'oggetto del decreto. Il modello ottenuto (chiamato 'baseline'), è risultato avere un'accuratezza predittiva del 76%; su tale modello *baseline* sono state applicate tecniche di 'regolarizzazione' (in particolare norma 2 e *dropout*), per migliorare l'accuratezza predittiva e limitare l'*over-fitting* sui dati di *training*. Il modello ottenuto è risultato avere un'accuratezza predittiva dell'80%; si è passati quindi alla valutazione della similarità tra le parole e dei suoi effetti sul miglioramento del potere predittivo della rete neurale (modello con '*embeddings*'). Il risultato ottenuto ha tuttavia mostrato una riduzione di accuratezza predittiva, diminuita al 72.5%; infine, l'utilizzo di 'reti neurali ricorrenti' e di 'reti neurali convoluzionali' per addestrare il classificatore, tramite la rappresentazione di ogni documento come una sequenza invece che come un insieme di parole o vettore sparso (modello basato su sequenze), ha condotto ad un risultato con accuratezza predittiva pari al 74%; in base ai risultati ottenuti negli ultimi tre punti, la scelta finale è ricaduta sul modello con regolarizzazione, che ha mostrato la migliore accuratezza in termini di capacità predittiva.

la capacità predittiva vari a seconda della struttura (con un colore più scuro sono rappresentate le strutture per le quali si registra una migliore predittività), e della numerosità complessiva dei Decreti da questa emessi (la previsione è più accurata se la struttura è caratterizzata da un numero elevato di Decreti).

Fig. 7 Matrice di confusione del modello risultato migliore al termine della sperimentazione.



Occorre tuttavia sottolineare che, trattandosi di un ambito sperimentale, l'algoritmo è stato addestrato, per ridurre i tempi necessari alla sua elaborazione, esclusivamente sui contenuti testuali dell'oggetto dei Decreti, anziché sui documenti nella loro interezza, e che, inoltre, ai fini della classificazione, le Direzioni regionali sono state trattate senza normalizzare i nomi delle classi (talvolta la stessa Direzione poteva infatti comparire con nomi diversi) e senza esclusione delle classi contenenti pochi elementi (l'algoritmo è più difficile da addestrare per quelle Direzioni che hanno emesso pochi atti). Si potrebbe ipotizzare, quindi, un incremento del livello di accuratezza se l'algoritmo fosse addestrato considerando tutto il contenuto del documento e trattando le sole classi (cioè le Direzioni regionali) a cui sono associati un certo numero di documenti, tralasciando quelle caratterizzate da scarsi livelli di emissione di atti.

Per di più, il sistema di classificazione reale potrebbe non seguire necessariamente criteri solo probabilistici: ad esempio, nei casi in cui i documenti presenti all'interno di una certa categoria presentassero una struttura regolare, nota a priori e caratteristica di quella classe, l'approccio probabilistico potrebbe essere accompagnato da un criterio di attribuzione deterministico.

In conclusione, gli aspetti che potrebbero rendere più complesso l'addestramento e l'applicazione dell'algoritmo ad ulteriori scenari reali,

quali quello del protocollo o della metadatazione di archivi, riguardano l'eventuale presenza in quei contesti di una classificazione multi-categoria (nella quale un documento può essere associato a più classi o strutture) e di tipo gerarchico (nella quale le classi possono essere tra loro in relazione 'padre-figlio'). Si tratta di circostanze non ravvisate nel caso dei Decreti e delle relative strutture emittenti presentato in questa sede, dove ogni documento afferisce ad un'unica struttura e le strutture non sono tra loro in relazione gerarchica, trattandosi sempre, nella fattispecie, di sole strutture di massima dimensione.

Ricerca dei principali argomenti trattati dai Piani energetici e dai Decreti Dirigenziali attraverso metodi non supervisionati (*topic modeling*) con algoritmo LDA

Il *topic modeling* è una tecnica di apprendimento non supervisionato che, attraverso la scansione di corpora di documenti, identifica raggruppamenti di parole che appaiono spesso insieme per rappresentare un tema o un argomento coerente.

Questa metodologia può aiutare ad avere una visione sintetica dei temi trattati da un corpus nel suo insieme e può essere applicata a documenti di vario tipo (*abstract* di progetti, atti di programmazione, atti amministrativi, ecc.), consentendo di confrontarli anche a prescindere dalle loro classificazioni originarie.

Rispetto all'utilizzo di tassonomie o vocabolari definiti a priori, che saranno esaminati nel paragrafo successivo, il *topic modeling* è una tecnica che permette di ottenere informazioni sul contenuto effettivo del documento. Esso è inoltre maggiormente scalabile e si presta ad individuare argomenti inter/multidisciplinari, informazioni latenti e temi emergenti non necessariamente ipotizzati prima di avviare l'analisi. Infine, rispetto all'utilizzo di tassonomie, il *topic modeling* si presta meno ad essere utilizzato in ambiti di monitoraggio, poiché i *topics* estratti variano con il corpus analizzato e, quindi, producono nel tempo risultati potenzialmente non confrontabili.

Di contro, il principale svantaggio risiede nella richiesta di un intervento ex post manuale per la verifica, l'interpretazione e la corretta etichettatura dei risultati ottenuti secondo terminologie intellegibili, poiché la tecnica restituisce per ciascun argomento (*topic*) insiemi di parole che il ricercatore è chiamato a interconnettere e a tradurre in significati reali ed espliciti.

Per ultimo, questa tecnica non può garantire che le associazioni tra le parole e gli argomenti estrapolati dall'algoritmo non presentino distorsioni e siano concettualmente e semanticamente correlate in modo corretto tra loro: a seconda della varietà e della complessità del linguaggio che caratterizza i testi elaborati, l'applicazione automatica dell'algoritmo può infatti cogliere 'rumore' o correlazioni bizzarre tra le parole presenti nei testi generando risultati 'sporchi' e non significativi.

Nell'ambito delle sperimentazioni condotte oggetto del capitolo, il *topic modeling* è stato applicato inizialmente al corpus dei 22 Piani energetici, per poi essere scalato su un corpus di 159.863 Decreti Dirigenziali (la cardinalità di questo secondo corpus è diversa rispetto a quello trattato nel paragrafo precedente, in quanto l'acquisizione dei Decreti oggetto di elaborazione e le sperimentazioni stesse sono state condotte in momenti differenti).

L'applicazione a entrambi i corpora del metodo LDA (*latent dirichlet allocation*) ha previsto tre successive iterazioni, rispettivamente di 20, 50 e 100 *topics* (il numero di *topics* ricercati è un parametro di input per l'algoritmo, pertanto la ricerca del numero ottimale può avvenire solo iterando l'elaborazione con numerosità differenti ed osservando i risultati).

Per l'interpretazione dei risultati ottenuti sui due diversi corpora si sono riscontrate complessità differenti, dato il loro diverso livello di omogeneità. Il primo corpus è costituito infatti di soli Piani energetici, quindi da documenti associati ad uno specifico dominio, mentre il secondo corpus comprende tutta la produzione amministrativa regionale, nelle diverse materie di competenza dell'ente.

Topic modeling dei Piani energetici

Nella Fig. 8 sono riportati i primi 10 argomenti in ordine decrescente di importanza rilevati dall'elaborazione del corpus dei Piani energetici impostando il numero di *topics* al valore di 50. Ciascun *topic* è descritto dai primi venti termini più significativi risultanti dall'analisi, che, come anticipato, devono essere sottoposti a interpretazione e classificazione a cura del ricercatore.

Fig. 8 Estratto dei primi dieci *topics* dai documenti di programmazione in ambito energetico.

rk	#	Termini
1	3	consumi produzione impianti settore dati emissioni consumo energetici pari base rinnovabili energetiche elettrici riduzione potenza quota media combustibili obiettivo fonte
2	38	energetico regionale regione energia produzione elettrica rinnovabili fonti territorio finali interventi impianti rete biomasse energetica ambientale dati residenziale intervento bilancio
3	46	piano riferimento risorse valore numero linee previsti piani processo caratteristiche corso attualmente relativa diversi disponibili viene programma pubblici considerazione pianificazione
4	9	sviluppo sistema nazionale obiettivi gestione nazionali principali termine condizioni fini contesto tipo integrazione infrastrutture specifiche guida ministero ambito fase livelli
5	26	settore elettrica emissioni scenario energetica energetico fonti obiettivi trasporti settori crescita trasporto attivitÃ livello tecnologie particolare riscaldamento riduzione periodo promozione
6	17	particolare valutazione possono livello totale punto uso risultati possibile prevede zone partire ridurre diverse tipologie interesse effetti servizio devono individuare
7	6	aree realizzazione calore territorio pubblico decreto lâ€™efficienza definizione diffusione nuove necessitÃ criteri pertanto termini nellâ€™ambito termica allâ€™interno opere relazione lâ€™obiettivo
8	43	energia interventi edifici impianti potenziale sistemi locali soggetti rifiuti costi potenza enti realizzazione azione finanziamento strumenti privati rete modalitÃ qualitÃ
9	7	fonte sistemi energetica impianti edifici calore efficienza utilizzo generazione monitoraggio rinnovabile sviluppo alimentati gas base trasporti pari trasporto superficie rapporto
10	33	energia risparmio fonti lâ€™energia energetica quantitÃ attivitÃ ambientale indicatori interventi stima prodotta province disponibilitÃ rapporto efficienza centrali riguarda caldaie combustibile

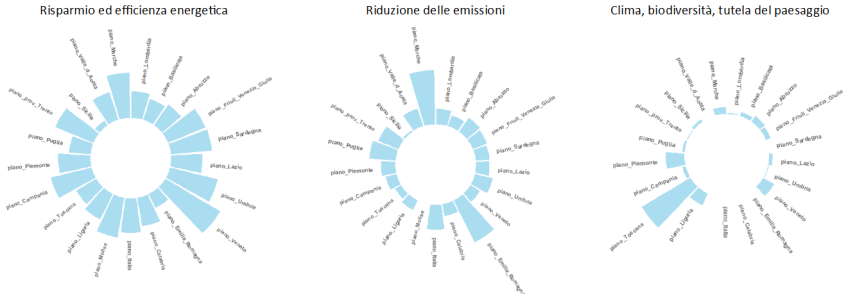
Nella successiva fase di analisi, ciascuno dei 50 *topics* suggeriti dall'algoritmo, fatta eccezione per quelli non significativi, è stato classificato rispetto a una o più dimensioni di riferimento tra le seguenti:

- focus di riferimento del *topic* in ambito energetico (qual è la tematica energetica specifica del *topic*?);
- interventi e soluzioni prospettate per il miglioramento (se il *topic* fa riferimento ad una o più soluzioni migliorative in ambito energetico, quali sono queste soluzioni?);
- strumenti attraverso i quali mettere in campo le azioni migliorative (per attuare l'intervento migliorativo in ambito energetico, quali modalità/strumenti di attuazione sono suggeriti?).

Ad ogni Piano energetico è stato quindi assegnato un livello di prossimità rispetto alle modalità assunte dalle tre dimensioni sopra illustrate e, infine, sono stati rappresentati graficamente i risultati ottenuti.

A titolo esemplificativo, si riportano i grafici relativi alla prima dimensione, quella cioè della focalizzazione del *topic*, per la quale sono state individuate le seguenti tre modalità, ciascuna con diversa rilevanza nei documenti esaminati: risparmio ed efficienza energetica; riduzione delle emissioni; clima, biodiversità, tutela del paesaggio. Si veda la Fig. 9, dove a ciascun documento corrisponde una barra in ciascuno dei tre *bar-plot* circolari.

Fig. 9 *Topic modeling* dei Piani energetici rispetto al focus di riferimento.



Da una prima osservazione dei grafici, si evince che il tema generale del risparmio e dell'efficienza energetica è presente con una intensità rilevante in tutti i Piani energetici, con varie regioni che mostrano un'attenzione al tema anche più accentuata rispetto al Piano nazionale. Sul tema della riduzione delle emissioni il livello di intensità è meno elevato ed omogeneo tra le varie regioni rispetto al caso precedente e particolarmente accentuato nel caso dell'Emilia-Romagna e delle Marche. Infine, la Regione Toscana sembra emergere sui temi connessi con la tutela del clima, del paesaggio e della biodiversità.

Come suggerito da questi primi risultati, l'applicazione del *topic modeling* ai documenti di programmazione ha consentito di effettuare un primo confronto con valore dimostrativo delle politiche regionali in ambito energetico, tramite una comparazione sincronica tra i livelli regionali e tra le regioni e il livello nazionale[4].

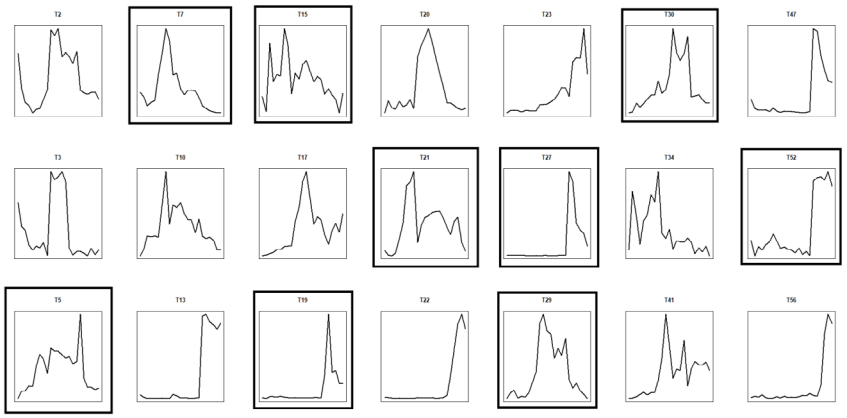
Topic modeling dei Decreti Dirigenziali

Relativamente all'applicazione della stessa metodologia al corpus dei Decreti Dirigenziali della Toscana, l'obiettivo non è quello di pervenire ad un confronto territoriale (la Regione è in questo caso la medesima), ma di ottenere serie storiche che esprimano la rilevanza di ciascun *topic* nell'ambito delle azioni intraprese nel tempo, di cui la produzione amministrativa può considerarsi diretta espressione nel caso di un'amministrazione pubblica.

[4] I piani esaminati non sono tutti riferiti allo stesso anno, ma sono costituiti dagli ultimi atti di programmazione recuperati al momento dell'analisi per ciascuna regione. In questo senso si assume, quindi, sebbene in via approssimata, che il confronto sia avvenuto solo tra aree territoriali.

Nella Fig. 10 sono rappresentati alcuni dei *topics* ottenuti e il relativo andamento nel corso dell'ultimo ventennio (i Decreti Dirigenziali elaborati coprono infatti, nel loro insieme, questo significativo intervallo di tempo).

Fig. 10 *Topic modeling* della produzione amministrativa e relativi trend temporali.

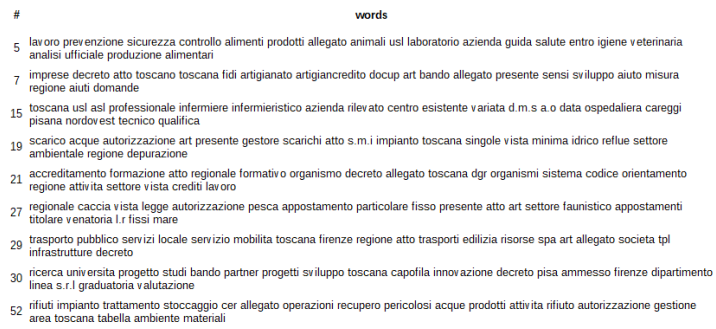


Si riportano di seguito (Fig. 11) i termini principali che consentono di descrivere alcuni dei *topics* precedentemente rappresentati e ritenuti di particolare interesse per l'amministrazione considerata (i quali appaiono, nella Fig. 10, evidenziati da una cornice più spessa).

Da una prima osservazione, si evince l'importanza negli ultimi anni del tema dei rifiuti, del loro trattamento, recupero e stoccaggio, con particolare riferimento alla tutela delle acque (*topic* #52). Nello stesso intervallo di tempo si è registrato un forte incremento dell'attenzione anche sui temi delle autorizzazioni agli scarichi da impianti della depurazione (*topic* #19).

In conclusione, le elaborazioni fin qui mostrate e i risultati ottenuti forniscono un primo supporto ad una migliore comprensione delle materie oggetto dell'amministrazione regionale e della loro rilevanza nel corso degli anni, suggerendo ulteriori percorsi di approfondimento e di studio dell'azione regionale su vari temi di interesse. Considerazioni e valutazioni quali quelle qui esposte possono essere infatti oggetto di sviluppo e approfondimento in svariati ambiti.

Fig. 11 Estratto di alcuni *topics* dal corpus dei Decreti Dirigenziali.



#	words
5	lavoro prevenzione sicurezza controllo alimenti prodotti allegato animali usl laboratorio azienda guida salute entro igiene veterinaria analisi ufficiale produzione alimentari
7	imprese decreto atto toscano toscana fidi artigianato artigiancredito docup art bando allegato presente sensi sviluppo aiuto misura regione aiuti domande
15	toscana usl asl professionale infermiere infermieristico azienda rilevato centro esistente variata d.m.s a.o data ospedaliera careggi pisana nordovest tecnico qualifica
19	scarico acque autorizzazione art presente gestore scarichi atto s.m.i impianto toscana singole vista minima idrico reflue settore ambientale regione depurazione
21	accreditamento formazione atto regionale formativo organismo decreto allegato toscana dgr organismi sistema codice orientamento regione attivita settore vista crediti lavoro
27	regionale caccia vista legge autorizzazione pesca appostamento particolare fisso presente atto art settore faunistico appostamenti titolare venatoria l.r fissi mare
29	trasporto pubblico servizi locale servizio mobilita toscana firenze regione atto trasporti edilizia risorse spa art allegato societa tpl infrastrutture decreto
30	ricerca universita progetto studi bando partner progetti sviluppo toscana capofila innovazione decreto pisa ammesso firenze dipartimento linea s.r.l graduatoria valutazione
52	rifiuti impianto trattamento stoccaggio cer allegato operazioni recupero pericolosi acque prodotti attivita rifiuto autorizzazione gestione area toscana tabella ambiente materiali

Ricerca dei Piani energetici e dei Decreti Dirigenziali che trattano un argomento specifico, definito tramite l'uso di un vocabolario

In questo paragrafo è presentata l'analisi dei due corpora costituiti dai Piani energetici e dai Decreti Dirigenziali di cui al precedente paragrafo secondo un approccio differente, tramite l'uso di un vocabolario controllato. In particolare, si ricerca se e con quale livello di intensità specifici argomenti siano trattati nei documenti esaminati (ad esempio, quanto viene trattato il tema energetico negli atti prodotti in un certo anno?). A differenza del *topic modeling*, l'obiettivo non è in questo caso estrarre dai corpora gli argomenti trattati, ma misurare se e a quale livello di intensità specifici *topics* predefiniti dal ricercatore trovino trattazione all'interno dei corpora.

Rispetto alla tecnica del *topic modeling*, l'approccio per vocabolari controllati si adatta maggiormente ad esigenze di monitoraggio, poiché il vocabolario viene definito a priori e il suo cambiamento rispetto al variare del corpus analizzato è assente o sotto controllo.

Per la creazione del vocabolario, nel quale ciascun argomento (*topic*) viene definito da una serie di termini chiave singoli o multipli, è utile affidarsi alla conoscenza di esperti, per definire e limitare l'ampiezza semantica di ogni argomento tramite l'identificazione di un insieme iniziale di termini chiave. Le parole chiave così definite potranno essere poi arricchite, affidandosi sia a metodi di apprendimento automatico, per recuperare tutti i sinonimi delle parole chiave iniziali, sia alle ontologie costruite sopra i *repository* di conoscenza, per raccogliere tutti i termini che sono collegati tra loro logicamente.

I documenti testuali sono stati elaborati attraverso tecniche di *Natural Language Processing* (NLP) e il loro contenuto è stato successivamente confrontato con il vocabolario, individuando così i documenti che sono in relazione più o meno stretta con il vocabolario stesso.

L'analisi dei risultati ottenuti è stata effettuata, infine, tramite derivazione di *plots* e indicatori, secondo processi e metodi tipici dell'analisi statistica.

I vantaggi della tecnica adottata riguardano principalmente la facilità con cui i vocabolari possono evolvere, essere opportunamente declinati in sottocategorie o combinati per abbracciare più dimensioni di interesse, e la facilità con cui essi si prestano a rappresentare in modo esaustivo un certo ambito di interesse, supportando al riguardo anche la costruzione di più evolute ontologie di dominio.

Gli aspetti meno favorevoli all'adozione di questa tecnica sono invece riconducibili alla necessità di ricercare o di creare di volta in volta il vocabolario per lo scopo prefissato e tale fase di creazione, oltre ad essere onerosa, è critica in quanto influenza fortemente il risultato finale.

Analisi dei Piani energetici tramite uso di vocabolario controllato

L'impiego della tecnica dei vocabolari controllati sul corpus dei Piani energetici ha riguardato il confronto di ciascun Piano rispetto all'Obiettivo di Agenda 2030 numero 13 'Lotta contro il cambiamento climatico'.

Ai fini dell'analisi è stato utilizzato un vocabolario controllato già disponibile online[5] per tutti i *Sustainable Development Goals* (*SDG*) di Agenda 2030, dal quale è stata estrapolata la sezione specifica riguardante l'Obiettivo 13. Il vocabolario ottenuto è stato sottoposto ad un'attività manuale di traduzione ed arricchimento, che ha portato ad un elenco finale di 643 termini ed espressioni, sia in lingua italiana che in inglese (Fig. 12).

Fig. 12 Vocabolario controllato sul *topic* 'lotta contro il cambiamento climatico'.

efficienza energetica	ghg	pollution	carbon dioxide	ossido nitroso
riduzione	energy management	costo ambientale	carbon leakage	paris agreement
risparmio energetico	gestione energetica	mitigazione dei cambiamenti climatici	energy policy	pollutant emission
co2	società di servizi energetici	global warming	temperatura globale	precarietà energetica
biogas	carbon tax	infrastruttura verde	eventi meteorologici estremi	rischio climatico
rischi	emissioni di metano	emissione di carbonio	impronta di carbonio	bilancio del carbonio
emissione	low carbon	economia a basse emissioni di carboni	incremento della temperatura	condizioni meteorologiche estreme
gas serra	acea	carbon footprint	variazione di temperatura	disaster risk reduction
bilancio energetico	siccità	climate action	architettura sostenibile	green climate fund
impatto ambientale	climate change	nuova energia	electric vehicle	greenhouse gas emissions
cambiamenti climatici	alluvione	bassa emissione di carbonio	new energy	inquinamento da co2
politica energetica	povertà energetica	veicolo elettrico	riscaldamento climatico	politiche sul clima
riqualificazione energetica	bioedilizia	emissione di inquinanti	sviluppo durevole	power purchase agreement
energia rinnovabile	edilizia sostenibile	air pollution	anidride carbonica equivalente	rimozione dell'anidride carbonica
sviluppo sostenibile	emissioni zero	calore rinnovabile	carbon dioxide equivalent	schema di riduzione
effetto serra	energy service company	erosione costiera	convenzione delle alpi	consumo mondiale di energia
prestazione energetica	lotta ai cambiamenti climatici	eventi alluvionali	credito di carbonio	pompa di calore ad aria
pompa di calore	condizioni meteo	greenhouse gas	accumulo di energia termica	società a basse emissioni di carbonio
anidride carbonica	emissions trading	riduzione di co2	floods	
green economy	energy efficiency	temperatura media globale	geoingegneria	
rendimento energetico	bioenergia	rimboschimento	hazards	
gas a effetto serra	audit energetico	co2 emission	permafrost	
cambiamento climatico	renewable energy	ondata di caldo	storm	
alluvioni	inondazione	azione per il clima	bus elettrico	
diagnosi energetica	pericoli	biossido di carbonio equivalente	cambiamento ambientale	
inquinamento atmosferico	aumento della temperatura	biofuel	certificato energetico	
compensazione	low carbon economy	green building	design sostenibile	
consumo di energia elettrica	economia verde	impronta ecologica	disaster resilience	
co2 equivalente	pannello solare	pannello fotovoltaico	energy poverty	
gas bio	emissioni di gas a effetto serra	sustainable development	environmental impact	
contaminazione	inquinamento dell'aria	scambio delle quote di emissione	facility management	
recupero energetico	doppi vetri	bioenergy	gas biologico	
biossido di carbonio	neutralità	maltempo	greenhouse effect	
gas ad effetto serra	energia bio	neutrale	impatto climatico	
adattamento ai cambiamenti climatici	inondazioni	terremoto	low-carbon economy	

[5] http://doi.org/10.5281/zenodo.3567769.

Ciascun termine o espressione presente nel vocabolario è stato successivamente ricercato all'interno dei documenti di programmazione regionali e nazionale, ottenendo una matrice termini-documenti con un numero di righe corrispondente al numero di espressioni o di singoli termini individuati e un numero di colonne pari a quello dei Piani energetici esaminati.

Si veda a tal riguardo la Fig. 13, che riporta un estratto della matrice, costituita complessivamente da 158 righe e 22 colonne. Nella matrice si trovano, in corrispondenza di ogni incrocio tra termine e documento, la frequenza assoluta con cui le parole o espressioni compaiono all'interno del documento.

Fig. 13 Matrice termini-documenti per i Piani energetici rispetto al *topic* 'lotta al cambiamento climatico'.

	Abruzzo	Basilicata	Calabria	Campania	Emilia_romagna	Friuli_V.G.	Italia	Lazio	Liguria	Lombardia	Marche	Piemonte	Prov_Trento	Puglia	Sardegna	Sicilia	Toscana	Umbria
efficienza energetica	23	45	62	149	55	279	180	381	273	166	154	130	47	94	134	27	47	42
riduzione	60	102	170	171	101	240	161	268	175	162	458	262	77	202	188	11	99	44
risparmio energetico	42	91	90	50	52	94	21	66	48	44	154	20	6	70	36	1	10	41
co2	111	68	77	143	51	192	64	100	44	66	254	49	48	114	112	38	34	26
biogas	9	13	7	58	14	15	10	83	136	171	26	69	54	28	33	52	4	1
rischi	15	11	15	27	3	53	41	69	34	22	59	80	1	23	13	5	101	5
emissione	5	7	22	15	4	119	9	9	19	13	118	14	8	53	39	4	4	13
gas serra	24	14	24	20	40	76	26	20	12	0	138	19	1	25	0	1	12	7
bilancio energetico	24	36	22	15	6	14	2	13	41	14	147	19	1	9	29	1	0	0
impatto ambientale	7	22	10	21	3	35	12	35	22	9	34	14	7	20	19	27	11	15
cambiamenti climatici	1	11	2	4	5	66	19	15	32	6	15	16	6	12	15	10	39	0
politica energetica	16	88	15	11	11	27	4	10	18	15	46	3	3	16	8	0	3	9
riqualificazione energetica	2	3	0	62	19	32	26	60	16	24	0	19	10	2	8	7	0	0
energia rinnovabile	1	2	1	17	6	35	47	26	38	9	12	22	5	3	8	6	4	10
sviluppo sostenibile	0	12	3	1	7	22	7	27	13	5	11	33	2	14	3	8	16	4
effetto serra	5	9	2	9	4	22	41	10	13	9	28	7	2	13	9	1	2	5
prestazione energetica	0	0	0	8	6	14	18	41	12	23	2	24	0	2	21	0	1	0
pompa di calore	0	3	39	18	0	16	6	10	20	4	1	15	4	0	29	2	1	0
anidride carbonica	0	5	15	5	0	1	2	4	6	5	65	6	27	22	13	0	5	7
green economy	0	0	0	5	8	2	0	83	24	12	0	35	1	0	0	0	31	0
rendimento energetico	4	1	2	3	0	9	3	4	7	8	19	9	15	12	6	0	0	13
gas a effetto serra	1	1	0	5	3	16	36	5	9	4	9	5	2	4	5	1	1	0
cambiamento climatico	1	4	2	2	5	5	4	5	8	7	11	29	0	4	1	7	10	2
alluvioni	1	1	0	0	0	0	1	17	7	0	0	13	0	0	0	0	19	0
diagnosi energetica	3	0	0	5	0	15	2	2	6	1	0	0	0	3	6	0	0	2
inquinamento atmosferico	4	2	8	5	0	8	2	14	2	3	4	6	4	5	1	0	7	3

A partire dalla matrice così ottenuta possono essere realizzate varie analisi. Nei successivi esempi (Fig. 14 e Fig. 15), tali analisi sono state realizzate in linguaggio R. La prima è rappresentata con mappe di calore (*heatmap*) nelle quali, ad una maggiore frequenza di ciascun termine o espressione, corrisponde una colorazione più intensa, consentendo di individuare i termini più rilevanti sia per i singoli Piani che per il corpus nel suo insieme. La Fig. 14 riporta, a destra, la *heatmap* complessiva, e a sinistra, per favorire la lettura, uno spaccato di sei termini principali.

Una diversa rappresentazione attraverso dendrogramma (Fig. 15), permette invece di stimare il livello di similarità (o di distanza) tra i contenuti dei Piani energetici rispetto alla trattazione delle tematiche caratterizzanti l'Obiettivo 13 dell'Agenda 2030, colte attraverso l'uso del vocabolario controllato descritto in precedenza.

Fig. 14 Mappe di calore dei Piani energetici rispetto al *topic* 'lotta al cambiamento climatico'.

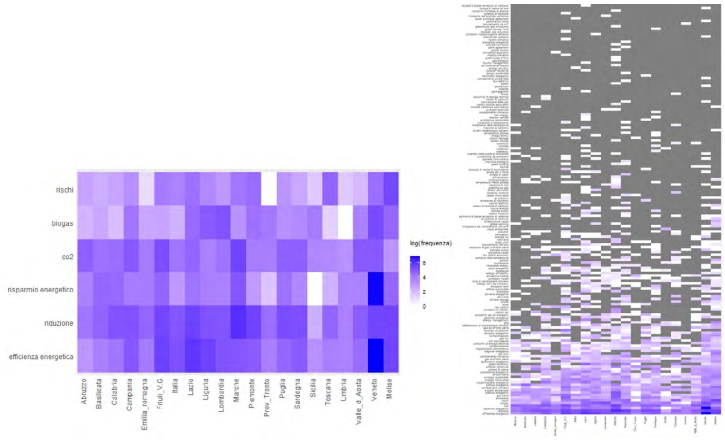
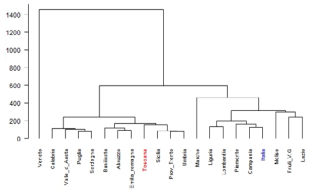


Fig. 15 Similarità dei Piani energetici rispetto alla trattazione del *topic* 'lotta al cambiamento climatico'.



Questa rappresentazione suggerisce, ad esempio, l'appartenenza del Piano energetico della Toscana ad un *cluster* che ricomprende anche quelli siciliano, umbro e della Provincia di Trento (tutti collocati ad una certa distanza rispetto al Piano nazionale, al contrario di quanto avviene in altre regioni, quali Piemonte, Campania, Lombardia e Liguria). Questo primo risultato suggerisce inoltre l'opportunità di approfondire i contenuti del Piano energetico veneto, che, come si può notare, sembra avere caratteristiche singolari e anomale rispetto a tutti gli altri.

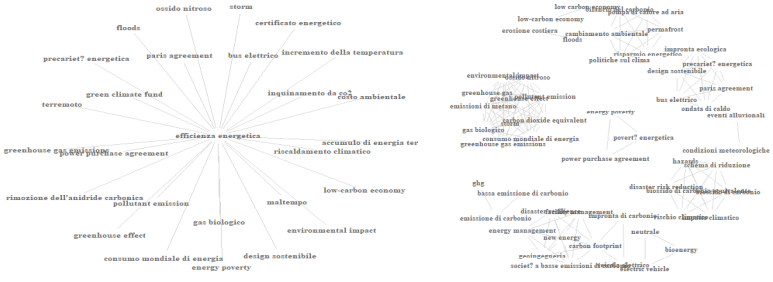
A partire dalla stessa matrice termini-documenti, è stato inoltre possibile costruire mappe semantiche sull'intero corpus dei Piani energetici in relazione all'Obiettivo 13 di Agenda 2030, nel tentativo di estrapolare significati latenti, non immediatamente rilevabili nelle analisi precedenti.

Una prima rappresentazione in questo senso è stata costruita sulla base di una matrice di distanze, calcolata a partire dalla matrice termi-

ni-documenti di cui sopra. Ognuno dei 158 termini disponibili nella matrice di origine può difatti essere interpretato come un vettore di 22 valori (un valore per ogni Piano), ciascuno dei quali è costituito dalla frequenza con cui ciascun termine compare all'interno di ogni documento (ad esempio, il vettore relativo al termine "efficienza energetica" mostrato alla riga 1 della Fig. 13 contiene i valori 23 per Abruzzo, 45 per Basilicata, e così via). Presa una qualunque coppia di termini presenti nella matrice è quindi possibile calcolare la distanza euclidea tra i relativi vettori e assumere che la distanza complessiva tra i due termini sia tanto minore quanto minore risulta la distanza euclidea calcolata tra i rispettivi vettori. Sulla base di questa assunzione, si è ottenuto un primo grafo (Fig. 16, parte sinistra) da cui si evince la complessiva centralità del tema di "efficienza energetica" che caratterizza i vari Piani e che si lega ad una serie di altri concetti, quali ad esempio quelli di "bus elettrico" o di "maltempo".

Nella Fig. 16, parte destra, viene invece mostrato il risultato ottenuto applicando un metodo di confronto tra i termini che, invece di utilizzare le distanze euclidee, è basato sul calcolo della matrice delle correlazioni positive tra i medesimi vettori. In questo caso, il risultato che si ottiene esprime la probabilità che coppie di termini siano presenti nello stesso Piano: si evince ad esempio che i PPA (*Power Purchase Agreement*, accordi di fornitura di energia elettrica a lungo termine tra due parti, che possono ridurre i rischi legati ai prezzi di mercato e sono utilizzati tra l'altro nel caso di investimenti per la costruzione o il mantenimento di impianti di energia rinnovabile) sono spesso associati nei Piani al tema della povertà energetica, oppure che il concetto di *floods* (inondazione) è spesso trattato congiuntamente a quello di erosione costiera.

Fig. 16 Mappe concettuali del *topic* 'lotta al cambiamento climatico'.



I risultati ottenuti non pretendono di essere esaustivi, ma hanno lo scopo di indagare in fase sperimentale le tecniche di analisi e suggerirne le potenzialità, tenendo conto che in una fase di concreta applicazione, esse dovranno essere accompagnate da una lettura critica e da un loro opportuno raffinamento.

Analisi dei Decreti Dirigenziali tramite uso di vocabolario controllato

Si riporta, infine, un breve cenno alle potenzialità del metodo basato sull'utilizzo di vocabolari controllati anche nell'ambito dell'analisi del corpus dei Decreti Dirigenziali di Regione Toscana. In questo caso è stato necessario l'impiego di tecniche di elaborazione altamente performanti, data l'ampia quantità di documenti processati (la base informativa, costituita da circa 150.000 documenti, è stata elaborata mediante le funzionalità esposte dal progetto Lucene[6]). Si è ottenuta anche in questo caso, come primo risultato intermedio, una matrice termini-documenti (di cui si riporta un estratto in Fig. 17), che in corrispondenza di ciascun incrocio tra termine o espressione (ad esempio "alluvioni" o "doppi vetri") e documento (ad esempio il Decreto n. 6704 del 12 maggio 2020) esprime un indice di associazione tra quel termine o espressione e quel documento, chiamato indice *tscore*.

Fig. 17 Estratto della matrice lemmi-documenti ottenuta elaborando la produzione amministrativa regionale.

RICERCA	NUMERO	DATA_ATTO	TSCORE
"alluvioni"	6.704	2020-05-12	2,8378
"alluvioni"	6.720	2020-05-12	2,5397
"alluvioni"	6.712	2020-05-12	2,5021
"alluvioni"	6.692	2020-05-12	2,5021
"alluvioni"	6.706	2020-05-12	2,5021
"condizioni meteo"	6.719	2020-05-12	2,3651
"condizioni meteo"	6.691	2020-05-12	2,2001
"contaminazione"	6.760	2020-05-12	3,9724
"contaminazione"	6.808	2020-05-12	3,1379
"contaminazione"	6.795	2020-05-12	0,5643
"doppi vetri"	6.757	2020-05-12	1,5478
"emissione"	6.753	2020-05-12	2,35107
"emissione"	6.808	2020-05-12	2,32043
"emissione"	6.757	2020-05-12	2,09931
"emissione"	6.756	2020-05-12	2,0849
"emissione"	6.793	2020-05-12	2,01841
"emissione"	6.795	2020-05-12	1,95677
"emissione"	6.752	2020-05-12	1,86116
"Inquinamento"	6.778	2020-05-12	2,21144
"Inquinamento"	6.790	2020-05-12	2,15343
"Inquinamento"	6.800	2020-05-12	2,14567
"Inquinamento"	6.776	2020-05-12	2,11423

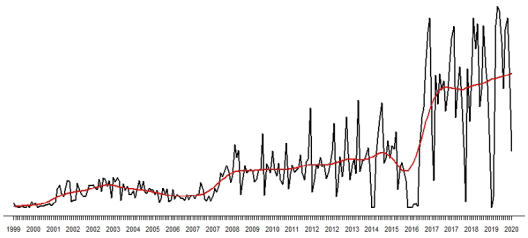
L'indice *tscore* è un indicatore statistico di rilevanza del documento rispetto all'espressione in esso ricercata, e il valore che assume è tanto

[6] https://lucene.apache.org/core/2_9_4/api/core/org/apache/lucene/search/Similarity.html, org.apache.lucene.search, Class Similarity.

maggiore quanto più il termine – sia nella sua forma primitiva che in quelle derivate (ed esempio, da "carbone" possono derivare "carbonifero", "carbonico", ecc.) – è ripetuto all'interno del documento. Nell'ottica di assumere un valore di soglia per il *tscore* oltre il quale ritenere significativa l'associazione termine-documento, nell'analisi condotta si è ipotizzata un'associazione significativa con un valore di *tscore* superiore a 0,75, e particolarmente rilevante in corrispondenza di un *tscore* superiore a 2,5.

A partire da questa base sono state ripetute anche per i Decreti Dirigenziali le analisi viste in precedenza per i Piani energetici: a titolo esemplificativo, uno dei risultati ottenuti è mostrato dal grafico in Fig. 18, che riporta l'andamento della serie storica del punteggio *tscore* riguardo agli atti regionali con contenuti compatibili rispetto all'Obiettivo 13 di Agenda 2030 "lotta al cambiamento climatico", suggerendo un rilevante aumento, nell'ultimo quinquennio, della trattazione di tali tematiche.

Fig. 18 Andamento del *topic* 'lotta al cambiamento climatico' nella produzione amministrativa regionale.



Un'interessante applicazione dello stesso metodo è stata condotta successivamente a tali sperimentazioni anche nell'ambito e per le finalità che rientrano tipicamente nella gestione di un sistema documentale, per il quale vi può essere l'interesse a catalogare un corpus di documenti disponibili rispetto ad un sistema di classificazione definito a priori, anche allo scopo di corredare ciascun documento di un opportuno sistema di metadati che ne consenta una ricerca più agevole e immediata. Il sistema di classificazione potrebbe esprimere, ad esempio, i molteplici ambiti di attività dell'ente (sistemi informativi, gestione del personale, finanziamenti alle imprese agricole, ecc.) e lo scopo potrebbe essere quello di elencare o conteggiare nel tempo quanti Decreti Dirigenziali sono stati adottati per ciascun ambito di attività (ad esempio quanti e quali atti fanno riferimento all'erogazione di finanziamenti o a concorsi pubblici, ecc.).

Il vantaggio nell'applicazione di queste tecniche consiste nel fatto che

con tutta probabilità i documenti da classificare non sono già stati classificati manualmente rispetto al medesimo sistema di classificazione, anche alla luce del fatto che il sistema stesso può nascere ed evolvere in un momento successivo all'emissione dell'atto, per finalità di classificazione e analisi che al momento in cui l'atto stesso è stato adottato non erano note o previste.

Si presenta di seguito una simulazione relativa alla ricerca delle attività svolte dalla Regione Toscana nell'ambito dei Sistemi informativi, che è stata costruita elaborando il corpus dei Decreti Dirigenziali regionali. La Fig. 19 riporta un estratto delle attività svolte da Regione Toscana nell'ambito dei Sistemi informativi, i cui codici costituiscono l'asse delle ascisse del grafico in figura 20.

Fig. 19 Elenco di attività svolte in ambito di Sistema informativo regionale.

E.070.010.010	Sistemi di elaborazione centrali
E.070.010.020	Infrastrutture di rete
E.070.010.030	Servizi di rete e comunicazione telematica
E.070.010.040	Sistemi e servizi per la sicurezza
E.070.010.050	Risorse e strumentazione informatica per gli uffici regionali
E.070.020.010	Applicativi, banche dati e sistemi gestionali
E.070.020.020	Applicativi, banche dati e sistemi di settore
E.070.020.030	Servizi trasversali su applicazioni e banche dati
E.070.020.040	Sistemi informativi di supporto delle decisioni
E.070.030.010	Sistema informativo statistico
E.070.030.020	Sistema informativo in materia di finanza delle autonomie locali
E.070.030.030	Sistema informativo in materia di appalti
E.070.030.040	Sistema informativo in materia di agricoltura e allevamento
E.070.030.050	Sistema informativo in materia di artigianato
E.070.030.060	Sistema informativo in materia di commercio
E.070.030.070	Sistema informativo in materia di industria e piccola e media impresa
E.070.030.080	Sistema informativo in materia di turismo
E.070.030.090	Sistema informativo sul territorio
E.070.030.100	Sistema informativo in materia di ambiente
E.070.030.110	Sistema informativo sociosanitario
E.070.030.120	Sistema informativo in materia di lavoro, orientamento e formazione professionale
E.070.030.130	Sistema informativo in materia di beni e attività culturali
E.070.030.140	Sistema informativo in materia di attività sportive e ricreative
E.070.030.150	Sistema informativo per l'Edilizia residenziale sociale
E.070.040.010	Tecnologie e servizi per l'inclusione e i diritti di cittadinanza
E.070.040.020	Tecnologie e servizi per la pubblica amministrazione
E.070.040.030	Tecnologie e servizi a supporto della competitività delle imprese
E.070.040.040	Architetture e infrastrutture abilitanti
E.070.060.010	Piano annuale di attività della rete telematica regionale - PAR
E.070.060.020	Coordinamento delle attività di supporto agli organismi di governo della Rete
E.070.060.030	Centri di competenza
E.070.060.040	Procedimenti di adesione
E.070.060.050	Raccordi istituzionali

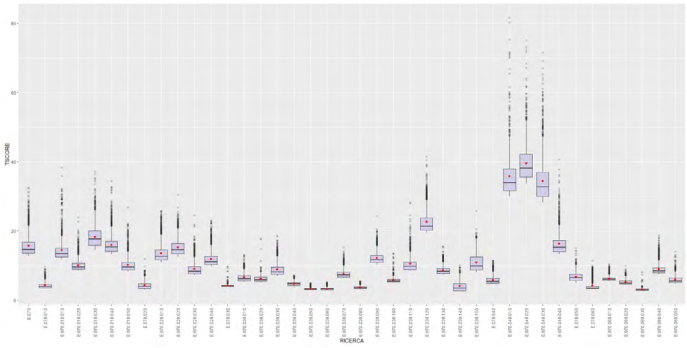
La Fig. 20 riporta, in corrispondenza di ciascuna tipologia di attività, un *box-plot* (grafico a scatola) che esprime come si distribuiscono gli atti rispetto al valore che assume l'indice *tscore* calcolato per quella tipologia di attività, con evidenza del valore medio, dei principali percentili (mediana, quartili), della variabilità e dell'asimmetria di ciascuna distribuzione (il grafico a scatola ha il vantaggio di dare evidenza in modo compatto ed efficace di molti indicatori statistici rilevanti per interpretare correttamente la forma della distribuzione e i suoi principali indici di posizione, variabilità e simmetria).

È utile precisare che ciascun Decreto può essere presente in più distribuzioni (ovvero in più *box-plot*) in quanto associato dall'algoritmo a più voci della classificazione.

Dalla figura 20 si evince la presenza di tre tipologie di attività particolarmente significative, che rientrano negli ambiti delle tecnologie e dei servizi per l'inclusione e i diritti di cittadinanza, per la pubblica ammini-

strazione e per il supporto alla competitività delle imprese (un'ulteriore materia di rilievo riguarda lavoro, orientamento e formazione professionale). Osservando ciascun *box-plot* si rileva che il livello di correlazione tra i Decreti e le aree di attività può variare anche considerevolmente. Per ciascun *box-plot* si possono ad esempio distinguere i decreti che risultano associati ad una certa area tematica con un livello di intensità (*tscore*) superiore alla media, oppure superiore alla mediana o al terzo quartile della distribuzione. Questa constatazione permette ad esempio di considerare significative e valide solo le relazioni tra Decreto e aree di attività che presentano un *tscore* superiore alla media di quell'area tematica (oppure alla mediana o al terzo quartile).

Fig. 20 Decreti Dirigenziali di Regione Toscana rispetto alle attività svolte in ambito di Sistemi informativi.



Confronto tra i Piani energetici sulla base del lessico utilizzato

In questo paragrafo i Piani energetici regionali e nazionale sono confrontati sulla base dei rispettivi profili lessicali, con lo scopo di mettere a confronto il lessico utilizzato.

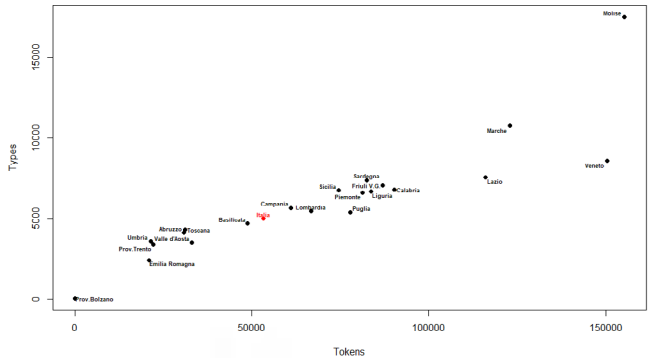
Tale analisi sperimentale è stata condotta solo sul corpus dei Piani energetici e non anche su quello dei Decreti, ed ha innanzitutto previsto, come attività preliminari, la selezione e il successivo trattamento delle sole parole piene.

Sono stati successivamente calcolati alcuni indicatori relativi alla struttura sintattica di ciascun documento, che hanno permesso di ottenere, come risultato, la misura delle distanze intertestuali tra i documenti analizzati.

Un primo risultato è rappresentato dal grafico in Fig. 21, che esprime il confronto tra i Piani energetici rispetto al numero di *tokens* (oc-

correnze delle forme grafiche presenti nel corpus ripulito, includendo quindi, nel conteggio, anche le forme ripetute) e di *types* (forme grafiche distinte, ottenute escludendo dal conteggio le ripetizioni presenti a parità di *token*).

Fig. 21 Primo confronto dei piani energetici sotto il profilo lessicale.

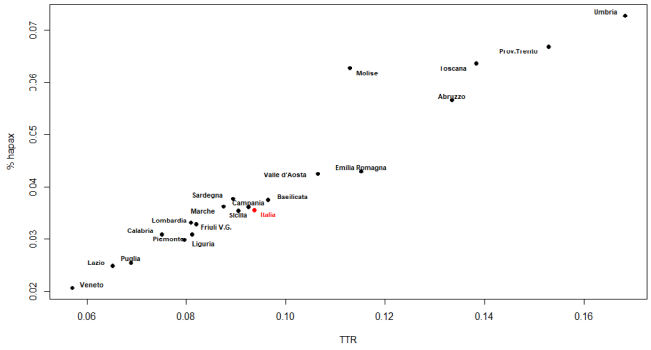


Nella figura, i Piani energetici posizionati a destra lungo l'asse delle ascisse (ad esempio i piani di Molise, Veneto, Marche e Lazio) sono caratterizzati da una maggior quantità di *tokens*, quindi da una maggior prolissità. Rispetto a questa dimensione di analisi, si evince quindi che il Piano energetico nazionale si colloca in una posizione mediana e che, ad esempio, il Piano toscano è meno prolisso di quello nazionale, essendo collocato più a sinistra. Tuttavia, non sempre e non necessariamente una maggior quantità di parole è associata ad una maggior ricchezza linguistica o di concetti espressi. Una prima quantificazione di tale ricchezza linguistica può essere stimata dal numero di *types*, che è la seconda dimensione espressa dal grafico tramite il posizionamento dei Piani rispetto all'asse delle ordinate. Confrontando ad esempio il Piano del Lazio con quello di Sardegna o con quelli corrispondenti agli altri punti vicini (Friuli-Venezia Giulia, Liguria, Piemonte, Calabria), si evince come il primo si presenti molto più prolisso, a fronte di una ricchezza lessicale sostanzialmente analoga.

A partire dai suddetti indicatori, costituiti dal semplice conteggio dei *tokens* e dei *types*, è stato calcolato un ulteriore indice chiamato *Types-Tokens Ratio* (TTR), espresso dal rapporto tra il conteggio dei *types* e dei *tokens* presenti in ciascun documento. L'indice TTR esprime in modo standardizzato, ossia confrontabile tra i vari documenti, la ricchezza lessicale che caratterizza ciascun Piano.

Quest'ultimo indice è stato messo infine in relazione con la percentuale di *hapax*, ossia di termini che compaiono nel testo una sola volta, senza ripetizioni. La percentuale degli *hapax* è il rapporto tra il numero dei lemmi che compaiono una sola volta in ciascun documento e il numero complessivo dei *tokens* presenti nello stesso documento. Tale indicatore è utile anche in via preliminare ad esprimere l'attitudine del documento ad essere trattato con algoritmi quantitativi di analisi testuale (maggiore la percentuale di *hapax*, minore l'attitudine del documento ad essere elaborato con algoritmi di text mining). Nella Fig. 22 è riportato il confronto tra i Piani energetici basato sugli ultimi due indici, dove entrambi sono espressione diretta del livello di ricchezza lessicale di ciascun Piano.

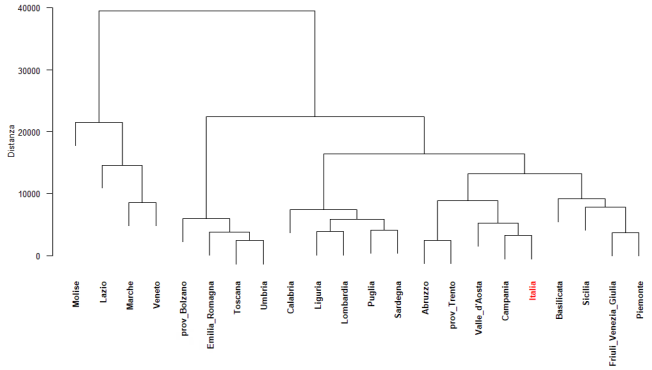
Fig. 22 Secondo confronto dei piani energetici sotto il profilo lessicale.



Dal confronto emerge che il Piano umbro, seguito da Provincia di Trento, Toscana e Abruzzo, presenta una ricchezza lessicale significativamente maggiore e che il Piano nazionale si colloca ancora in una posizione intermedia rispetto agli altri. In questo senso, l'analisi fornisce un primo indizio anche sulla qualità degli strumenti di pianificazione adottati dalle diverse amministrazioni.

Il grafico (dendrogramma) in Fig. 23 mostra, infine, l'albero di raggruppamento tra i Piani energetici in base alla somiglianza o distanza dei loro profili lessicali sulla base dei precedenti indici, ottenuto attraverso l'analisi della distanza intertestuale tra i documenti. Il Piano toscano risulta simile a quello umbro ed appartiene ad un *cluster* che ricomprende anche i Piani energetici di Emilia-Romagna e della Provincia di Bolzano (si noti invece come i Piani di Molise, Lazio, Marche e Veneto costituiscano assieme un *cluster* distante rispetto agli altri Piani, analogamente a quanto osservato in Fig. 21).

Fig. 23 Similarità dei Piani energetici rispetto al profilo lessicale.



Le potenzialità del text mining nell'ambito dell'analisi della comunicazione social (*social mining*)

In quest'ultimo paragrafo si intende presentare l'ultimo caso di studio, relativo all'analisi del sentimento desumibile dai social network. L'aspetto su cui si ritiene interessante focalizzare l'attenzione del lettore è che, seppure in un ambito molto diverso rispetto ai precedenti casi esaminati, anche questo caso vede l'applicazione di tecniche di analisi testuale sostanzialmente analoghe (la struttura di un *post* non è del resto così diversa rispetto a quella dell'oggetto di un decreto, ad esempio).

Il corpus è in questo caso costituito da dati provenienti da Twitter, su cui è stato implementato un prototipo sviluppato nelle seguenti fasi:

- recupero dei tweet a partire da una lista di *hashtag* o parole chiave, secondo un criterio assimilabile a quello dei vocabolari controllati presentato in precedenza;
- raggruppamento dei tweet raccolti in insiemi omogenei di tematiche trattate, secondo un approccio non supervisionato, paragonabile alle tecniche di *topic modeling* trattate in precedenza;
- analisi del sentimento che caratterizza il corpus dei *tweet* raccolti ed analisi statistica degli utenti e del loro livello di interesse sulle specifiche tematiche trattate dal corpus stesso;
- raccolta dei risultati in una *dashboard* interattiva che mostra alcune rappresentazioni grafiche utili a fornire un'idea di insieme.

Oggetto della sperimentazione sono stati 4.479 *tweet*, contenenti complessivamente 750 differenti *hashtag*. L'ipotesi di base del lavoro è che gli

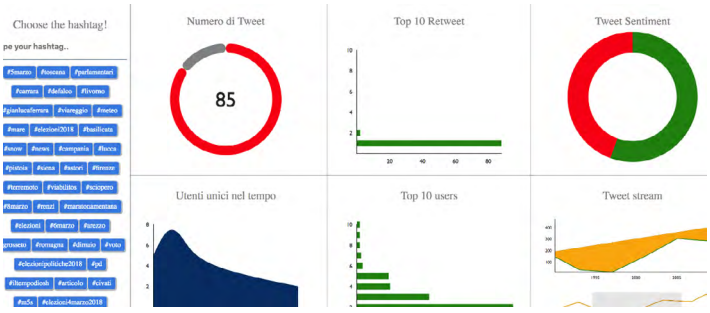
hashtag intercettati possano essere considerati rappresentativi del contenuto del *tweet* e che il *topic* sia individuato da un insieme di *hashtag*.

In prima istanza, secondo un approccio di tipo *unsupervised*, è stato implementato un grafo (Fig. 24) nel quale ciascun *tweet* viene rappresentato da un punto e due *tweet* risultano tra loro connessi se condividono un *hashtag*. Ogni *tweet* è colorato in base al *cluster* di appartenenza, quindi è possibile individuare a colpo d'occhio la quantità dei principali argomenti trattati nei *tweet* e il volume di *post* che hanno generato.

Fig. 24 Raggruppamento dei *tweet* raccolti rispetto agli *hashtag* in insiemi di tematiche.

Nella seconda fase di sviluppo, è stata messa a disposizione una *dashboard* interattiva (Fig. 25), contenente, per gli *hashtag* o i *topics* selezionati, alcuni dati statistici di possibile interesse, quali, ad esempio, il numero di *tweet* ricompresi nell'*hashtag* o nel *topics* selezionato, l'andamento nel tempo della numerosità dei *tweet*, il numero di *re-tweet*, la numerosità di utenti distinti che hanno trasmesso i *tweet* presi in esame.

Fig. 25 *Dashboard* interattiva per la consultazione dei dati statistici sui *tweet* estratti.



Al fine di completare i contenuti della *dashboard* con una *sentiment analisys* (approccio *supervised*), ogni termine contenuto in ciascun *tweet* è stato classificato come positivo, negativo o neutro secondo un vocabolario controllato. Si tratta di un approccio più semplice e in linea teorica più limitato rispetto all'addestramento di un algoritmo di intelligenza artificiale, in quanto non consente di desumere il livello di positività o meno del *tweet* tenendo conto dei significati desumibili dal testo del *tweet* considerato nel suo insieme. In questo senso, l'utilizzo in questo contesto di un approccio *machine learning* come quello presentato per il primo caso di studio (reti neurali) avrebbe consentito di ottenere livelli di accuratezza più alti, presentando di contro la necessità di costruire il dataset di *training* su cui addestrare l'algoritmo avendo preventivamente classificato un adeguato numero di occorrenze.

Conclusioni

A corollario delle suddette esperienze, che si sono svolte nell'arco dei mesi da marzo a maggio 2018 e di cui sono stati descritti i tratti salienti, si ritiene utile trarre alcune considerazioni di carattere generale che appaiono significative.

In primo luogo, l'attività svolta e la condivisione di conoscenza che ne è derivata sono state possibili grazie all'utilizzo di dati in formato aperto, software liberi e forte spirito di collaborazione tra amministrazioni pubbliche, in forza della condivisione di progetti di comune interesse, per quanto sperimentali.

Il carattere 'libero' dei software e dei linguaggi utilizzati, peraltro in linea con gli artt. 68 e 69 del Codice dell'amministrazione digitale, dovrebbe consentire anche una rapida trasferibilità dei progetti e costituisce, più in generale, una forte spinta all'innovazione in un contesto, come quello della pubblica amministrazione, nel quale vi è grande necessità e vantaggio ad aggiornarsi e ad apprendere gli uni dagli altri.

In secondo luogo, al fine di condurre ricerche e sperimentazioni orientate alla realizzazione di sistemi di intelligenza artificiale per i vari domini di interesse della pubblica amministrazione, si ritiene che l'ambito delle tecnologie linguistiche sia uno di quelli più promettenti.

In questo capitolo sono stati toccati contesti di applicazione caratterizzati in generale da volumi di dati elevati, variabilità e mancanza di struttura nei dati, nonché rapidità di elaborazione richiesta a fronte dei caratteri precedenti: siamo, in pochi termini, nel contesto dei *big data*, per il cui trattamento, in un ambito di considerazioni attinenti al trasferi-

mento e alla messa a regime delle esperienze fatte, è opportuno e urgente dotarsi di appropriate piattaforme tecnologiche e competenze.

Per operare in ambiti di applicazione quali quelli oggetto delle sperimentazioni descritte, costituisce infatti una criticità la scarsa disponibilità, sul mercato in generale e nella pubblica amministrazione in particolare, di figure professionali adeguate, quale quella del *data scientist*. Giova considerare al riguardo l'attenzione rivolta anche dal recente Piano Nazionale di Ripresa e Resilienza (PNRR) alle attività di *upskilling* e *reskilling* per lo sviluppo di competenze in tema di digitale. Le possibili linee di azione che si intravedono per l'amministrazione pubblica nell'attuale contesto sono quelle di formare personale già presente al suo interno, acquisire servizi da soggetti terzi tramite gare d'appalto o, infine, attivare collaborazioni tra amministrazioni in forza della condivisione di progetti di comune interesse, mettendo così a fattor comune le professionalità disponibili.

Riguardo all'ultimo punto, si sottolinea come la disponibilità di competenze interne costituisca un fattore determinante per l'introduzione equilibrata delle nuove tecnologie, per le quali è importante che le amministrazioni mantengano un livello adeguato di conoscenza e non si trovino ad essere dipendenti da piattaforme e soluzioni chiuse e fornite da un ristretto numero di soggetti.

Del resto, anche in ambito di normativa *privacy* è ribadito il principio di trasparenza che deve ispirare la progettazione di servizi basati su soluzioni di intelligenza artificiale, sia riguardo agli algoritmi utilizzati che riguardo alle logiche di costruzione delle basi dati su cui essi operano, mettendo in grado l'amministrazione di motivare i suoi provvedimenti anche nella parte elaborata da sistemi di intelligenza artificiale, nonché di rendere consapevoli i responsabili dei procedimenti amministrativi dei criteri di elaborazione adottati.

Valorizzazione di una fonte archivistica: i verbali della Commissione araldica veneta

Salvatore Alongi[1]

Venezia, Archivio di Stato, Commissione araldica, Verbali, Genealogia, Storia.

Introduzione

Oggetto di trattazione nel presente capitolo è un intervento di riproduzione, trascrizione e infine analisi automatica tramite l'utilizzo del software Iramuteq del contenuto dei verbali delle riunioni della Commissione araldica veneta svolte nei due quadrienni 1889-1893 e 1938-1942, progettato ed eseguito tra l'aprile e il settembre 2020 a corredo dell'attività di riordinamento e inventariazione dell'intero archivio della Commissione, un'istituzione risalente al periodo post-unitario (1889-1946) incaricata della compilazione dell'elenco delle famiglie italiane residenti nella regione storica della Venezia in legittimo possesso di titoli nobiliari, e il cui fondo documentario è conservato oggi dall'Archivio di Stato di Venezia, Istituto dipendente dal Ministero della cultura.

[1] Dal 2009 al 2018 è stato archivista libero professionista, collaborando con numerosi enti e istituti culturali tra Bologna e Venezia. Dal 2018 è funzionario archivista presso l'Archivio di Stato di Venezia, dove è responsabile della Gestione informatica corrente e del Servizio di fotoriproduzione. È inoltre titolare degli insegnamenti di Archivistica generale nelle Scuole di APD di Bologna e di Modena, e di Legislazione dei beni culturali e di Lineamenti di diritto amministrativo nella Scuola di APD di Venezia.

L'obiettivo del lavoro è quello di valorizzare una fonte archivistica, tramite l'applicazione di tecniche innovative per l'analisi del contenuto dei testi, dando particolare rilievo al dato storico e all'evoluzione temporale dei fenomeni, offrendo agli storici uno strumento di orientamento e di studio dei temi, problemi e campi di interesse toccati dall'attività della Commissione, e all'Archivio di Stato la possibilità di coniugare in maniera innovativa le funzioni di conservazione, fruizione e valorizzazione del patrimonio documentario conservato.

Duplice, dunque, il risultato atteso che tale intervento dovrebbe apportare nel lungo periodo: il ricorso alla riproduzione e alla trascrizione dovrebbe contribuire a ridurre (fin quasi a eliminarla del tutto) la necessità di consultare in originale la documentazione, assicurandone intatta la conservazione, mentre il ricorso agli esiti dell'analisi testuale dovrebbe consentire di abbattere i costi della ricerca (con un risparmio in termini di risorse e di tempo) grazie a una maggiore velocità e una più rigorosa sistematicità delle operazioni di spoglio e di sintesi delle informazioni di interesse.

Il progetto di analisi automatica del contenuto dei verbali prodotti dalla Commissione araldica veneta oggetto del capitolo rientra in una più ampia politica culturale, avviata dall'Istituto fin dal 2003, di recupero e valorizzazione delle fonti per la storia biografico-famigliare e araldico-genealogica.

Tale politica si sostanzia, nello specifico, in due principali filoni di attività.

Il primo, e certamente più complesso e sfaccettato, oltre che dalle ampie ricadute extra-istituzionali, riguarda il trattamento della documentazione anagrafico-militare. L'interesse verso tale genere di fonti archivistiche ha difatti conosciuto un incremento esponenziale negli ultimi decenni, legato – da una parte – al mutamento della tipologia del pubblico degli utenti degli archivi, che sempre più spesso conduce ricerche rapide e puntuali sulla proprietà privata o sulla storia famigliare (basti pensare che nel lustro 2015-2020 ben il 24% degli iscritti italiani alla Sala di studio dell'Archivio di Stato di Venezia ha dichiarato come motivo della richiesta di ammissione un "interesse culturale personale") e – dall'altra – all'aumento dei procedimenti di riconoscimento del possesso della cittadinanza per filiazione '*ius sanguinis*' agli stranieri discendenti da un avo italiano emigrato in Paesi dove vige lo '*ius soli*', procedimenti che necessitano dell'accertamento del filo genealogico fino all'ultimo antenato con cittadinanza italiana.

Per venire incontro a un così sensibile e diversificato aumento delle richieste di ricerca, l'Istituto ha progressivamente messo in campo differenti strumenti, graduati sulle varie esigenze e sulle disuguali capacità dell'utenza di orientarsi tra le fonti. Il progetto indubbiamente più ambizioso e complesso è rappresentato dalla realizzazione di una banca dati delle liste di leva, ottenuta tramite il recupero dei dati anagrafici presenti appunto nelle liste di leva, compilate dai comuni della Provincia di Venezia e versate, per il tramite dell'Ufficio di leva, all'Archivio di Stato, una fonte indispensabile per l'avvio di qualsiasi ricerca di natura biografica per i cittadini maschi nati a partire dal 1825, soprattutto quando manchino i corrispondenti registri anagrafici comunali o i registri parrocchiali (basti ricordare che il Veneto, soprattutto durante la Prima guerra mondiale, ha subito ingenti danni al patrimonio archivistico a causa dei bombardamenti austriaci, trovandosi molti centri abitati sulla linea del fronte italiano).

Sul versante della predisposizione di strumenti di descrizione tradizionali, l'Istituto ha inoltre rilasciato nel settembre 2019 l'inventario aggiornato della serie dei registri di leva: rispetto allo strumento pubblicato nel 2010 (che comprendeva le classi di nascita fino al 1911), l'inventario del 2019 giunge fino alla classe di nascita 1948, assicurando di fatto la completa conoscenza e la totale disponibilità alla consultazione della serie delle liste di leva conservata dall'Archivio di Stato, considerato che "le liste di leva e di estrazione sono versate settant'anni dopo l'anno di nascita della classe cui si riferiscono" (D.lgs. 22 gennaio 2004, n. 42, art. 41, c. 1 "Codice dei beni culturali e del paesaggio").

Più in generale, l'intera Amministrazione archivistica statale, al cui vertice è la Direzione generale archivi del Ministero della cultura, è da tempo impegnata nella valorizzazione e nel potenziamento della fruizione, anche a distanza, delle fonti genealogiche italiane: nel 2010 è stato così inaugurato il 'Portale Antenati' che consente all'utente di consultare gratuitamente le riproduzioni digitali dei registri dello stato civile e, più raramente, altri documenti di carattere genealogico e anagrafico, conservati presso i singoli archivi di Stato italiani; l'Archivio di Stato di Venezia ha aderito al progetto e sul portale è così possibile consultare i registri dello stato civile napoleonico prodotti tra il 1806 e il 1815.

A completare il quadro di questo primo indirizzo di attività concorre la recente presentazione, da parte dell'Istituto, del progetto 'Viaggio nelle radici: alla scoperta delle fonti della genealogia', che si colloca sulla scia del più ampio intervento promosso dal Ministero degli affari esteri e della

cooperazione internazionale sul 'Turismo delle Radici', che ha visto finora la pubblicazione del primo volume della 'Guida alle radici italiane. Un viaggio sulle tracce dei tuoi antenati'.

Il secondo filone, nel quale si colloca il progetto presentato in questo capitolo, interessa invece le fonti per la storia delle famiglie nobiliari e comprende il recupero, il riordinamento e la descrizione degli archivi di enti pubblici e privati che, in vario modo, interessano quelle scienze sussidiarie della storia che sono l'araldica, l'onomastica e la genealogia.

L'ultimo significativo intervento in tal senso era stato condotto sul finire del XIX secolo dall'archivista Pietro Bosmin, che aveva curato l'indice dell'archivio della Commissione araldica austriaca, un complesso risalente al periodo 1816-1828 e che testimonia dell'attività dell'organismo istituito con notificazione del 28 dicembre 1815, durante il Regno Lombardo-Veneto (1814-1848), per l'esame e l'eventuale riconoscimento dei titoli nobiliari di tutto l'ex Stato veneto (Veneto, Friuli, Istria, Dalmazia, Albania, Brescia, Bergamo e Crema).

In ideale collegamento con l'impresa ottocentesca, più recentemente è stata condotta una campagna di censimento dei fondi di natura araldico-genealogica ancora in disordine o privi di strumenti di descrizione al fine di sottoporli ad un'analisi storico-critica per ricostruirne e ripristinarne l'ordine, laddove fosse stato alterato, e predisporre i mezzi di corredo indispensabili alla consultazione da parte dei ricercatori.

Nel corso del 2018 è stato così prodotto l'inventario dell'archivio dello Studio araldico genealogico Giovanni De Pellegrini in Venezia, un'impresa nata nel 1882 e attiva fino al 1916 specializzata nella ricerca araldico-genealogica e che ha lasciato una rara e preziosa collezione di stemmi di Stati e di Comuni, di famiglie nobili, di famiglie cittadinesche veneziane e di famiglie patrizie veneziane.

Sulla medesima scia, il lavoro di riproduzione, trascrizione e analisi automatica del contenuto del corpus dei verbali delle riunioni della Commissione araldica veneta oggetto del capitolo ha lo scopo di valorizzare tale fonte archivistica, sia nei confronti dell'utenza dell'Archivio, offrendo agli storici uno strumento di orientamento e di studio innovativo, sia nei confronti della pubblica amministrazione, fornendo la possibilità di coniugare in maniera innovativa le funzioni di conservazione, fruizione e valorizzazione del patrimonio documentario.

Il capitolo è organizzato in due sezioni principali. Nella prima sezione è condotto un breve *excursus* storico sulla Commissione araldica veneta. Nella seconda sezione sono invece descritte le fonti documentali oggetto

di analisi testuale automatica con applicazione di diverse tecniche (da un'analisi del *network*, ad un'analisi delle corrispondenze e ad una *topic detection* tramite metodo Reinert). L'ultimo paragrafo è dedicato alle considerazioni conclusive.

Nota storico-istituzionale e archivistica sulla Commissione araldica veneta

La vicenda storica della Commissione araldica veneta

Il regio decreto (r.d.) 15 giugno 1889 (pubblicato nella Gazzetta ufficiale del Regno d'Italia n. 174 del 23 luglio 1889), col quale fu approvato il regolamento per le iscrizioni d'ufficio nei registri della Consulta araldica, introdusse l'istituto degli elenchi regionali (art. 4) da formare per ciascuna delle regioni storiche italiane (art. 5). Tale formazione fu demandata a commissioni locali, che si sarebbero radunate presso le prefetture o gli archivi di Stato (art. 9), nominate dal Presidente del Consiglio, udita la Consulta araldica, e composte da membri della stessa Consulta, funzionari giudiziari, archivisti, studiosi di storia e legislazione nobiliare, rappresentanti del patriziato locale.

La Commissione locale per la compilazione degli elenchi relativi alla Regione veneta fu formata con Decreto del Presidente del Consiglio dei Ministri dell'8 novembre 1889 con la seguente composizione: Nicolò Barozzi, Guglielmo Berchet, Andrea Marcello, Girolamo Soranzo, Federico Stefani e Lodovico Valmarana.

Il r.d. 5 marzo 1891, n. 115 stabilì che le commissioni deputate alla formazione degli elenchi regionali sarebbero divenute permanenti, con il compito anche di dare pareri sulle materie riguardanti la legislazione e la materia nobiliare del proprio rispettivo territorio, su richiesta del Ministro, della Consulta o del Regio Commissario.

La Commissione araldica veneta cessò di fatto le proprie attività con il referendum istituzionale del 2 giugno 1946 (in quell'anno terminarono le registrazioni nel protocollo della corrispondenza) e formalmente con l'entrata in vigore della Costituzione repubblicana (in base alla 14a disposizione transitoria della Carta fondamentale "i titoli nobiliari non sono riconosciuti"), quantunque l'ultima riunione verbalizzata dell'organo risalga al 12 dicembre 1942.

Sedimentazione e conservazione delle carte prodotte dalla Commissione

La Commissione araldica veneta ebbe stabile sede presso l'Archivio di Stato di Venezia e, per tale ragione, è più che ragionevole ipotizzare che il complesso delle carte prodotte nel corso della propria attività sia stato ininterrottamente custodito ai Frari, dove è ancora oggi conservato.

L'archivio della Commissione, invero di dimensioni medio-piccole se confrontato con il fondo dell'analoga Commissione araldica di periodo austriaco, è costituito principalmente:

- dalla corrispondenza ordinata secondo un piano di classificazione (12 buste) e dai relativi protocolli e rubriche (8 registri);
- dalle pratiche di riconoscimento aperte su istanza dei nobili interessati all'iscrizione nell'elenco (42 buste e relativo indice a schede);
- dalla raccolta ufficiale degli stemmi prodotti dalle famiglie nobili del Veneto (15 volumi circa);
- da una piccola biblioteca di opere a stampa a uso dei commissari.

I verbali della Commissione araldica veneta

In questo paragrafo è presentato il lavoro di riproduzione, trascrizione e analisi automatica del contenuto del corpus dei verbali delle riunioni della Commissione araldica veneta.

A seguito della descrizione dei verbali e delle fasi di riproduzione e trascrizione dei testi, saranno illustrate un'analisi lessicale dei verbali delle riunioni della Commissione (che ha visto la ricerca dei *network* delle parole contenute nel corpus), una lettura dell'evoluzione temporale dei campi di interesse e di attività dell'ente e una misurazione della distanza tra i diversi periodi presi in considerazione (tramite un'analisi delle corrispondenze). Infine, saranno discussi i *topics* rappresentativi delle aree semantiche presenti nei testi analizzati (identificati con l'uso del metodo Reinert).

Il corpus dei verbali sottoposto ad analisi del contenuto rappresenta soltanto una porzione (e, incidentalmente, un campione) del più ampio complesso delle registrazioni delle sedute della Commissione araldica veneta, che operò ininterrottamente dal 1889 al 1942: i verbali si riferiscono difatti solamente ai due quadrienni 1889-1893 e 1938-1942, vale a dire, rispettivamente, ai primi e agli ultimi quattro anni di attività documentata

della Commissione.

La scelta dei testi con i quali comporre il corpus è stata, per certi aspetti, obbligata: attualmente, difatti, i verbali dell'ente utilizzati sono gli unici disponibili. Non è escluso, tuttavia, che nel corso dei lavori di riordinamento del fondo archivistico in corso di esecuzione altri verbali possano essere recuperati alla consultazione.

Il nucleo noto consta, complessivamente, di 101 documenti, e per l'esattezza 91 verbali prodotti nel corso del primo quadriennio (1889-1893) e 10 verbali relativi alle riunioni tenute nel secondo quadriennio (1938-1942).

I verbali sono numerati da 1 a 91 per il primo quadriennio e da 1 a 10 per il secondo, e le due serie di registrazioni risultano integre e complete.

La grande differenza numerica tra la prima e la seconda serie di registrazioni, effettuate in archi cronologici pressoché identici, è determinata dalla frequenza delle riunioni della Commissione:

1) nel corso del primo quadriennio l'ente si riunì mediamente ogni 16 giorni, arrivando a convocare i suoi membri ogni 7 giorni (così come previsto dalla stessa Commissione in una sua deliberazione registrata nel verbale n. 3 del 13 marzo 1890, che fissava nel giovedì alle due pomeridiane il giorno e l'ora degli incontri). Sono tuttavia presenti tre significative lunghe interruzioni dei lavori, ossia dal 4 settembre al 4 dicembre 1890 (91 giorni), dal 27 agosto 1891 al 28 gennaio 1892 (154 giorni) e dal 25 agosto al 15 dicembre 1892 (112 giorni), coincidenti con quelle che venivano allora definite le 'vacanze autunnali';
2) nell'ultimo quadriennio le sedute si diradarono sensibilmente, fino a raggiungere una distanza media di 195 giorni l'una dall'altra. In questo quadriennio, la Commissione si riunì da 1 a 3 volte l'anno, ma tra l'incontro del 12 dicembre 1938 e quello del 15 dicembre 1939 corsero ben 368 giorni.

È tuttavia da rilevare come, a fronte di ben 91 verbali per il primo quadriennio rispetto a soli 10 per l'ultimo quadriennio, le due diverse serie di registrazioni si equivalgano in termini di occorrenze presenti nei rispettivi vocabolari, con una leggera superiorità, anzi, della seconda serie sulla prima: nel corpus preso in esame, che registra nel complesso 93.065 *word tokens*, le registrazioni del periodo 1889-1893 ricomprendono 44.893 parole (con una media, quindi, di 493 parole per verbale, con una oscillazione dalle 86 parole del verbale n. 15 alle 2.026 parole del verbale n. 65), mentre quelle del periodo 1938-1942 includono 48.172 parole (con

una media di 4.817 parole per verbale, con punte estreme di sole 1.111 parole nel verbale n. 1 e ben 10.599 parole nel verbale n. 7).

Le ragioni di una tale difformità lessicale sono da ricercarsi nel mutamento delle finalità della Commissione dall'uno all'altro arco temporale, e nel conseguente radicale aggiornamento delle proprie modalità di lavoro.

Dalla loro istituzione con r.d. 15 giugno 1889, e fino all'emanazione del nuovo ordinamento e relativo regolamento della Consulta araldica (avvenuta coi due rr.dd. n. 313 e 314 del 2 e 5 luglio 1896), le commissioni araldiche furono fondamentalmente incaricate della sola redazione degli elenchi regionali delle famiglie italiane in legittimo possesso di titoli nobiliari.

Quello della Commissione araldica veneta nella prima fase della sua esistenza si presenta dunque per lo più come un lavoro di natura compilatoria: l'organismo avrebbe dovuto difatti stilare un nuovo elenco aggiornato a partire dall'ultimo repertorio ufficiale delle famiglie nobili del Veneto, pubblicato nel 1841. Quest'ultimo strumento, redatto quasi cinquant'anni prima rispetto all'istituzione della Commissione, fu nel corso delle frequenti riunioni sistematicamente ragguagliato e integrato attraverso lo spoglio della documentazione prodotta dalla Commissione araldica austriaca successivamente al 1841 e fino al 1866, e degli elenchi trasmessi dal Ministero dell'interno relativi alle famiglie già iscritte presso la Consulta araldica.

Dal 1896, con l'istituzione, accanto agli elenchi nobiliari regionali, del Libro d'oro della nobiltà italiana, l'impegno richiesto alla Consulta araldica e, per riflesso, alle commissioni araldiche regionali, crebbe in maniera rilevante. Come meglio specificato nel successivo r.d. 21 gennaio 1929, n. 61, che approvò l'Ordinamento dello stato nobiliare italiano, se nell'elenco sono riportati solamente "i nomi e cognomi per ordine alfabetico di tutte le persone che si trovano nel legittimo e riconosciuto possesso di titoli e attributi nobiliari" (art. 102), nel Libro d'oro sono registrate "le famiglie italiane che ottennero la concessione, la rinnovazione, l'autorizzazione o il riconoscimento di titoli e attributi nobiliari" (art. 98). Dall'iscrizione nel Libro d'oro sarebbero dunque risultati il Paese d'origine, la dimora abituale, i titoli e gli attributi nobiliari con le indicazioni di provenienza e di trasmissibilità, i provvedimenti regi o governativi, la descrizione dello stemma e la parte di genealogia che veniva documentata.

Ogni istanza avanzata alla Consulta araldica per l'iscrizione nel Libro d'oro doveva essere ora trasmessa alla competente commissione araldica regionale (art. 130), che di norma la sottoponeva "all'esame di uno o più

commissari per farne speciale relazione" (art. 94), per restituirla poi entro due mesi dal ricevimento corredata da un proprio parere.

Tali nuove disposizioni richiesero un'importante revisione delle modalità di lavoro della Commissione araldica veneta: le numerose domande, distribuite più o meno equamente tra i commissari, esigevano un approfondimento e uno studio che non potevano essere condotti nell'arco di pochi giorni o settimane; spesso era necessaria la consultazione di antica documentazione conservata presso l'Archivio di Stato di Venezia prodotta dagli organi di governo e dalle magistrature feudali della Repubblica veneta, o di manoscritti e pubblicazioni a stampa disseminati tra le più importanti biblioteche civiche della Regione.

L'analisi dei commissari era rivolta essenzialmente alla verifica della regolarità, della completezza e della coerenza dei documenti presentati rispetto all'oggetto della domanda (secondo l'art. 111 dell'Ordinamento dello stato nobiliare italiano, alla domanda dovevano essere allegati "la documentazione della esistenza dei titoli, predicati o stemmi e quella dell'attacco fra il richiedente e il concessionario e l'ultimo investito o riconosciuto, la dimostrazione per linea e grado del diritto di succedere nel titolo, nonché il diploma di concessione o di conferma e lo stemma a colori con la descrizione in termini araldici"), ma le loro relazioni, soprattutto nei casi più complessi e articolati, assumevano spesso la veste di brevi trattati di storia araldico-genealogica sulla famiglia candidatasi all'iscrizione al Libro d'oro. Tutto ciò determinò la dilatazione dei tempi delle convocazioni che, da "una volta ogni bimestre" (come previsto dall'art. 93 dell'Ordinamento dello stato nobiliare italiano), per la Commissione araldica veneta arrivarono a verificarsi all'incirca una volta ogni semestre.

Infine, tutti i verbali, sia quelli appartenenti alla prima che alla seconda serie, in ossequio ad una prassi consolidata, riportano nel protocollo documentale il numero di repertorio, il luogo, la data e l'ora, nonché l'elenco dei commissari presenti e degli assenti giustificati; solamente i verbali della seconda serie riferiscono, in forma estremamente essenziale, l'ordine del giorno, composto quasi sempre dai soli due punti "approvazione del verbale della seduta precedente" e "relazioni dei commissari" (a eccezione dei verbali 4, 5, 6 e 7, che al primo punto prevedono "comunicazioni del Presidente" in merito, rispettivamente, alla nomina del nuovo direttore dell'Archivio di Stato di Venezia, alle dimissioni del commissario Mario Nani Mocenigo, alla commemorazione del defunto commissario Girolamo Marcello e alla nomina del nuovo commissario Alessandro Marcello). Nell'escatocollo dei verbali è inoltre indicato l'orario di chiusura della seduta, seguito dalle sottoscrizioni del Presidente e del Segretario.

Per quanto riguarda il supporto scrittorio, è interessante rilevare che dei primi 60 verbali (23 dicembre 1889-14 luglio 1892) ci sono pervenute esclusivamente le copie litografiche, ottenute ossia mediante stampa con un particolare tipo di pietra levigata e poi incisa o disegnata con un segno grasso in grado di trattenere l'inchiostro. Le copie furono eseguite a Roma dalla Consulta araldica per essere successivamente trasmesse (a mo' di lettere circolari) a tutte le commissioni regionali, compresa quella Veneta. Dal verbale n. 82 del 27 luglio 1893 si apprende che, a partire da quella data, la Consulta determinò di cessare l'attività di confezione e spedizione delle copie litografate: il divario di un anno tra l'ultimo verbale litografato disponibile della Commissione araldica veneta e l'interruzione della pratica sopra descritta fa ritenere che la Commissione si trovasse in forte ritardo nella trasmissione alla Consulta degli originali dei resoconti delle proprie riunioni.

Dei verbali 61-91 (28 luglio 1892-7 dicembre 1893) si sono invece conservati gli originali manoscritti a penna in bi-fogli sciolti.

In conclusione, da un punto di vista squisitamente archivistico, i verbali 1-91 sono collocati all'interno della serie Corrispondenza, classificati sotto il titolo III ("Atti e pubblicazioni"), rubrica 1 ("Copia verbali delle adunanze della Commissione veneta"). I verbali 1-10 (19 febbraio 1938-14 dicembre 1942) sono pervenuti infine sempre sotto forma di originali manoscritti, annotati però in registro e non in fogli sciolti.

Acquisizione e normalizzazione del testo dei verbali

La creazione del corpus ha implicato una lunga, complessa e minuta attività di preparazione dei documenti all'analisi automatica.

Per l'acquisizione del contenuto dei verbali è stato difatti necessario procedere alla trascrizione del testo mediante un programma di video scrittura.

Tale operazione, che ha rappresentato l'occasione per condurre un indispensabile processo di normalizzazione del testo, rappresenta una pratica familiare ai cultori di discipline umanistiche come la filologia e la paleografia. Tuttavia, in considerazione delle specifiche finalità della trascrizione (non un'edizione di fonte ma una *content analysis*), a interventi tradizionali quali lo scioglimento degli acronimi e delle abbreviazioni, il controllo di apostrofi e accenti, la modernizzazione dell'interpunzione e dell'uso delle maiuscole e delle minuscole, l'armonizzazione delle scritture (interventi che potrebbero essere compresi sotto la comune definizione

di *punctuation control*), si è affiancato un processo di *tokenization*, che si è essenzialmente concentrato sull'individuazione e la demarcazione di parole multiple chiave (*multi-word expressions*) tramite l'utilizzo del separatore "_" (trattino basso o *underscore*) al posto dello spazio libero.

Al fine di ridurre le varianti morfologiche individuando le forme base è stata inoltre condotta in maniera automatica, con l'ausilio del software, un'operazione di lemmatizzazione.

Inoltre:

- è stato evitato l'uso del grassetto e del corsivo nella formattazione del testo;
- sono stati eliminati i trattini presenti originariamente nel testo e sostituiti con trattini bassi (nel caso in cui le parole unite dal trattino siano state considerate *multi-word expressions* e dunque meritevoli di essere sottoposte ad analisi specifica) oppure con virgole (quando il trattino costitutiva un mero segno di interpunzione);
- tutti i verbi con pronomi personali sono stati sciolti e il pronome personale è stato posto in posizione proclitica (ad es. "apresi la seduta" in "si apre la seduta"; "deve attenersi" in "si deve attenere");
- tutti i numeri sono stati resi in cifra;
- sono stati eliminati gli apostrofi, le virgolette e i puntini di sospensione.

Il corpus è stato infine classificato in base ad alcune variabili e relative modalità, utili ai fini dell'analisi: il periodo di riferimento dei verbali ("Serie_Prima", che comprende i verbali prodotti nel periodo 1889-1893, e "Serie_Seconda", che aggrega i verbali risalenti al quadriennio 1938-1942); la data ("data") e l'originale numero progressivo attribuito ai verbali.

Osservare il complesso

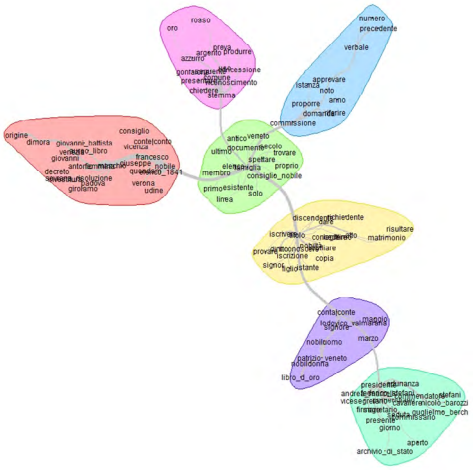
Una prima lettura d'insieme del corpus oggetto di esame è stata attuata per mezzo di un'analisi delle relazioni tra le forme più frequenti nell'intero corpus (Fig. 1) eseguita tramite il software Iramuteq, che presenta sotto forma di *network* di relazioni le diverse forme (selezionate in questo caso tra quelle che compaiono almeno 70 volte nel corpus) e i legami tra loro. Nella figura le forme appaiono raggruppate entro nuvole colorate sulla base delle similarità tra i segmenti di testo analizzati, consentendo di cogliere, nel corpus analizzato – caratterizzato da una forte rilevanza

del tema temporale, trattando due serie storiche di verbali – l'evoluzione delle diverse attività svolte dalla Commissione araldica veneta nel corso della propria storia pluridecennale e, allo stesso tempo, di rilevare come il vocabolario possa specializzarsi e differenziarsi in modo da provvedere alle esigenze di 'documentalità' da una parte (per produrre un oggetto non è sufficiente che un atto sia proferito, è necessario che sia scritto e registrato, da qui le formalità del verbale) e di 'funzionalità' dall'altra.

Così, alle estremità del grafico in Fig. 1 si collocano due insiemi di parole più marcatamente caratterizzati: nella parte bassa, quello connotato dalle forme riconducibili alla specifica attività di registrazione e illustrazione dello svolgimento delle sedute, scandite dall'intervento dei vari commissari dei quali vengono enunciate le rispettive cariche, e nella parte alta, il tema specifico della descrizione araldica degli stemmi con i loro smalti e dei gonfaloni con i loro colori. Al centro, all'interno della nuvola verde, il lemma "famiglia", che fa da collante tra le due estremità e rappresenta la radice dell'albero, rispecchiando in tutto e per tutto il principale mandato, e dunque la ragion d'essere, dell'istituzione. È difatti nella ricostruzione e nella valutazione dei titoli nobiliari gentilizi, narranti la storia delle famiglie, che si sostanzia il compito fondamentale della Commissione, così come dichiarato fin dall'inaugurazione dei suoi lavori e registrato nel primo verbale del 23 dicembre 1889:

> la «compilazione degli elenchi nobiliari [...] è un compito molto utile servendo alla storia delle famiglie che è la storia della Patria, e al prestigio della Monarchia la quale ama circondarsi delle rappresentanze sincere delle glorie italiane».

Fig. 1 Analisi del *network*.



Individuare gli insiemi

Al corpus dei verbali è stata successivamente applicata un'analisi delle corrispondenze, con l'uso del software Iramuteq, rispetto alla variabile temporale "Serie", allo scopo di individuare somiglianze o differenze nel lessico caratterizzante le due serie di verbali. Quest'operazione è stata eseguita, infatti, a partire dall'ipotesi che le due serie di verbali corrispondano a due insiemi lessicalmente e tematicamente ben distinti.

La letteratura di settore infatti da sempre individua nell'anno 1896 una soluzione di continuità nella vicenda istituzionale e nelle prerogative delle commissioni araldiche regionali: la transizione da soggetti investiti della sola compilazione – sulla base di repertori già esistenti – degli elenchi regionali delle famiglie italiane in legittimo possesso di titoli nobiliari a protagonisti attivi nella redazione del nuovo Libro d'oro della nobiltà italiana deve avere inevitabilmente lasciato tracce profonde nella documentazione prodotta nel corso delle due diverse fasi di vita di questa particolare tipologia di enti consultivi.

Le tue tabelle che seguono (Tab. 1 e 2) riportano un estratto delle principali forme attive, selezionate tra quelle con almeno 50 frequenze, per ciascuna delle due serie di verbali.

Nello specifico, la Tab. 1 mostra le forme attive dei verbali appartenenti alla prima serie (relativa agli anni 1889-1893), con l'indicazione, nell'ultima colonna, del confronto in termini di frequenza delle medesime forme con la seconda serie.

Tab. 1 Principali forme attive della prima serie di verbali e confronto con la seconda serie.

Forma attiva	Serie Prima	Serie Seconda
Maschio	693	95
Nobile	588	214
Femmina	500	56
Elenco 1841	441	0
Famiglia	432	283
Commendatore	354	31
Nobiluomo	334	63
Presidente	325	60
Venezia	314	36
Titolo	308	335
Conte	308	181
Commissione	305	216
Elenco	167	11
Origine	166	17
Dimora	160	2

Verbale	159	33
Libro d'oro	156	72
Nobildonna	152	14
Signore	147	64
Seduta	142	44
Adunanza	139	12
Patrizio veneto	134	23
Approvare	126	42
Pomeridiano	123	0
Riconoscere	119	110
Presente	110	31
Vicesegretario	107	1
Iscrivere	97	110
Cavaliere	96	10
Nobiltà	79	125
Atto	75	214
Aureo libro	72	0
Segretario	71	25
Presentare	67	143
Discendente	65	43
Consueto	63	1
Veneto	63	33
Commissario del Re	63	6
Nobiliare	63	54
Iscrizione	62	108
Aperto	62	10
Casato	61	6
Archivio di Stato	61	14
Sovrana risoluzione	60	16
Investitura	59	39
Commissario	59	45
Comunicare	58	2
Residenza	57	2
Osservazione	57	3
Linea	55	25
Decreto	55	16
Oggi	54	3
Consulta araldica	54	14
Ascrizione	53	5
Provincia	53	16
Documento	52	113
Estinto	52	5
Proposta	50	11

Ciò che spicca già a un'osservazione macroscopica è la presenza, nelle prime 5 posizioni, di forme con valori di frequenza superiori a 400, e riconducibili alla principale attività della Commissione araldica nel suo primo quadriennio di attività, ossia la formazione dell'elenco delle famiglie

italiane residenti nella regione storica della Venezia in legittimo possesso di titoli nobiliari: "nobile maschi e femmine (Elenco 1841)" costituisce difatti la formula canonica che nei verbali segue il nome della famiglia che viene registrato nell'elenco in formazione, principalmente a seguito dello spoglio dell'ultimo elenco ufficiale austriaco (Elenco dei nobili e titolati delle venete provincie, Venezia, Per Francesco Andreola Tipografo, 1841). L'assenza dell'espressione "Elenco 1841" nella seconda serie di verbali, redatta a cavallo tra gli anni Trenta e Quaranta del Novecento, trova la sua ragione nella tendenza della Commissione araldica a rifarsi a quell'epoca oramai solo al nuovo elenco ufficiale della nobiltà italiana del Regno d'Italia.

Variamente legati a questa principale attività redazionale, che catalizzò quasi completamente le forze della Commissione, sono i lemmi "nobiluomo"/"nobildonna", "titolo", "conte", "elenco", "origine", "dimora", "Libro d'oro"/"Aureo libro", "patrizio veneto", "riconoscere", "iscrivere"/"iscrizione"/"ascrizione", "nobiltà"/"nobiliare", "presentare", "discendente", "casato", "sovrana risoluzione", "investitura", "osservazione", "linea", "decreto", "provincia", "estinto".

Emerge poi un notevole numero di parole specificamente legate al funzionamento interno della Commissione araldica: termini come "commendatore"/"cavaliere"/"signore", "presidente"/"segretario"/"vicesegretario", "commissione"/"commissario", "verbale", "seduta"/"adunanza", "approvare", "pomeridiano", "presente", "consueto", "aperto", "Archivio di Stato", "comunicare", "residenza", "oggi", "documento", "proposta", testimoniano delle formalità insite nella natura stessa del verbale come tipologia documentaria che esige una particolare ritualità legata alla registrazione dei membri dell'organo presenti nel luogo deputato, all'accertamento del numero legale, alla presentazione e alla discussione di mozioni, ecc.

I verbali della prima serie risentono inoltre di una maggiore rigidità stilistica e contenutistica rispetto alle successive registrazioni della seconda serie, una rigidità che, associata a una generale minore consistenza dei singoli testi rispetto a quelli che costituiscono il *subcorpus* novecentesco, contribuisce a presentare la maggior parte dei verbali ottocenteschi come piccole costruzioni sempre uguali, divergenti solo per i lunghi elenchi di famiglie iscritte di volta in volta nell'elenco regionale.

Nella Tab. 2 sono elencate le forme attive presenti con almeno 50 frequenze nei verbali appartenenti alla seconda serie (relativa agli anni 1938-1942), con l'indicazione, nell'ultima colonna, del confronto in termini di frequenza delle stesse forme con la prima serie.

Tab. 2 Principali forme attive della seconda serie di verbali e confronto con la prima serie.

Forma attiva	Serie seconda	Serie prima
Stemma	463	17
Titolo	335	308
Famiglia	283	432
Commissione	216	305
Nobile	214	588
Atto	214	75
Conte	181	308
Riconoscimento	162	36
Comune	162	16
Presentare	143	67
Chiedere	141	21
Azzurro	141	0
Nobiltà	125	79
Argento	120	0
Figlio	118	23
Istante	114	0
Documento	113	52
Rosso	112	1
Copia	110	10
Riconoscere	110	119
Iscrivere	110	97
Iscrizione	108	62
Domanda	104	36
Provare	101	13
Proporre	99	15
Oro	97	0
Maschio	95	693
Diritto	95	49
Uso	93	21
Matrimonio	91	20
Prova	86	6
Gonfalone	83	0
Concessione	82	16
Legittimo	72	24
Produrre	72	7
Libro d'oro	72	156
Istanza	70	2
Richiedente	69	1
Risultare	68	10
Signore	64	147
Riferire	63	23
Nobiluomo	63	334
Campo	63	3
Naturale	62	5

Figura	60	0
Libro d'oro della nobiltà italiana	60	2
Presidente	60	325
Fratello	58	6
Nero	58	0
Femmina	56	500
Antico	54	29
Podestà	54	0
Nobiliare	54	63
Relatore	51	4
Secolo	51	31
Nascita	50	10
Ricordare	50	19
Spettare	50	19
Verde	50	0

La specificità del *subcorpus* in esame emerge soprattutto dal confronto tra le due serie di registrazioni: dei verbali novecenteschi, difatti, sono quasi esclusivi termini come "stemma", "comune", "azzurro", "argento", "rosso", "oro", "gonfalone", "campo", "naturale", "figura", "nero", "podestà", "verde", legati all'attività consultiva della quale furono investite le commissioni regionali, a partire dal 1896 e più ancora dal 1929, per il riconoscimento o la concessione del blasone di famiglie ed enti pubblici. All'iscrizione al nuovo "Libro d'oro della nobiltà italiana" e al riconoscimento dei titoli nobiliari sono connessi lemmi quali "figlio", "diritto", "uso", "matrimonio", "prova", "legittimo", "fratello" e "nascita".

Più in generale, verbi come "presentare", "chiedere", "provare", "proporre", "produrre", "risultare", "ricordare" e "spettare", nonché sostantivi come "atto", "riconoscimento", "istante", "documento", "copia", "domanda", "concessione", "richiedente" e "istanza", testimoniano di una modalità di lavoro radicalmente mutata rispetto ai primi anni di attività delle commissioni regionali, che dall'iscrizione d'ufficio in un nuovo elenco che si andava componendo di tutte le famiglie in legittimo possesso di titoli nobiliari (attività svolta per lo più mediante lo spoglio di precedenti repertori), passarono all'approfondito e meticoloso esame (condotto da relatori appositamente incaricati) delle istanze per l'inserimento nel Libro d'oro della nobiltà italiana avanzate dalle famiglie alla Consulta araldica e trasmesse a Venezia per competenza territoriale.

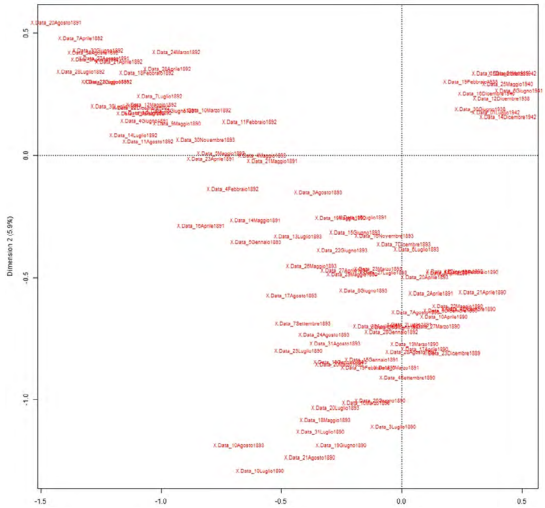
Una seconda analisi delle corrispondenze (Fig. 2), condotta questa volta rispetto alla data di riferimento dei verbali (variabile "Data"), ha ulteriormente confermato l'ipotesi sopra formulata, vale a dire l'esistenza

di due insiemi lessicalmente e tematicamente distinti: sugli assi cartesiani del grafico generato dal software le etichette si concentrano difatti in due ammassi molto ben caratterizzati, il minore dei quali (nel quadrante in alto a destra, decisamente circoscritto e compatto) è costituito dagli anni compresi tra il 1938 e il 1942, mentre il maggiore (chiaramente più ampio e frastagliato, dall'andamento quasi sinuoso) è formato dagli anni racchiusi tra il 1889 e il 1893.

Da un'osservazione più accurata dell'agglomerato principale pare possa cogliersi una disposizione delle etichette indicativamente cronologica, con le date più risalenti collocate nella parte alta e quelle più recenti nella parte bassa della scia. Emerge nondimeno che i primi verbali dell'ente, redatti ossia fino a tutta la primavera del 1890, occupano una posizione decisamente eccentrica della curva, posizionandosi in quella che si potrebbe definire la parte esterna del "gomito": si tratta difatti delle registrazioni delle sedute preparatorie dei futuri lavori della Commissione, durante le quali i membri definivano i criteri generali per la redazione dell'elenco nobiliare, anche alla luce delle indicazioni che venivano loro fornite dal Commissario del Re presso la Consulta araldica, e mancano dunque del tutto della caratteristica che accomuna tutti gli altri verbali della prima serie, ossia l'enumerazione delle famiglie iscritte di volta in volta nel catalogo.

Si tratta di una conferma empirica che restituisce un ordine cronologico chiaro dal punto di vista dei profili lessicali per anno. Questo significa che temi e parole utilizzate differiscono notevolmente in base ai diversi periodi.

Fig. 2 Distribuzione sugli assi cartesiani della variabile "Data".



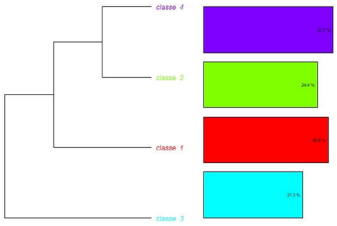
Gli argomenti dei verbali e le attività della commissione

Al corpus dei verbali è stata applicata, tramite l'uso del software Iramuteq, una *topic detection* basata sul metodo Reinert, volta all'individuazione dei *topics* caratterizzanti i verbali e alla ricerca, alla luce delle analisi eseguite nel paragrafo precedente, di come a temi specifici si colleghino distinte attività della Commissione.

L'analisi automatica è risultata nell'identificazione di 4 *cluster*.

Da una prima analisi del dendrogramma che illustra la suddivisione del corpus in classi e la percentuale delle unità testuali contenuta al loro interno (Fig. 3), si evince che i *topics* numero 2 e 4 sono i più simili tra loro, mentre una somiglianza meno marcata si può rilevare tra questi ultimi e il *topic* numero 1. Quantitativamente, i gruppi non presentano significative differenze: il *cluster* che contiene il maggior numero di segmenti di testo è il numero 4 (con il 27,7%), seguito dal numero 1 (con il 26,6%) e dal numero 2 (con il 24,4%). In coda si colloca il *cluster* numero 3 (con il 21,3%), che è anche il *topic* maggiormente distinto a livello tematico.

Fig. 3 Dendrogramma della classificazione dei verbali in 4 *topics*.



Come emerge dal grafico in Fig. 4 – che indica, rispetto alla precedente Fig. 3, anche i termini maggiormente significativi associati a ciascuna classe – i *cluster* numero 2 e 4 sono identificativi di quella parte di attività della Commissione araldica veneta legata al riconoscimento della nobiltà gentilizia e all'approvazione degli stemmi in uso presso le famiglie o gli enti (o alla concessione di nuovi).

In particolare, il *cluster* semantico numero 2 sembra più orientato a restituire la specifica procedura di ricognizione della documentazione di concessione o di conferma degli stemmi: parole come "stemma", "insegna" (parola esclusiva di questo *cluster* poiché ricorre solamente nei segmenti di testo che lo compongono), "corona", "sigillo", "elemento",

"fioroni", "bandiera" (forme molto ricorrenti), sono riconducibili alla descrizione degli ornamenti dei blasoni in termini araldici.

Il *cluster* numero 4, dal profilo molto ben marcato, è invece riconducibile all'attività di iscrizione delle famiglie nel Libro d'oro della nobiltà italiana che si svolgeva, come già più volte ricordato, non più d'ufficio ma su istanza di parte: tra le forme più frequenti ("provare", "istante", "copia", "legittimo", "matrimonio", "nascita", "stato civile", "risultare", "discendenza", "richiedente", "ascendente", "sovrana risoluzione austriaca"), alcune sono associate quasi esclusivamente a questa classe e rappresentano in maniera nitida il processo di presentazione della domanda di riconoscimento di titoli e predicati nobiliari, accompagnate dalle prove genealogiche (copie autentiche degli atti legali di nascita, di matrimonio e morte, grado per grado, di tutti gli individui compresi nella dimostrazione genealogica) utili a provare il diritto.

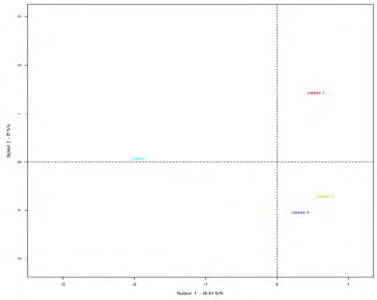
Il *cluster* numero 3, che, come si accennava, appartiene a un ramo differente rispetto agli altri, presenta una specifica connotazione legata alla prima fase di attività della Commissione araldica, quella ossia inerente la redazione del nuovo elenco nobiliare della regione storica della Venezia: "maschio", "femmina", "elenco 1841", "dimora", "origine", "nobildonna", "nobiluomo", "patrizio veneto", "Aureo libro", costituiscono forme quasi esclusivamente presenti in questo gruppo semantico e ne caratterizzano il dizionario, marcando sensibilmente la distanza tra questa e le altre classi, oltre che confermare, ancora una volta, le profonde differenze che intercorrono tra la Commissione delle origini e quella dell'epilogo.

Infine, il *cluster* numero 1 si colloca tra i due rami (quello relativo alle classi 2 e 4 da una parte, quello relativo alla classe 3 dall'altra), sia in termini grafici che tematici, vista la sua natura trasversale di contenitore legato al funzionamento interno della Commissione araldica, le cui riunioni sono oggetto delle registrazioni trasposte nei verbali analizzati. Il vocabolario è in questo gruppo costituito da forme come "presidente", "commissione", "approvare", "commendatore", "adunanza", "verbale", "precedente", "seduta", "comunicare", "commissario", "leggere" (in particolare la forma associata del participio passato "letto"), "proposta", "tornare" (lemmatizzazione del participio passato "tornata", da intendere come seduta), "delibera", "aperto", "discussione", "deliberazione", "voto", "prendere", ecc., che restituiscono il lavoro dell'organismo teso a dare forma giuridica, e dunque legittimità, alle proprie deliberazioni.

Fig. 4 Dendrogramma della *topic detection* con indicazione per ogni *cluster* delle forme più significative.

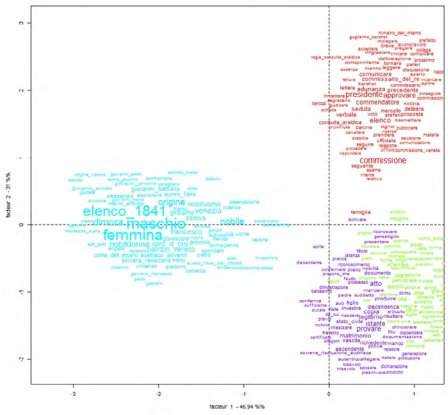
Le Fig. 5 e 6 mostrano i risultati di un'analisi delle corrispondenze applicata ai *cluster*, ossia il posizionamento dei *cluster* su un piano cartesiano volto a misurare la vicinanza o lontananza dei gruppi tematici precedentemente descritti. Dal grafico in Fig. 5 risalta in particolar modo come i *cluster* 2 e 4 diano vita ad un insieme separato dai restanti due (numero 3 e 1). In base all'analisi sopra esposta, si può ragionevolmente supporre che la linea di discriminazione sia, da una parte, lo spazio quarantennale che separa le due serie temporali di verbali (presentate nelle classi 2, 4 e 3), e, dall'altra, il tenore stesso delle registrazioni, interessate da ricorrenti formalismi (classe 1) contrapposti al merito delle trattazioni.

Fig. 5 Disposizione sugli assi cartesiani dei quattro *topics*.



Nella Fig. 6, che riporta, per ogni classe del grafico precedente, le parole più significative, è evidente che, mentre i *cluster* numero 1 e 3 vedono le loro forme nettamente discoste da quelle delle altre classi, nei *cluster* numero 2 e 4, accantonati e contermini, alcune parole tracimano le une nel campo delle altre.

Fig. 6 Composizione sugli assi cartesiani dei quattro *topics*.

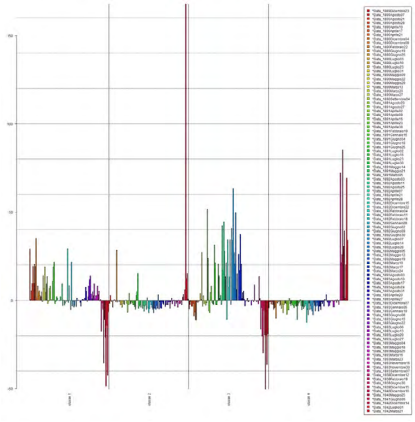


Rilevanza degli argomenti nel tempo

In ultimo, al fine di scoprire la rilevanza dei vari *topics* negli anni, è stata eseguita tramite il software un'analisi χ^2 del livello di associazione tra le 4 classi e la variabile "data", relativa alle date dei verbali.

Il grafico riportato in Fig. 7 illustra il legame che intercorre nello specifico tra i quattro *cluster* semantici individuati dal software e le due serie di verbali, evidenziando quali argomenti siano caratteristici di quali anni e, di conseguenza, la maggiore o minore associazione della singola data al *topic*.

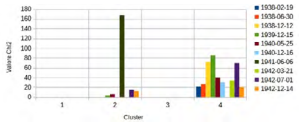
Fig. 7 Distribuzione della variabile "Data" nei quattro *cluster*.



Nonostante la scarsa leggibilità del grafico, dovuta all'estensione delle serie temporali considerate, è subito evidente come i verbali degli ultimi

anni (posizionati, in figura, sull'estremità destra di ciascuna delle quattro classi, con colori dal rosa scuro al rosso), relativi al quadriennio 1938-1942, siano fortemente caratterizzanti i *cluster* 2 e 4, nei quali a registrare valori positivi sono difatti quasi esclusivamente le date comprese tra il 1938 e il 1942, e scarsamente distintivi dei *cluster* 1 e 3, ossia come le classi numero 2 e 4 siano rappresentative dell'attività svolta dalla Commissione araldica veneta nel suo ultimo quadriennio. Si riporta in Fig. 8 un estratto dei valori di associazione del χ^2 delle classi con la variabile temporale per il solo periodo 1938-1942, corrispondente alla seconda serie dei verbali.

Fig. 8 Distribuzione della seconda serie di verbali (anni 1938-1942) nei quattro *cluster*.



Al contrario, le classi 1 e 3 vedono la preminenza, nella gamma dei valori positivi, degli anni compresi tra il 1889 e il 1893. Tale superiorità si manifesta tuttavia in maniera diseguale nei due *cluster*. Nello specifico, la classe numero 3 è associata agli anni del primo periodo di attività della Commissione con l'esclusione dei primi 15 verbali circa, ossia di quelli che, come è già stato detto esaminando la disposizione sugli assi cartesiani della variabile "Data" (cfr. Fig. 2 del presente capitolo), si differenziano nettamente dai successivi nel contenuto perché principalmente dedicati a pianificare e disciplinare i lavori futuri dell'ente. Il *cluster* numero 1, definito 'trasversale' ai due *subcorpora* per il vocabolario riferibile precipuamente al funzionamento interno della Commissione araldica, presenta invece una maggiore omogeneità nei valori, eccezion fatta per gli anni finali di attività, che precipitano ben al di sotto della linea dello zero: ciò è plausibilmente riconducibile alla consistenza globale della seconda serie di registrazioni (48.172 parole, distribuite in soli 10 testi), che determina una minore incidenza numerica nel *subcorpus* delle forme legate alle formule iniziali e finali e una loro concentrazione di gran lunga inferiore rispetto alla prima serie di verbali (composta da 44.893 parole distribuite in ben 91 testi).

Conclusioni

L'analisi testuale automatica applicata ai verbali prodotti dalla Commissione araldica veneta e conservati presso l'Archivio di Stato di Vene-

zia ha indubbiamente consentito di verificare la convergenza dei risultati dell'analisi automatica con le ipotesi avanzate in sede storiografica sulla base della lettura delle norme generali che nel tempo hanno disciplinato il funzionamento delle diverse commissioni araldiche regionali.

Sebbene abbia operato per poco meno di sessant'anni in un regime di formale continuità istituzionale, l'attività della Commissione araldica veneta è stata scandita in fasi nettamente differenti, contraddistinte da un mandato, una composizione e una prassi operativa sensibilmente mutati in base all'evoluzione della normativa in materia nobiliare e alle nuove urgenze che questa dettava.

L'analisi del contenuto operata nel corso del capitolo ha confermato il divario in termini di lessico e di tematiche tra i due periodi storici e i relativi testi presi in esame:

1) la prima serie di verbali testimonia di un organismo sostanzialmente introflesso, i cui ritmi di lavoro erano determinati dalla ricognizione sistematica e monocorde di antichi elenchi e registrazioni per il loro aggiornamento in un nuovo elenco nobiliare (attività che escludeva a priori la possibilità per la Commissione di dialogare con l'esterno, se non con il Commissario del Re presso la Consulta araldica o con i propri corrispondenti nelle province, e per i privati di interloquire con l'ente);
2) la seconda serie di verbali è al contrario la traccia eloquente della nuova condizione nella quale la Commissione si viene a trovare, una sorta di triangolo i cui vertici sono rappresentati dalla Consulta araldica, dalle commissioni regionali e dall'universo dei candidati (persone ed enti), intorno al riconoscimento o alla concessione del titolo o dello stemma, un circuito decisamente più aperto e dinamico, nel quale alle commissioni regionali è attribuito un ruolo consultivo dirimente nell'istruttoria delle pratiche.

L'auspicato recupero e la conseguente analisi anche dei verbali delle sedute svolte tra il 1894 e il 1937 potrebbe contribuire a colmare il consistente divario rilevato tra i due periodi esaminati, il passaggio tra i quali è risultato all'esame automatico decisamente marcato per via della distanza cronologica ultra decennale e della mutata temperie politica e culturale.

Alla luce dell'esperienza e dei risultati maturati nello studio, appare evidente come la valorizzazione di una fonte archivistica tramite tecniche quali quelle qui presentate, possa condurre ad un rafforzamento dell'offerta istituzionale, e ad un'effettiva sintesi delle funzioni di conservazio-

ne, fruizione e valorizzazione del patrimonio documentario dell'Archivio di Stato.

Sembra infine più che ragionevole considerare la possibilità di estendere l'applicazione del metodo di analisi automatica dei contenuti ad altri complessi documentari disponibili presso l'Archivio di Stato. Gli istituti di conservazione archivistica rappresentano difatti un potenziale di conoscenza senza eguali, ma il numero incalcolabile di informazioni che custodiscono è registrato in forma non strutturata. La sperimentazione di nuovi metodi di raccolta ed elaborazione di dati storici, sfruttando le possibilità offerte dal *machine learning* e dal *semantic web*, deve pertanto tenere in conto un lungo processo preliminare di preparazione applicato a una documentazione manoscritta, che non sempre le istituzioni sono in grado di sostenere direttamente, sia in termini economici che organizzativi.

L'applicazione del metodo di analisi automatica del contenuto a fonti antiche quali ad esempio la serie degli Annali della Repubblica di Venezia (1549-1719), recanti la registrazione delle materie considerate degne di essere ricordate, se da una parte potrebbe difatti restituire, per un lasso di tempo plurisecolare, l'evoluzione della sensibilità pubblica sul concetto di "memorabile", d'altra parte dovrà scontare le difficoltà costituite dalla grafia cinque-seicentesca, intelligibile solo a esperti paleografi, e dalla lingua, sì volgare ma non ancora canonizzata nell'italiano contemporaneo, e dunque sconosciuta ai vocabolari dei software di analisi, salvo specifici, complessi e costosi adattamenti, come il recupero retrospettivo di dizionari dialettali a stampa.

Non mancano tuttavia esempi virtuosi nei quali la collaborazione tra diversi soggetti (istituti di conservazione, centri e gruppi di ricerca universitari, fondazioni, ecc.) ha consentito di affrontare e superare gli aspetti critici legati al reperimento sia delle risorse economiche che delle competenze necessarie: nel dominio delle discipline umanistiche difatti i metodi e gli strumenti tradizionalmente a disposizione della pratica storica, filologica, linguistica, paleografica e archivistica hanno da tempo trovato applicazione nel trattamento digitale di fonti inedite, e sempre più numerosi sono i progetti di schedatura, trascrizione e analisi automatica di documentazione manoscritta in grado di rendere disponibili e analizzabili dati in serie, premessa indispensabile per più articolate ricerche di storia economica, sociale e giuridica.

Dalla pergamena al digitale. Conservazione e valorizzazione del patrimonio archivistico

Andrea Erboso[1]

Archivi, Diplomatica, Valorizzazione, Conservazione, Digital humanities.

Introduzione

La sfida principale nella gestione del patrimonio culturale italiano è quella di cercare un punto di equilibrio tra due istanze all'apparenza antitetiche: da un lato, le esigenze collegate alla conservazione dei beni culturali e quindi a tutti gli aspetti conoscitivi, descrittivi e di cura materiale necessari a preservare l'integrità del bene per le generazioni future, dall'altro, le esigenze legate alla valorizzazione del bene, cioè alla sua fruibilità da parte dei cittadini e all'estrinsecazione del suo potenziale culturale. Tale polarità, nata dapprima nella riflessione teorica, è assurta a norma con l'entrata in vigore del D.lgs. 22 gennaio 2004, n. 42 (D.lgs. 42/04), "Codice dei beni culturali e del paesaggio", che disciplina le due attività, affidando agli istituti e ai luoghi della cultura l'onere di armonizzare con questi principi le proprie attività.

Questo lavoro vuole offrire un contributo alla risoluzione di questo problema, utilizzando gli strumenti forniti dall'analisi testuale: la loro ap-

[1] Dal 2018 è funzionario archivista di Stato nell'Archivio di Stato di Venezia. Dallo stesso anno insegna diplomatica presso la Scuola di Archivistica, Paleografia e Diplomatica dell'Istituto.

plicazione alle fonti archivistiche può, infatti, rappresentare una strada per tutelare le condizioni materiali del patrimonio favorendo e aumentando, al contempo, la possibilità di fruizione e i servizi offerti dall'amministrazione archivistica.

L'Archivio di Stato di Venezia

L'Archivio di Stato di Venezia è un ufficio periferico del Ministero della cultura, definito *Istituto della cultura* dall'art. 101 del citato D.lgs. 42/04, la cui *mission* è conservare la documentazione prodotta dagli uffici periferici dello Stato italiano presenti nel territorio della Città metropolitana di Venezia (es. Prefettura, Questura, Tribunale ordinario, Corte d'appello, Procura della Repubblica, ecc.). Trascorsi 30 anni dall'esaurimento delle pratiche delle varie amministrazioni, il legislatore ritiene infatti esaurito l'interesse amministrativo della documentazione, a favore di un crescente interesse storico nei suoi confronti.

Negli Archivi di Stato si conserva, inoltre, la documentazione più antica afferente al territorio di riferimento. Nel caso di Venezia, capitale di uno stato preunitario, gli archivi prodotti dalla Serenissima rappresentano la parte più significativa del patrimonio, cui si aggiungono le carte provenienti dalle corporazioni religiose soppresse in epoca napoleonica e gli atti notarili.

Complessivamente, l'Archivio di Stato di Venezia conserva, nel solo deposito monumentale della sede principale di Santa Maria Gloriosa dei Frari, documentazione che copre un arco cronologico che va dal IX al XX secolo e occupa circa 80 km di scaffali.

La sala di studio dell'Istituto, dove viene erogato il servizio di consultazione del materiale archivistico, ospita una media di 1500 studiosi all'anno, di cui il 20% stranieri, e ogni giorno vengono consultati circa 100 pezzi archivistici che vengono prelevati dai depositi e consegnati direttamente all'utenza per lo studio [2].

Il servizio di consultazione rappresenta il fulcro delle attività dell'Istituto, ma anche il momento più delicato per la fragile documentazione archivistica. Il materiale, infatti, trova le migliori condizioni di conserva-

[2] I dati sulla presenza degli utenti e sulla consultazione del patrimonio provengono dalle rilevazioni statistiche del Ministero della cultura e dal software interno all'Istituto, Archimatic, che gestisce l'anagrafica degli ammessi alla sala di studio e permette loro di richiedere la consultazione della documentazione. Alcuni dati statistici generali sono presenti anche nel sito dell'Ufficio statistica del Ministero, all'indirizzo http://www.statistica.beniculturali.it/).

zione nei depositi che, seppur non dotati di un controllo di temperatura e umidità, preservano la carta e la pergamena da mutamenti repentini delle condizioni ambientali. Durante il tempo della consultazione, che può variare da alcune ore fino a dei mesi, a seconda del tipo di materiale e del tipo di ricerca, la documentazione è sottoposta allo *stress* dovuto al rapidissimo mutamento di temperatura e umidità che comporta il suo trasferimento dai depositi agli ambienti accessibili al pubblico e allo *stress* meccanico dovuto alla consultazione stessa da parte dell'utente.

La digitalizzazione del patrimonio: una via da seguire

Per essere consultati, i beni archivistici (assieme a quelli librari) vengono maneggiati direttamente dai fruitori, operazione che inevitabilmente ne compromette la conservazione e, di fatto, ne accorcia la vita. Da qui l'esigenza di trovare sempre nuove modalità per garantire e aumentare la fruibilità del patrimonio senza comprometterne le caratteristiche materiali.

L'amministrazione archivistica italiana ha individuato da tempo, nel processo sistematico di digitalizzazione, la via maestra per garantire la conservazione del patrimonio e la sua accessibilità, un indirizzo di policy sfociato nel 2017 nel Piano nazionale di digitalizzazione del patrimonio culturale[3].

Su questa scia, con intuizione antesignana, il laboratorio interno di fotoriproduzione dell'Archivio di Stato di Venezia ha avviato la creazione, fin dagli esordi dell'era digitale, di una banca dati che supera, oggi, gli 80 terabyte di immagini digitali di documenti.

Questo patrimonio immateriale permette di non movimentare gli originali dal deposito se non per valide e motivate ragioni, non solo garantendo la possibilità di una fruizione adeguata delle fonti storiche, ma ampliandola, poiché le immagini digitali possono essere visionate da più utenti contemporaneamente e rese accessibili anche da remoto.

Il patrimonio digitale, inoltre, favorisce la diffusione dei beni culturali anche tra il pubblico dei non specialisti, costituendo le basi per una divulgazione di qualità.

In questo scenario, volto a favorire le sperimentazioni di tecniche e strumenti di valorizzazione che non impattino sulle condizioni materiali del patrimonio archivistico, da diversi anni gli archivi sono oggetto di studio da parte del mondo accademico e, in particolare, di settori come le *di-*

[3] Per le caratteristiche e l'attuazione del piano si veda http://pnd.beniculturali.it/il-piano/.

gital humanities e la ricerca sulle intelligenze artificiali e l'uso dei *big data*.

Procedendo su piani paralleli, l'Istituto Centrale per il Catalogo e la Documentazione (ICCD) ha predisposto, a livello centrale, una serie di ontologie informatiche per supportare l'ingresso degli archivi e dei beni culturali in genere nel mondo dei *big data* e del *semantic web,* mentre negli istituti periferici, che conservano il patrimonio e possono sperimentare su di esso, si è proceduto invece a tentativi, a volte pionieristici, di applicazione dei metodi quantitativi e delle ICT alla ricerca storica[4].

Il solo Archivio di Stato di Venezia può vantare, negli anni, diverse collaborazioni ad alto livello per la creazione di banche dati, a partire dai progetti decennali con la Hedgelawn Foundation fino al più recente progetto Garzoni, Apprenticeship, Work and Society – GAWS (Bellavitis et al. 2017)[5].

La digitalizzazione come sintesi tra conservazione e valorizzazione del patrimonio archivistico

Questo lavoro ambisce a dimostrare come le tecniche di analisi testuale nella gestione archivistica possano migliorare l'offerta di servizio degli Archivi, offrendo nuove modalità di studio e di approccio al patrimonio archivistico. Le potenzialità offerte da queste metodologie, infatti, sembrano produrre risultati pienamente soddisfacenti sul piano scientifico, e possono pertanto rappresentare una possibile alternativa ai tradizionali strumenti di indagine storica.

L'analisi testuale è ormai molto diffusa nel campo delle discipline umanistiche, ma incontra subito un ostacolo all'applicazione immediata alle fonti storiche. Si tratta, come è facile immaginare, della 'lettura' dei testi, per la quasi totalità manoscritti e redatti con grafie molto varie: diverse nei secoli, in prospettiva diacronica, e diverse da *scriptor* a *scriptor*, soprattutto a partire dal XVI secolo in poi, quando la pratica della scrittura diventò diffusissima.

Questo primo grande ostacolo ha finora orientato la ricerca princi-

[4] Il progetto principale in tale senso è *Progetto ArCo - Architettura della conoscenza,* http://www.iccd.beniculturali.it/it/progetti/4597/arco-architettura-della-conoscenza-ontologie-per-la-descrizione-del-patrimonio-culturale.

[5] La collaborazione tra l'Archivio di Stato di Venezia e la Hedgelawn Foundation ha dato vita a numerosi interventi, tra tutti il progetto *Cives veneciarum,* http://www.civesveneciarum.net/ e, in generale, la messa a disposizione online di numerose riproduzioni digitali e schedature oggi consultabili dal sistema informativo dell'Istituto *moreveneto,* http://asve.arianna4.cloud/. Per il progetto GAWS si veda anche https://garzoni.hypotheses.org/.

palmente verso le fonti librarie a stampa e lo sviluppo di strumenti di lettura automatizzata, moderni OCR in grado di elaborare le immagini e trasformarle in testi analizzabili con tecniche quantitative[6].

Pur in considerazione di tale ostacolo, tramite il presente lavoro si tenterà di mostrare come l'introduzione di tecniche di analisi testuale nell'amministrazione archivistica, fuori pertanto dalla pura sperimentazione accademica, potrebbe trovare spazio in diversi ambiti, migliorando concretamente l'offerta all'utenza, in particolare offrendo nuove risorse alla ricerca storica e all'insegnamento della diplomatica.

La ricerca storica riguarda il principale bacino di fruitori dell'Istituto, cioè studenti, ricercatori e docenti di discipline storiche. L'insegnamento della diplomatica pertiene invece alla *mission* formativa dell'Archivio di Stato di Venezia che, a norma del Regio Decreto 2 ottobre 1911, n. 1163, con altri 16 istituti italiani, è dotato di una Scuola di archivistica, paleografia e diplomatica. Tale Scuola, al termine di un biennio di studi e di un esame finale, rilascia un titolo necessario ad operare sugli archivi vigilati dalla Stato italiano e dalle Soprintendenze archivistiche e bibliografiche. Nell'ambito della formazione dell'archivista, la *diplomatica*, cioè la disciplina che studia la genesi e le forme del documento (soprattutto medievale), sembra la più affine all'approccio della *text analysis* poiché inferisce i suoi precetti dallo studio di formulari, ricorrenze e strutture dei testi, condividendone dunque l'impostazione metodologica seppure con strumenti tradizionali.

In entrambi i casi, l'implementazione di tecniche di analisi testuale si inserisce nel più ampio contesto della valorizzazione e della tutela dei beni culturali. È evidente, infatti, che offrire all'utenza possibilità di ricerca innovative e allineate con i nuovi indirizzi metodologici dell'analisi del documento e del testo costituisce un'occasione di valorizzazione molto interessante, soprattutto perché porrebbe l'amministrazione archivistica al centro di un processo di innovazione e promozione di nuovi approcci di studio nei confronti di un bacino di utenti in cui dominano gli approcci qualitativi e tradizionali.

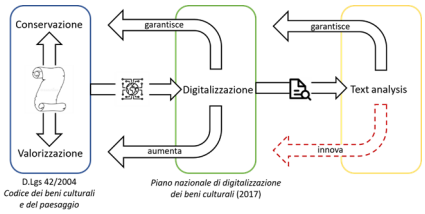
Sul versante della tutela, è opportuno ribadire che ogni metodologia in grado di offrire opportunità di ricerca scientificamente valide senza richiedere la consultazione materiale del patrimonio va perseguita con ogni forza, poiché evita la movimentazione di centinaia di documenti dai

[6] Tra i molti progetti di lettura automatizzata si segnala READ, promosso da un consorzio di più di 80 istituti europei e che fornisce la piattaforma *Transkribus*, https://eadh.org/projects/read.

depositi alla sala di studio con conseguente risparmio in termini di tempo e risorse e, soprattutto, riducendo il rischio di danneggiamento e usura del patrimonio.

In definitiva, la ricerca mira, dunque, a mostrare come l'applicazione di tecniche di analisi testuale ad una specifica tipologia di materiale archivistico, appositamente individuata, possa produrre un miglioramento nell'erogazione dei servizi dell'Archivio di Stato di Venezia, in ottemperanza agli obbiettivi di valorizzazione imposti dalla norma e nel rispetto di tutte le accortezze in ordine alla tutela del patrimonio. Si veda la Fig. 1 di modellizzazione grafica della ricerca.

Fig. 1 Modello della proposta di ricerca.



Il capitolo è organizzato in diverse sezioni. Nel prossimo paragrafo sarà presentato il *Codice diplomatico veneziano*, la raccolta da cui provengono i testi che hanno costituito il punto di partenza per l'analisi del caso di studio. Successivamente, uno specifico paragrafo sarà dedicato alla formazione del corpus, alla scelta delle variabili e alle criticità emerse nel lavoro di raccolta e *pre-processing* dei testi. Verranno poi presentati i risultati della *topic detection* effettuata sui testi con il software Iramuteq e l'interpretazione dei *topics* alla luce delle attuali conoscenze archivistiche e diplomatistiche. Infine, si proporrà una riflessione sulle possibili applicazioni dei risultati dell'analisi e della metodologia utilizzata ai servizi offerti dall'amministrazione archivistica, con particolare attenzione alle potenzialità, anche metodologiche, offerte alla ricerca storica, principale interesse degli utenti dell'Archivio, e allo studio e insegnamento della diplomatica.

Il Codice diplomatico veneziano

Il corpus per l'analisi testuale è stato individuato nel *Codice diplomatico veneziano*, il monumentale strumento di descrizione messo a punto da Luigi Lanfranchi (Lanfranchi 1944-1986), archivista e poi Direttore

dell'Archivio di Stato di Venezia, avviato nel 1944 e continuato dallo studioso per tutta la vita (Lanfranchi 1942, 1984).

Innanzitutto, va specificato che il Codice non è una fonte archivistica propriamente detta, bensì uno strumento di descrizione che consente, attraverso la sua mediazione, di apprendere il contenuto delle fonti documentarie che descrive e, solo in seconda istanza, di giungere alla loro eventuale consultazione. La scelta di adoperare l'analisi testuale sugli strumenti di ricerca e non direttamente sulle fonti si basa su alcune valutazioni preliminari:

- le fonti sono manoscritte e la loro trascrizione è un'operazione scientifica complessa, estremamente onerosa in termini di tempo e non quantificabile in una *timeline* di lavoro (troppe variabili, infatti, ne rendono impossibile la definizione: diversa lunghezza dei testi, differenti scritture, possibile illeggibilità dovuta allo stato di conservazione);
- le fonti sono in latino medievale o in volgare per cui richiederebbero l'uso di dizionari non ancora disponibili nella libreria del software utilizzato per le analisi;
- gli strumenti di descrizione sono molti e costituiscono essi stessi un patrimonio da valorizzare;
- gli strumenti di descrizione sono manoscritti, dattiloscritti e, in minima parte, anche scritti in formato digitale.

La scelta di realizzare il corpus a partire dallo strumento del Codice è quindi dovuta alla necessità, in questo primo tentativo sperimentale, di ridurre alcune macro-criticità di cui si darà conto più ampiamente in seguito, nel rispetto degli assunti di partenza, cioè la valorizzazione del patrimonio archivistico e la preservazione della sua integrità materiale, evitando di lavorare sulla documentazione originale.

Il lavoro di Lanfranchi si ispira alla tradizione dei 'diplomatici', ossia raccolte sistematiche di pergamene – solitamente le più antiche – accomunate dalla medesima provenienza, riferite al medesimo contesto, o semplicemente riunite in ordine cronologico. La formazione dei diplomatici, molto diffusa nel XVIII secolo, è un'operazione considerata oggi anti-archivistica e di fatto non consentita dalle norme vigenti, poiché implica l'estrazione dei singoli documenti dai fondi archivistici di provenienza e la conseguente distruzione del vincolo archivistico esistente tra le carte, decontestualizzando così la fonte storica[7].

[7] A tal proposito l'art. 20, comma 2 del D.lgs. 42/2004 ordina che "Gli archivi non possono

Queste operazioni, spesso irreversibili, erano tuttavia dettate dal gusto museografico e antiquario dell'epoca e hanno prodotto strumenti formidabili per lo studio della diplomatica e della paleografia, consentendo agli studiosi di allora, privi di mezzi di riproduzione, di consultare e confrontare le più antiche testimonianze scrittorie e gettare di fatto le basi della paleografia e della diplomatica come le conosciamo oggi.

Il caso più celebre di diplomatico è senz'altro quello voluto dal Granduca di Toscana, Pietro Leopoldo, realizzato a partire dal 1778 e più volte implementato in seguito, oggi conservato presso l'Archivio di Stato di Firenze[8].

Una simile operazione fu tentata, qualche decennio dopo, da Napoleone Bonaparte, il quale promosse, durante il dominio sui territori italiani, la creazione di un immenso *Archivio diplomatico del Regno italico* da realizzare a Milano presso l'Archivio di San Fedele, trasferendovi anche le pergamene provenienti dai territori veneti. Il progetto fu avviato nel 1807 dalla Direzione generale del demanio e diritti uniti che, tramite le sue articolazioni periferiche nei singoli dipartimenti amministrativi, ordinò l'estrapolazione di migliaia di pergamene (tutte quelle prodotte tra l'VIII secolo e l'anno 1400) dai fondi di pertinenza, preparandole per il viaggio verso Milano.

La conclusione dell'epoca napoleonica portò all'abbandono del progetto dell'Archivio diplomatico, ma lasciò le pergamene, ormai estratte e riunite nella sede veneziana della Direzione dipartimentale del demanio e diritti uniti in un edificio in contrada San Provolo, in stato di disordine.

All'opera di descrizione e riordino di questa documentazione sovrintese, a un secolo e mezzo di distanza, Luigi Lanfranchi, che vi affiancò un'operazione di individuazione e schedatura della documentazione membranacea più antica conservata in Archivio. Tale lavoro, e qui risiede la grande intuizione dello studioso, fu condotto in maniera virtuale, senza cioè modificare l'ordinamento dei fondi archivistici di provenienza che, anzi, andavano col tempo a ritrovare il loro ordinamento originario.

Lanfranchi realizzò un corpus di trascrizioni, di regesti e di riproduzioni analogiche in bianco e nero (di straordinaria qualità per l'epoca), e avviò dunque la formazione del Codice diplomatico veneziano.

Il Codice, rimasto dattiloscritto e inedito, fu costantemente ampliato, integrato e corretto, rimanendo incompiuto con la morte del suo autore

essere smembrati".

[8] Il Diplomatico fiorentino è integralmente consultabile online, https://www.archiviodistato.firenze.it/pergasfi/.

nel 1986. La natura 'virtuale' dello strumento consentì, inoltre, a Lanfranchi, di ricercare e inserire nella raccolta anche quei documenti relativi a Venezia – o comunque pertinenti ai fondi archivistici veneziani – che, per le più varie ragioni, non erano conservati presso l'Archivio di Stato, ma presso biblioteche e altri istituti, sia italiani che stranieri.

La struttura del Codice

La struttura del *Codice* si articola in 3 serie principali, cui si aggiunge la raccolta delle fotoriproduzioni (Erboso 2019):

- trascrizioni di documenti dell'XI e XII secolo, dal 1000 al 1199, in 32 volumi;
- regesti di documenti dell'XI e XII secolo, in 10 volumi;
- regesti di documenti del XIII secolo, realizzati in schede riunite in 145 raccoglitori.

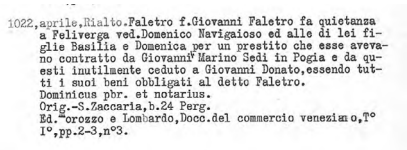
La serie più adatta per la sperimentazione è quella dei regesti di documenti del XI e XII secolo, i cui dattiloscritti originali sono conservati alla Biblioteca del Museo civico Correr di Venezia, cui furono donati dagli eredi Lanfranchi ma che l'Archivio possiede in copia.

Il regesto è, in diplomatica, il riassunto di un documento redatto secondo specifiche regole, formalizzato e normalizzato. Queste caratteristiche, oltre al fatto di essere redatti in lingua italiana, rendono i regesti particolarmente interessanti in quanto contengono tutte le informazioni necessarie a individuare un documento e a comprenderne la tipologia e il contenuto giuridico

La formazione del corpus

Il corpus testuale impiegato per l'analisi è stato realizzato trascrivendo manualmente i primi due volumi dei regesti di documenti dell'XI e XII secolo del Codice diplomatico veneziano, corrispondenti alle descrizioni di tutte le pergamene veneziane dall'anno 1000 fino a tutto il 1134, per un totale di 727 testi (Fig. 2).

Fig. 2 Esempio di un regesto tratto dal Codice diplomatico veneziano.


1022,aprile,Rialto.Faletro f.Giovanni Faletro fa quietanza a Feliverga ved.Domenico Navigaioso ed alle di lei figlie Basilia e Domenica per un prestito che esse avevano contratto da Giovanni^V Marino Sedi in Pogia e da questi inutilmente ceduto a Giovanni Donato,essendo tutti i suoi beni obbligati al detto Faletro.
Dominicus pbr. et notarius.
Orig.-S.Zaccaria,b.24 Perg.
Ed.-Corozzo e Lombardo,Docc.del commercio veneziano,T° I°,pp.2-3,n°3.

Benché i testi del Codice siano dattiloscritti, non è stato possibile automatizzare la trascrizione con software OCR, essendo caratterizzati da un ampio uso di abbreviazioni e di termini specifici, spesso non tradotti dalla lingua latina, che comportano, durante il lavoro di trascrizione, un intervento impegnativo di mediazione e verifica, anche ricorrendo alla consultazione di bibliografia specifica o di altre serie del Codice come le Trascrizioni. Su queste e altre difficoltà si tornerà a breve.

Anche il tentativo di adoperare un software di scrittura vocale, come per esempio quello in dotazione nella *suite* di Chrome, non ha presentato vantaggi significativi in termini di tempo necessario alla trascrizione del corpus, poiché il tempo risparmiato nella digitazione è compensato dal lavoro di revisione e di correzione.

Lo schema di redazione dei regesti è quello consolidato dalla diplomatica e prevede che di ogni documento venga indicata:

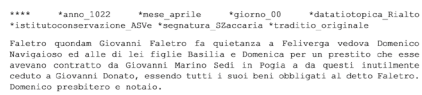
- la cosiddetta *datatio*, ossia l'indicazione della data (*datatio cronica*) e del luogo (*datatio topica*) in cui è stato redatto il documento;
- il riassunto del contenuto giuridico secondo una struttura costante: *autore* dell'azione giuridica, verbo dispositivo (che indica la natura giuridica dell'atto che si documenta), destinatario dell'azione. A queste componenti fondamentali si aggiungono tutte le clausole che limitano o rendono efficace il contenuto giuridico del documento;
- la trascrizione del nome del funzionario che ha redatto il documento secondo le caratteristiche formali che ne garantiscono la validità. Si tratta, nella maggior parte dei casi, del nome del notaio o del cancelliere, riportato così come si trova nel documento e dunque in latino medievale;
- informazioni relative alla cosiddetta *traditio* del documento, cioè al fatto che esso sia un originale o una copia, dando eventualmente notizia anche di copie note del documento;
- segnatura del documento, cioè la collocazione logica della pergamena all'interno del fondo archivistico di provenienza;
- indicazioni bibliografiche.

Per quanto riguarda la compilazione del corpus, il dato indicato al punto 3, relativo all'ufficiale che roga l'atto, è stato tradotto di volta in volta in italiano, per consentire al dizionario del software di analisi di elaborarne il contenuto. I punti 4 e 5 non sono stati trascritti, ma sono stati individuati come variabili per la classificazione dei testi. Il punto 6 non è

invece stato trascritto in quanto non pertinente alle finalità dell'analisi.

Il corpus è stato poi formattato per l'analisi con il software Iramuteq e strutturato attraverso un *set* di metadati, individuando delle variabili a cui sono associate le modalità caratteristiche di ogni testo (Fig. 3).

Fig. 3 Esempio di un regesto trascritto per il corpus testuale da analizzare.



Le variabili individuate sono le seguenti:

- anno, mese, giorno, per individuare la datazione cronica del documento;
- datazione topica, cioè il luogo dove è stato redatto il documento;
- istituto di conservazione, che nella quasi totalità dei casi è l'Archivio di Stato di Venezia, ma non solo (alla data di redazione del Codice diplomatico veneziano, infatti, alcuni documenti si trovavano ancora in altri istituti archivistici del Veneto in quanto il processo di concentramento degli archivi a Venezia non era ancora stato ultimato);
- *segnatura* (il termine non è inteso qui col significato tecnico sopra riportato, ma in un'accezione più ampia, che individua il soggetto produttore del documento, in larga misura coincidente con la segnatura, ma non in tutti i casi);
- *traditio*, con le sole modalità originale/copia.

Questi concetti, va chiarito, non rappresentano un gradiente di valore della documentazione, ma sono elementi specifici di valutazione nell'esegesi storica e dunque fondamentali anche per le analisi quantitative.

Il corpus è stato sottoposto ad un successivo lavoro di bonifica e normalizzazione delle parole in modo tale da risultare trattabile statisticamente. Dopo i primi tentativi di analisi, infatti, è stato rilevato un ampio numero di *hapax,* riconducibili principalmente ai nomi di persona e ai toponimi presenti nei regesti, spesso molto distanti dagli esiti linguistici contemporanei.

Il lavoro massiccio di normalizzazione è stato eseguito a posteriori, con l'inserimento anche di alcune *multi-word expressions*: in modo particolare, si sono trasformati in un'unica parola tutti i toponimi relativi ai

nomi di chiese e monasteri, si è unito il numero ordinale al nome proprio di papi imperatori e sovrani e, in generale, i nomi di località composti da più parole.

Criticità nella formazione del corpus

La costruzione del corpus ha richiesto, in ultima analisi, il ricorso ampio alle competenze multidisciplinari che sono richieste all'archivista, rivelandosi un'operazione non solo onerosa in termini di tempo, ma anche altamente specializzata e solo in parte, dunque, automatizzabile.

In particolare, le difficoltà riscontrate sono le seguenti:

- traduzione: come già ricordato, anche i regesti contengono parole o addirittura frasi in latino medievale o in volgare veneziano che devono necessariamente essere tradotti per essere analizzati. L'assenza di vocabolari in queste lingue è la ragione principale che ha portato il presente lavoro a concentrarsi sugli strumenti di descrizione e, tra essi, sui regesti anziché sulle fonti o sulle loro trascrizioni. L'analisi testuale su testi latini potrebbe rivelarsi molto efficace, soprattutto perché la lemmatizzazione, applicata ad una lingua flessiva com'è il latino, dovrebbe funzionare al meglio. Tuttavia, Iramuteq non dispone ancora di questo strumento e, in generale, i dizionari online non contemplano il latino medievale (che ha una varietà lessicale molto più ampia di quello classico);
- scioglimento delle abbreviazioni: i regesti presentano molte abbreviazioni, necessarie, a suo tempo, per velocizzare il lavoro di dattiloscrittura a fronte dell'immensa mole di documenti da descrivere. Le medesime abbreviazioni, tuttavia, non si sciolgono sempre allo stesso modo e richiedono, di volta in volta, di valutare il contesto. Un esempio è l'abbreviazione "ab." adoperata sia per "abate" che per "abitante";
- termini tecnici: nei regesti si fa spesso ricorso a termini tecnici o specifici del contesto cronologico o topografico di riferimento e che richiedono, per essere meglio contestualizzati o sostituiti con termini correnti presenti nel vocabolario di Iramuteq, di ricorrere alla lettura della trascrizione del documento o a bibliografia specifica. È il caso, per esempio, di termini come "fondamento", che nel contesto delle saline si riferisce alle strutture di compartimentazione delle basse aree lagunari trasformate in vasche per la raccolta del sale, o a termini come

"repromissa" o "dimissoria", che fanno riferimento a specifici istituti giuridici del diritto intermedio;

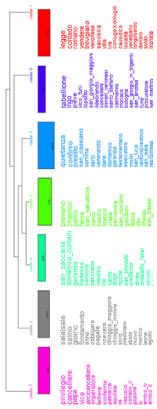
- normalizzazione: la redazione del Codice è stata condotta nell'arco di decenni, per cui gli stessi luoghi e le stesse persone non sono sempre decritti in modo uniforme e risentono, anzi, della varietà di occorrenze con cui questi si ritrovano nei documenti. Il caso preponderante è quello dei nomi e dei cognomi che non sono ancora fissati in forme stabili e che si ritrovano nei regesti con molte varianti da ricondurre, per il buon funzionamento dell'analisi, ad un'unica variante standard.

Analisi dei *topics*: metodologie innovative per il miglioramento della fruizione del patrimonio archivistico

L'applicazione delle tecniche di analisi automatica del contenuto al Codice diplomatico veneziano, presentata in questo paragrafo, ha prodotto risultati che rappresentano quanto già noto in letteratura, oltreché fornire nuova conoscenza, mostrando chiaramente come il prodotto di tali modalità di analisi possa effettivamente costituire la base su cui costruire un'offerta innovativa per migliorare la fruizione del patrimonio da parte dell'utenza (o almeno offrire nuove possibilità ai ricercatori di studiare il patrimonio documentario veneziano), e favorire la conoscenza del patrimonio stesso, sia ai fini della didattica erogata dall'Istituto, sia per la formazione del personale stesso.

La *clusterizzazione* del corpus con il metodo Reinert ha individuato la presenza di 7 classi semantiche, riassunte nella Fig. 4, la quale indica, per ogni classe, il processo di classificazione, la percentuale di testi racchiusi in ciascun *cluster* e le parole più significative ad essi associate.

Fig. 4 Classi semantiche con rappresentazione dei termini più significativi.



L'analisi dei singoli *cluster* consente di interpretare le 7 classi come segue:

1) documenti privati che riguardano negozi tra soggetti sottoposti al diritto romano o al diritto longobardo;
2) documenti privati relativi alle saline di Chioggia;
3) documenti privati che attestano passaggi di proprietà che interessano principalmente il monastero di San Cipriano di Murano;
4) documenti privati che attestano passaggi di proprietà che interessano principalmente il monastero di San Zaccaria di Venezia;
5) documenti di quietanza relativi a prestiti e all'amministrazione del diritto successorio;
6) documenti privati relativi agli affari del monastero di San Giorgio Maggiore in terraferma;
7) documenti pubblici.

Come premessa generale, l'esame delle classi semantiche indica che il corpus documentario esaminato è composto in larga misura da documenti che attestano titoli di proprietà e che sono in buona parte riconducibili ai fondi archivistici prodotti dalle corporazioni religiose. Dal punto di vista storico e archivistico, questa specificità si spiega con il perdurante interesse degli istituti religiosi a mantenere prova dei diritti connessi ai propri possedimenti. Si tratta, infatti, di istituti che spesso sopravvivono per secoli ai mutamenti politici che avvengono attorno ad essi e che hanno sempre avuto cura dei propri archivi ordinandoli, conservandoli al meglio e predisponendo copie autentiche dei documenti deperiti o illeggibili.

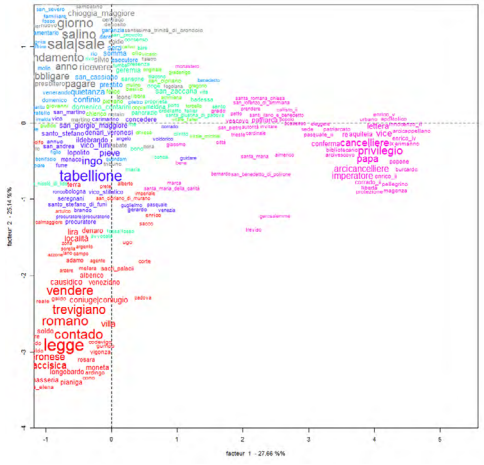
Il software ha distinto nettamente la documentazione pubblica (classe 7) da quella relativa ai negozi giuridici tra privati (classi 1-6). Si tratta di una suddivisione classica dello studio diplomatistico, che spicca in modo ancora più marcato nella Fig. 5, relativa ad un'analisi delle corrispondenze applicata ai *cluster*.

La documentazione pubblica è infatti quella prodotta da entità sovrane (papi, imperatori, re, ecc.), e convalidata da figure professionali specifiche, i cancellieri, laddove invece la documentazione tra privati si avvale della professionalità dei notai.

Anche sul piano della tipologia documentaria c'è una grande differenza: le autorità sovrane emanano privilegi, concedono diritti ed esenzioni, promettono protezione; gli atti tra privati sono, invece, espressione della molteplicità di istituti giuridici presenti nel diritto romano, nel diritto intermedio e nel diritto veneto.

Altra suddivisione netta proposta dal software è quella tra i documenti in cui una delle parti è veneziana (soggetto privato o monastero) da quella tra persone esterne al ducato (classe 1). Tale distinzione viene individuata proprio a partire dal piano giuridico, poiché il diritto applicato nella terraferma veneta dell'XI-XII secolo, prima dunque della conquista veneziana, è quello imperiale-romano o, in alternativa, quello salico o longobardo, mentre nelle isole lagunari è andata formandosi una forma propria e specifica di diritto.

Fig. 5 Rappresentazione dei *cluster* su piano cartesiano.



Va ricordato, infatti, che in quel periodo vigeva ancora la personalità del diritto e ogni individuo agiva giuridicamente secondo le leggi del popolo cui apparteneva, e non secondo quelle del territorio in cui si trovava.

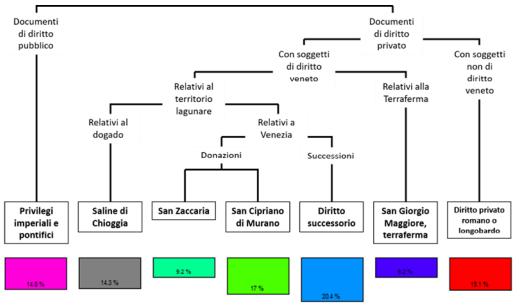
Le altre classi semantiche individuano nuclei specifici di documentazione legati all'attività, per esempio, dei monasteri di San Giorgio Maggiore, San Cipriano di Murano e di San Zaccaria. Si tratta delle corporazioni religiose di cui si conserva la documentazione più antica e che rappresentano, con le loro pergamene, una parte quantitativamente rilevante del corpus esaminato, tanto da portare all'individuazione di specifici *cluster*.

Di interesse è anche il *cluster* 5, che riunisce una serie di documenti che attestano passaggi di proprietà tra privati, ma a differenza di quelli compresi nelle altre classi, specificatamente legati al diritto successorio. In questa classe non troviamo dunque donazioni, ma principalmente eredità, esecuzioni testamentarie e qualche caso di cessione di beni posti a garanzia di prestiti non onorati per la morte del contraente.

La Fig. 6 riassume il processo di *clusterizzazione* compiuto da Iramuteq, ed evidenzia come il software abbia inferito dalle descrizioni delle pergamene una serie di distinzioni di carattere giuridico, territoriale e tipologico. Si tratta in tutti i casi di suddivisioni coerenti e appropriate,

confermate anche dalle valutazioni di tipo qualitativo che si sono svolte durante la trascrizione del corpus.

Fig. 6 Schema di organizzazione delle classi semantiche.



Migliorare e ampliare la valorizzazione e la fruibilità del patrimonio archivistico

Le analisi di cui al precedente paragrafo consentono di ipotizzare delle modalità concrete di implementazione dei risultati e, in generale, del metodo dell'analisi testuale, nei servizi resi all'utenza da parte dell'amministrazione archivistica, in particolare degli Archivi di Stato che conservano il patrimonio e si occupano della sua fruizione e valorizzazione. Nei paragrafi seguenti si analizzeranno le possibili applicazioni dell'analisi testuale per migliorare la conoscenza del patrimonio da parte degli storici, principali fruitori dell'Archivio, e per l'insegnamento e lo studio della diplomatica, coerentemente con la *mission* didattica delle Scuole di archivistica, paleografia e diplomatica.

La ricerca storica

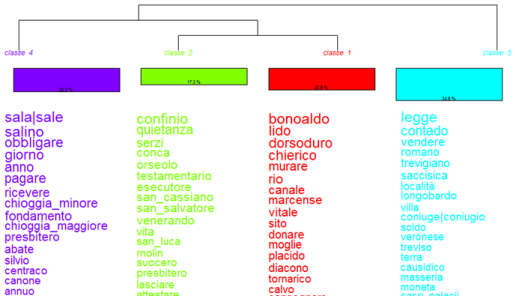
Il principale bacino di utenza degli archivi di Stato, per lo meno di quelli di grandi dimensioni che hanno sede nelle ex capitali degli stati preunitari, è costituito da ricercatori altamente specializzati, provenienti dal mondo accademico e con interessi di ricerca vari, che possiamo sinteticamente ricondurre nell'alveo delle discipline storiche e delle scienze del documento.

Le possibili ricerche che si possono condurre a partire da un corpus testuale come quello oggetto di questo capitolo sono potenzialmente infinite, e ci si limiterà pertanto ad individuare alcuni spunti che si ritiene siano resi possibili solo dall'utilizzo di tecniche di analisi automatiche,

utili a evidenziare la portata innovativa dello strumento.

Il caso più semplice è senz'altro quello legato alla possibilità di estrarre dal corpus dei *subcorpora* statisticamente trattabili sulla base di specifici interessi di ricerca. Per proporre un esempio 'classico' per gli studi storici, si è creato un *subcorpus* utilizzando, tra i metadati, la variabile "segnatura". Sono stati così selezionati tutti i documenti prodotti dai monasteri benedettini – istituti a cui va il merito di aver colonizzato e popolato la laguna di Venezia dando impulso all'organizzazione territoriale e all'economia delle isole realtine (Fig. 7) – sui quali è stata eseguita tramite il software Iramuteq una *topic detection* con il metodo Reinert.

Fig. 7 Classi semantiche estratte dal *subcorpus* dei documenti prodotti dai monasteri benedettini.



Anche in questo caso, la divisione semantica rispetta criteri diplomatistici basati su distinzioni di natura giuridica e territoriale. Essa consente inoltre di far emergere nuove domande di ricerca e filoni di indagine, legati per esempio al legame tra l'ordine benedettino e l'estrazione del sale in laguna (classe 4), o alla politica di acquisizioni terriere che i monaci compierono con assiduità e insistenza in terra padovana, in particolare in Saccisica (classe 3).

L'analisi testuale offre dunque la possibilità di interrogare centinaia – e, in prospettiva, migliaia – di documenti contemporaneamente, e di individuare, attraverso lo studio del corpus, quali documenti possano essere maggiormente di utilità alla propria ricerca.

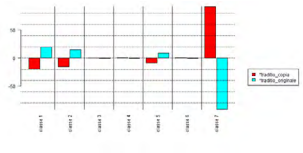
Successivamente, il ricercatore potrà ampliare l'indagine passando alla consultazione dei documenti, che potrà essere condotta nelle trascrizioni del Codice diplomatico e, quando disponibili, nelle riproduzioni digitali, ricorrendo agli originali solo in via residuale, consentendo all'amministrazione archivistica di perseguire così l'intento iniziale di valorizzare le fonti limitandone il più possibile la consultazione diretta, e ampliando al contempo le possibilità di fruizione.

La facilità con cui il software utilizzato consente di organizzare il corpus in *subcorpora* dà luogo a possibilità vastissime. Anche solo le variabili individuate in questa fase sperimentale consentono di indicare archi cronologici specifici, di definire ambiti territoriali, di selezionare specifiche categorie di soggetti produttori.

Altre possibilità di analisi riguardano, per esempio, la misura del valore χ^2 di associazione di singole parole o dei metadati con i *cluster* originati dalla *topic detection*, o di singole parole o gruppi di parole con le variabili con cui è stato classificato il corpus.

Un esempio è mostrato nella Fig. 8, che rappresenta il rapporto χ^2 tra i 7 *cluster* e la variabile "*traditio*", ovvero la variabile che classifica il documento sulla base che sia un originale o una copia.

Fig. 8 Associazione della variabile "traditio" con i 7 *cluster*.



Il grafico mostra come la classe 7, composta dai documenti pubblici, raccolga principalmente documenti in copia, un dato interessante e confermato anche dall'esperienza diretta.

Pur in assenza di dati quantitativi circa il rapporto tra originali e copie all'Archivio di Stato di Venezia (dati che un eventuale completamento del corpus potrebbe fornire), si può asserire che la maggior parte della documentazione prodotta da soggetti pubblici tra l'XI e il XII secolo sia pervenuto a noi tramite copie più tarde.

Gli originali di questo periodo, infatti, risultano essere piuttosto rari e spesso in cattive condizioni. La quasi totalità di essi, inoltre, è conservata nei fondi archivistici delle corporazioni religiose, spesso caratterizzate da una storia istituzionale travagliata, che si traduceva con la dispersione o lo smembramento dell'archivio. Per questa ragione, poiché le pergamene prodotte dalla cancelleria imperiale e da quella pontificia erano considerate le più importanti in virtù del prestigio del loro produttore, gli istituti religiosi ne curavano frequentemente la copia, non appena le condizioni materiali degli originali andavano deperendo a causa dell'usura o dei frequenti casi di sottrazione, distruzione o dei tentativi di manomissione.

Si deve attendere la seconda metà del XIII secolo per un cambio di tendenza: il consolidarsi di una cancelleria vera e propria, intesa come or-

gano burocratico a supporto delle attività di uno Stato, quello veneziano, ormai formato e indipendente dall'Impero Romano d'Oriente, rappresenta la nascita di una struttura incaricata di garantire la certezza del diritto per tutti i cittadini del Comune.

Da questo momento in poi si assiste, innanzitutto, a una massiccia operazione di copia dei documenti più antichi presenti sul territorio lagunare, che venivano trascritti dagli ufficiali pubblici in registri da conservare nei locali della cancelleria ducale e che spesso costituiscono l'unica testimonianza di tale antica documentazione. Molte delle copie dei documenti più antichi sono dunque riconducibili a questa attività e a questo periodo (si pensi al caso dei *Pacta*, una raccolta di sette registri iniziata nel XIII secolo col preciso intento di tramandare i più antichi trattati internazionali della Repubblica, con trascrizioni di documenti dal IX secolo in poi i cui originali si sono perduti).

In secondo luogo, sempre a partire dalla metà del XIII secolo, la cancelleria e lo Stato diventarono corrispondenti delle altre autorità pubbliche dell'Europa del tempo, per cui molti originali giunsero a Venezia indirizzati direttamente al Doge: questi documenti venivano conservati direttamente in Palazzo ducale e archiviati con grande cura. Nel progresso di tempo, dunque, il rapporto tra originale e copia si invertì e le pergamene originali divennero la maggioranza.

Di maggior interesse e più difficile spiegazione è, invece, la grande quantità di originali concentrati nelle classi 1, 2 e 5 e riconducibili agli atti rogati secondo diritto imperiale, agli atti relativi alle saline di Chioggia e a quelli di diritto successorio. Un rapporto tra il livello di conservazione di questa documentazione e il suo contenuto e tipologia costituisce senz'altro un campo di studio inedito e potenzialmente fruttuoso.

Impatto sullo studio e l'insegnamento della diplomatica

La diplomatica rappresenta uno dei principali insegnamenti nelle Scuole di archivistica, paleografia e diplomatica e, benché i programmi ministeriali prevedano di lavorare su documentazione medievale, la disciplina fornisce le basi per comprendere molte delle scelte del legislatore in ambito di documento digitale e gestione documentale in genere (Duranti 1998, Penzo Doria 2020).

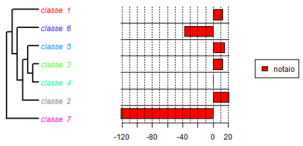
Recentemente, dopo anni di confinamento nelle Scuole di archivistica e nei corsi di laurea in storia o archivistica, la diplomatica sta riscoprendo una certa vitalità proprio grazie al supporto di nuove tecnologie informatiche e alla collaborazione multidisciplinare (Nanetti et al. 2021).

La specificità della disciplina, che studia gli aspetti formali del documento unitamente alle caratteristiche materiali, richiede la consultazione diretta delle fonti e, per il caso veneziano, ha sempre trovato nel *Codice diplomatico* di Luigi Lanfranchi lo strumento principale per l'individuazione e lo studio della documentazione.

L'analisi testuale consentirebbe di proporre, in sede didattica, modalità innovative per presentare e sedimentare i contenuti della materia, avvalendosi anche della facilità con cui il software utilizzato in questa sede elabora i contenuti in formato grafico.

Due casi a titolo puramente esemplificativo riguardano la presenza della figura del notaio come redattore e garante del valore e della pubblicità dei documenti e il rapporto tra ecclesiastici e laici nello svolgimento di tale professione. Si tratta di temi di specifico interesse per la diplomatica, che potrebbero essere analizzati grazie alla possibilità di realizzare grafici dei singoli *cluster* e alla già citata analisi del χ^2 di specifici termini.

Fig. 9 Associazione della parola notaio con i singoli *cluster*.



La Fig. 9 presenta l'analisi del valore χ^2 di associazione della parola "notaio" con i 7 *cluster* precedentemente individuati. Risalta immediatamente come nella classe 7 la parola non sia quasi mai presente, come pure sia scarsamente rappresentata nella classe 6. Tale grafico consente di visualizzare immediatamente concetti non sempre chiari sul piano didattico, come il fatto che la documentazione pubblica e quella privata sono rogate da figure differenti sul piano istituzionale benché simili su quello professionale, i cancellieri da una parte e i notai dall'altra: questo spiega la quasi assenza del termine "notaio" dalla classe 7. Per quanto riguarda la classe 6, relativa ai possedimenti del monastero di San Giorgio Maggiore in terraferma, invece, il termine ha scarsa significatività perché, come si può verificare analizzando ulteriormente il *cluster*, questi documenti sono rogati da notai laici che preferiscono definirsi "tabellioni".

Un secondo esempio riguarda un classico argomento di studio della diplomatica veneziana, cioè il rapporto tra notai laici e notai ecclesiastici (Bartoli Langeli 2006, pp. 59-86).

Una *network analysis* rappresentativa delle relazioni tra le parole all'interno del *cluster* 2, relativo ai documenti di compravendita e cessio-

ne delle saline di Chioggia, mostra chiaramente come in area veneta il notariato sia di esclusiva pertinenza ecclesiastica, tanto che risulta evidente la correlazione tra il termine "notaio" e "presbitero", una dittologia che perdurò nei territori della Serenissima per secoli rispetto al resto d'Europa (Fig. 10).

Fig. 10 Relazioni delle parole all'interno del *cluster* 2.

In entrambi questi esempi, le informazioni ricavate dall'analisi non costituiscono certo una novità per la storiografia, ma oltre a costituire la conferma della validità generale dell'analisi restituita dal software, presentano indubbi vantaggi sul piano didattico, sia per l'efficacia della resa grafica dei contenuti, sia per la possibilità di studiare questi contenuti con una logica di *learning by doing*.

Conclusioni

L'analisi testuale del corpus di documenti storici individuati nel Codice diplomatico veneziano e presentata nel capitolo si inserisce nel dibattito sulla valorizzazione e conservazione del patrimonio archivistico custodito negli Archivi di Stato, offrendo nuovi strumenti di sintesi tra le due esigenze.

I risultati ottenuti costituiscono un esempio concreto e scientificamente valido di come queste tecniche consentano di ricavare dai documenti nuove informazioni e nuovi indirizzi di ricerca altrimenti di difficile individuazione con metodi esclusivamente qualitativi e manuali, se non a fronte di un dispendio di tempo incompatibile con le odierne necessità della ricerca di ambito accademico. Al contempo, l'analisi del corpus preso in esame ha consentito di approfondire e facilitare l'accesso e la conoscenza dei contenuti dei documenti, avvalendosi anche di restituzioni grafiche utili sia sul piano didattico che sul piano della formazione del personale.

È chiaro, dunque, che sulla base di questi primi risultati, seppur sperimentali, si può affermare che lo strumento dell'analisi testuale sia in

grado di offrire modalità concrete di miglioramento dei servizi resi all'utenza, costituita principalmente da studiosi e studenti, ma anche della formazione del personale interno, tenuto alla conoscenza più approfondita del patrimonio della cui tutela è responsabile.

La possibilità più immediata per conseguire effettivamente tale policy di miglioramento dell'offerta è quella di restituire agli utenti e diffondere, attraverso il sito istituzionale o il sistema informativo dell'Archivio di Stato di Venezia, i risultati delle analisi sul Codice diplomatico veneziano in modo che, offerti all'occhio di vari specialisti, possano esprimere al massimo il loro potenziale.

Tale operazione, semplice ed economica sul piano della realizzabilità, richiede però di essere adeguatamente contestualizzata, tenendo conto dell'utenza di riferimento, molto specializzata nell'ambito storiografico, ma poco propensa alle analisi più di stampo quantitativo.

Esiste, tuttavia, anche una parte di ricercatori già abitualmente impegnati su questo tipo di indagini, alla quale potrebbe essere fornito anche il corpus di testi in modalità *raw*, in una logica di *open data*. Questa soluzione richiede che l'utente sappia adoperare il software Iramuteq o, in generale, le tecniche dell'analisi testuale, ma lo lascia poi libero di sperimentare a proprio piacimento le possibilità della *text analysis*.

In questo modo, l'archivista realizza a pieno il suo ruolo di mediatore, offrendo strumenti di ricerca che consentano di fruire del patrimonio anche in maniera innovativa e aggiornata, ma senza venire meno al principio deontologico di non entrare nel merito delle ricerche e dell'esegesi che pertengono allo storico.

La quantità di testi da trascrivere ancora disponibili lascia stimare che il corpus completo potrebbe superare le 7000 unità, considerando solo i documenti fino al XII secolo, e sarebbe ampliabile con quelli del secolo successivo di almeno cinque volte tanto.

L'implementazione della *text analysis* nel servizio di consultazione offerta agli studiosi presenta diversi ulteriori vantaggi anche sul piano della sostenibilità dei costi, nodo cruciale per le amministrazioni pubbliche. L'uso della *text analysis* ha l'indubbio vantaggio di non richiedere costi per acquisire infrastrutture poiché la dotazione hardware di un istituto come l'Archivio di Stato di Venezia è più che sufficiente a sostenere un lavoro di questo tipo. Anche i software utilizzati, come già ricordato, sono tutti *open source* e non richiedono costi aggiuntivi.

Altro elemento a favore della sostenibilità delle proposte è il fatto che, lavorando su strumenti di ricerca, si evita o si riduce al minimo indispen-

sabile la consultazione della documentazione originale. In questo senso, la sperimentazione si è dimostrata perfettamente in linea con l'obiettivo iniziale di fornire una soluzione al problema della gestione fisica dei beni archivistici, aumentando la loro fruibilità e valorizzazione senza comprometterne la conservazione (e i costi conseguenti).

Certamente i principali elementi a sfavore di un'analisi di questo tipo sono rappresentati dal tempo necessario alla predisposizione dei testi e dalla qualifica medio-alta necessaria per prendere parte al lavoro di creazione/ampliamento del corpus testuale.

Tali limiti potrebbero essere superati con finanziamenti *ad hoc* per la creazione di corpora di testo (anche nella forma di borse di studio per archivisti in formazione), che possono essere stanziati, tuttavia, solo a fronte della manifestazione d'interesse da parte dell'utenza. Da qui l'importanza di sollecitare questo interesse anche a partire dai risultati di questa ricerca.

Sul versante della formazione, si è visto come l'analisi testuale – e in particolare la *topic detection* – consenta di inferire da un corpus testuale una serie di concetti che rappresentano i punti salienti dello studio e dell'analisi diplomatistica del documento. Il software utilizzato ha consentito, in altre parole, di ripercorrere rapidamente e con un valido supporto grafico il percorso di analisi e comparazione che ha portato alla formulazione degli assunti teorici della disciplina, offrendo un approccio conoscitivo e didattico diretto e chiaro.

L'introduzione di questi strumenti, come anche l'organizzazione di laboratori didattici specifici, nell'insegnamento della Scuola di archivistica, paleografia e diplomatica, rappresenta un obbiettivo facilmente perseguibile e la cui validità si può misurare con gli strumenti consueti di valutazione della didattica e del rendimento scolastico. Per quanto riguarda quest'ultimo punto, infine, il miglioramento della qualità dell'insegnamento dovrà essere necessariamente verificato nell'arco del biennio di studi e probabilmente anche oltre ma, già in via preliminare, possiamo assumere che l'introduzione della *text analysis* nel percorso didattico di studio della diplomatica risponda all'esigenza di offrire un insegnamento aggiornato che consenta agli archivisti di domani di tessere relazioni con il mondo della ricerca contemporaneo e contribuire alla ridefinizione del ruolo degli archivi nella società.

L'uso dei social network per valutare la performance e la qualità dei servizi culturali digitali. Il caso delle Gallerie degli Uffizi

Monica Ibba[1]

Gallerie degli Uffizi, Cultural heritage, Valutazione di performance, Social mining.

Introduzione

Il capitolo presenta un caso di applicazione dell'analisi testuale alla valutazione della performance con particolare riguardo alla qualità di servizi culturali digitali. Oggetto del presente lavoro è infatti l'analisi, mediante l'applicazione di tecniche quali-quantitative (Tuzzi 2003) di text mining, del contenuto di commenti rilasciati dagli utenti ai contenuti social-culturali pubblicati dalle Gallerie degli Uffizi nella loro pagina ufficiale sul social network Facebook[2], allo scopo di valutare la performance dell'ente tramite la realizzazione di un'analisi di *customer satisfaction* della policy di digitalizzazione attuata dal museo sul *social media*.

La *customer satisfaction* può essere definita, secondo l'approccio sostenuto dalla teoria della conferma delle aspettative (*expectations-confirmation theory*), come la percezione della qualità del prodotto o servizio ricevuto, ed equivale al risultato della comparazione tra le aspettative pre-consumo e la qualità percepita (se la qualità percepita è maggiore delle aspettative,

[1] Assegnista di ricerca presso l'Università degli Studi di Padova.

[2] https://www.facebook.com/uffizigalleries/.

l'utente è soddisfatto, se è più bassa, è insoddisfatto), (Zhao et al. 2019).

In base al Decreto legislativo 27 ottobre 2009, n. 150 (D.lgs. 150/09)[3], che ha introdotto in Italia il concetto di performance organizzativa,[4] e alle successive modifiche attuate dal D.lgs. 74/17[5], l'affermazione del principio della partecipazione dei cittadini al ciclo della performance, tramite l'espressione delle proprie percezioni rispetto al servizio ricevuto, è presupposto affinché possano essere realizzate, da parte delle pubbliche amministrazioni, valutazioni delle proprie performance organizzative "secondo criteri strettamente connessi al soddisfacimento dell'interesse del destinatario dei servizi e degli interventi"[6]. A tal riguardo, uno degli ambiti in cui le amministrazioni pubbliche devono valutare annualmente le proprie performance organizzative, cioè misurare i risultati ottenuti, è "la rilevazione del grado di soddisfazione dei destinatari delle attività e dei servizi"[7].

A partire dal 2020, fra gli obiettivi di performance assegnati dal Piano triennale della performance del Ministero per i beni e le attività culturali e per il turismo[8] (oggi Ministero della cultura) agli Istituti dotati di autonomia speciale di livello dirigenziale generale – che ricomprendono molti musei e aree archeologiche, fra cui le Gallerie degli Uffizi – è stato previsto un obiettivo relativo alla predisposizione di appositi strumenti di verifica del grado di soddisfazione degli utenti, da realizzarsi mediante la messa in opera di almeno un apposito strumento permanente di verifica (*customer satisfaction*, *survey*, uso statistico dei *social media*, ecc.), al fine di disporre di risultati confrontabili annualmente.[9] A partire dal 2021, pertanto, il Pia-

[3] D.lgs. 150/09 "Attuazione della legge 4 marzo 2009, n. 15, in materia di ottimizzazione della produttività del lavoro pubblico e di efficienza e trasparenza delle pubbliche amministrazioni".

[4] Accanto a quello di performance individuale, che non sarà trattato nel presente lavoro.

[5] D.lgs. 74/17 "Modifiche al decreto legislativo 27 ottobre 2009, n. 150, in attuazione dell'articolo 17, comma 1, lettera r), della legge 7 agosto 2015, n. 124".

[6] Art. 3, comma 4, D.lgs. 150/09 e ss.mm.ii.

[7] Art. 8, comma 1, lett. c), D.lgs. 150/09 e ss.mm.ii.

[8] Piano della performance per il triennio 2020-2022. Ai sensi del D.lgs. 150/09, il Piano della performance è un documento programmatico a valenza triennale, che viene adottato entro il 31 gennaio di ogni anno dall'organo di indirizzo politico-amministrativo in coordinamento con l'Organismo indipendente di valutazione della performance (OIV) e in collaborazione con i vertici dell'amministrazione. Esso contiene gli obiettivi strategici ed operativi e i relativi indicatori e *target* per la misurazione della performance organizzativa e individuale.

[9] Piano della performance per il triennio 2020-2022. Priorità politica II: Promozione dello sviluppo della cultura. Obiettivo specifico triennale n. 3: Potenziare la qualità, le modalità di fruizione e l'accessibilità dei luoghi della cultura garantendo i necessari livelli di sicurezza nei luoghi della cultura a seguito dell'emergenza sanitaria Covid-19. Uno degli

no triennale della performance del Ministero della cultura[10] ha previsto la realizzazione di successive e annuali rilevazioni.[11]

Tramite il presente lavoro è presentata, nello specifico, una strategia di analisi della *digital customer satisfaction*, attinente cioè alla soddisfazione degli utenti circa la condivisione, da parte delle Gallerie degli Uffizi, di contenuti culturali digitali – prevalentemente nella forma di foto, video o dirette *streaming* – nella pagina Facebook del museo. Quello della digitalizzazione del patrimonio culturale è infatti un tema che acquista sempre maggiore rilevanza e urgenza tra le priorità politiche del Ministero della cultura,[12] sia come strumento di valorizzazione e conservazione, sia come principio di indirizzo delle strategie di crescita e investimento, in considerazione dell'ampio spazio dedicato agli interventi di digitalizzazione del patrimonio culturale all'interno del Piano nazionale di ripresa e resilienza (PNRR). Come conseguenza, visti anche gli obiettivi di potenziamento dell'offerta culturale digitale posti dai Piani triennali[13] di cui sopra agli Istituti dotati di autonomia speciale di livello dirigenziale generale, per il prossimo futuro ci si aspetta un incremento delle policy e delle azioni di

obiettivi annuali riguarda la predisposizione di appositi strumenti di verifica del grado di soddisfazione degli utenti, il cui indicatore è la realizzazione e la messa in opera di almeno uno strumento permanente di verifica del grado di soddisfazione degli utenti (indagini di *customer satisfaction, survey*, uso statistico dei *social media*, ecc.) al fine di disporre di risultati da confrontare con l'anno precedente.

10 Piano della performance per il triennio 2021-2023.

11 Piano della performance per il triennio 2021-2023. Priorità politica II: Promozione dello sviluppo della cultura. Obiettivo specifico triennale n. 3: Potenziare la qualità, le modalità di fruizione e l'accessibilità dei luoghi della cultura anche attraverso l'utilizzo delle nuove tecnologie di digitalizzazione in conformità con il Piano Nazionale di Ripresa e Resilienza. Garantire i necessari livelli di sicurezza nei luoghi della cultura a seguito dell'emergenza sanitaria Covid-19. Uno degli obiettivi annuali riguarda la predisposizione di appositi strumenti di verifica del grado di soddisfazione degli utenti, il cui indicatore è l'elaborazione di un report dettagliato sul grado di soddisfazione degli utenti rispetto alla consultazione online del sito dell'istituto e del patrimonio museale attraverso lo strumento di rilevazione implementato nel 2020.

12 Decreto ministeriale (D.M.) 2 aprile 2021, n. 148 "Atto di indirizzo concernente l'individuazione delle priorità politiche da realizzarsi nell'anno 2021 e per il triennio 2021-2023".

13 Piano della performance per il triennio 2020-2022 e Piano della performance per il triennio 2021-2023. Priorità politica II: Promozione dello sviluppo della cultura. Obiettivo specifico triennale n. 3: Potenziare la qualità, le modalità di fruizione e l'accessibilità dei luoghi della cultura garantendo i necessari livelli di sicurezza nei luoghi della cultura a seguito dell'emergenza sanitaria Covid-19. Uno degli obiettivi annuali riguarda il miglioramento della qualità e della fruizione dei luoghi della cultura anche attraverso l'ampliamento dei circuiti integrati e la collaborazione con gli enti locali, rendendo disponibile anche la fruizione tramite strumenti di accesso web.

digitalizzazione del patrimonio culturale e, dunque, maggiori auspicabilità ed esigenza di rilevazioni della soddisfazione degli utenti orientate anche alla fruizione digitale.

Per misurare la *customer satisfaction* degli utenti che interagiscono con i contenuti digitali, sono state ascoltate le voci degli stessi nel dialogo instaurato con il museo tramite il *social media* e il loro mutare nel tempo. In particolare, mediante il software Iramuteq, è stata realizzata un'analisi automatica del contenuto di 41.225 commenti rilasciati dagli utenti[14] alle pubblicazioni dell'istituzione sulla propria pagina Facebook tra il 10 marzo 2020 e il 30 giugno 2021, al fine di individuare il lessico e la tendenza dei discorsi degli utenti nel periodo considerato e, quindi, le loro opinioni in merito ai contenuti digitali condivisi dal museo.

L'opportunità di applicare tecniche di text mining ai dati provenienti dai social network deriva dall'enorme disponibilità di dati prodotti e conservati nei *social media* (Smyrnaios et al. 2013, Ceron et al. 2014, Criado et al. 2013), che non solo consente, a livello tecnico, di realizzare analisi automatiche di corpora di medie o grandi dimensioni, ma, a livello sostantivo, di includere nelle analisi un vasto numero di opinioni, nel tentativo di ridurre il più possibile il rischio di parzialità delle informazioni che può derivare da errate selezioni dei soggetti da coinvolgere nelle analisi.

Il capitolo è organizzato in due sezioni. Nel prossimo paragrafo sarà brevemente presentata la policy di digitalizzazione delle Gallerie degli Uffizi, sia nell'insieme sia con specifico riferimento al social network Facebook. A seguire, sarà illustrata un'analisi lessico-testuale automatica del corpus dei commenti, eseguita mediante il software Iramuteq, al fine di valutare la performance dell'istituzione museale relativamente alla percezione degli utenti di un servizio digitale. L'ultimo paragrafo è dedicato alle considerazioni conclusive.

Misurare le percezioni degli utenti per valutare la performance

La policy di digitalizzazione delle Gallerie degli Uffizi

Le Gallerie degli Uffizi sono un museo dotato di autonomia scientifica, amministrativa, finanziaria e contabile. Esse sono un ufficio periferico dirigenziale di livello generale del Ministero della cultura, e comprendono la Galleria degli Uffizi, il Corridoio Vasariano, il Giardino di Boboli e Palazzo

[14] La raccolta dei dati è avvenuta in data 18 agosto 2021.

Pitti, con i relativi Istituti e luoghi della cultura.[15]

La politica di digitalizzazione dell'ente prese avvio a partire dal 2016, con l'apertura del sito web e dei primi canali social.[16] Oggi le Gallerie degli Uffizi sono presenti su Google Arts & Culture, hanno account attivi sulle piattaforme social Instagram, Twitter, YouTube, TikTok e Facebook e un sito web dotato di archivi digitali e *tour* virtuali. Tali *digital media* sono utilizzati dal museo come strumenti permanenti di promozione del patrimonio culturale. Come riportato sulla pagina del sito web[17] delle Gallerie dedicata alla *social media* policy dell'ente,

> «i canali social delle Gallerie degli Uffizi sono utilizzati per promuovere il patrimonio culturale e le attività dei musei *in primis* presso i cittadini e gli utenti nonché per instaurare una relazione di contatto, ascolto, invito alla partecipazione, confronto/dialogo e rilevamento del *feedback* nell'ottica della trasparenza e della condivisione».

L'*account* Instagram delle Gallerie conta 686 mila *follower*, quello Twitter 58,8 mila e YouTube 5,03 mila. I più recenti *account* Facebook e TikTok contano, rispettivamente, 127,31 mila e 99,4 mila *follower*.[18] Questi ultimi sono stati aperti parallelamente allo scoppiare della pandemia Covid-19[19] e al primo *lockdown* nazionale. Se il canale TikTok è rivolto ad avvicinare al museo un pubblico giovanissimo,[20] il canale Facebook è stato creato per dare un vero contributo alla comunità, principalmente durante le settimane di chiusure nazionali, ma con una prospettiva di lungo periodo.[21]

Nel periodo compreso tra il 10 marzo 2020[22] e il 30 giugno 2021[23], le Gallerie degli Uffizi hanno pubblicato nella loro pagina Facebook 558 contenuti tra aggiornamenti di profilo, foto, video e dirette. Alla data del 18 agosto 2021, tali contenuti contavano in totale 64.489 commenti.

Il grafico in Fig. 1 mostra come, in media, tra le varie forme di condi-

[15] Art. 1, commi 1, 2, 3 dello Statuto delle Gallerie degli Uffizi, allegato al D.M. 27 novembre 2017, n. 517 "Approvazione dello Statuto delle Gallerie degli Uffizi".

[16] Da un'intervista al Corriere della Sera del direttore delle Gallerie degli Uffizi Eike Schmidt del 1° aprile 2020. Si veda https://www.facebook.com/corrieredellasera/videos/981605565574486.

[17] https://www.uffizi.it/pagine/social_media_policy_uffizigalleries.

[18] I dati sono stati raccolti il 17 gennaio 2022.

[19] L'Organizzazione mondiale della sanità (OMS) definì l'epidemia Covid-19 una pandemia l'11 marzo 2020.

[20] https://forbes.it/2021/04/28/ilde-forgione-la-social-manager-che-ha-rilanciato-gli-uffizi-grazie-a-tiktok/.

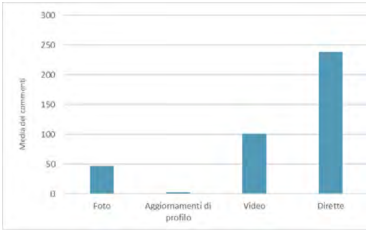
[21] https://www.uffizi.it/news/uffizi-facebook-2020.

[22] Data di apertura dell'*account*.

[23] Data in cui ha preso avvio il lavoro.

visione dei contenuti digitali utilizzate dal museo (foto, video, dirette *streaming* e, in piccolissima parte, *post* dedicati ad aggiornamenti del profilo social), siano le dirette a riscuotere il maggior numero di commenti, configurandosi di fatto come il mezzo digitale di condivisione del patrimonio culturale più adatto ai fini di successive realizzazioni di analisi di *customer satisfaction* che siano basate sull'esame dei commenti online come principale fonte delle opinioni degli utenti. In particolare, le indagini di *customer satisfaction* potrebbero essere condotte sia con un approccio simile a quello utilizzato nel presente lavoro, volto a sfruttare i *social data* così come offerti dai social network, sia con approcci più di stampo tradizionale, rivolgendo direttamente agli utenti in ascolto domande riguardo la soddisfazione al termine delle conferenze.

Fig. 1 Media dei commenti per modalità di condivisione dei contenuti sul social network.



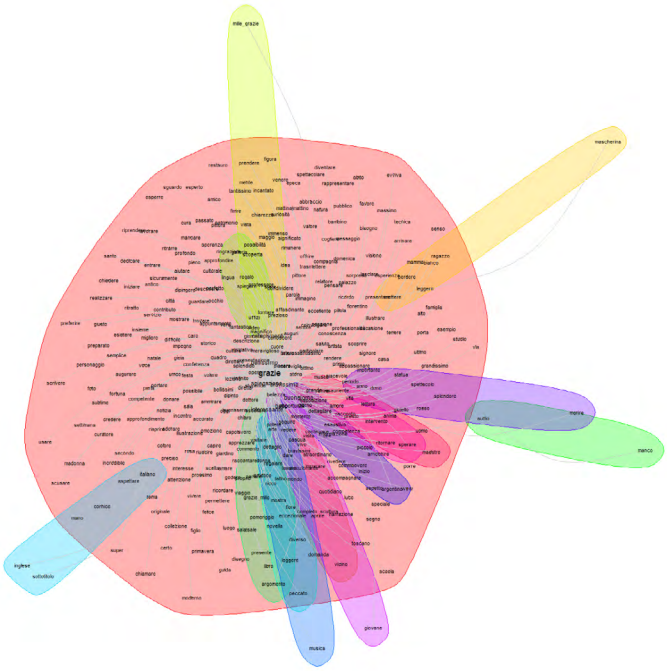
La trama della soddisfazione: *network analysis* e connessioni di parole

Oggetto delle prossime pagine sarà l'analisi del contenuto dei commenti rilasciati dagli utenti ai contenuti pubblicati dal museo nella propria pagina Facebook tra il 10 marzo 2020 e il 30 giugno 2021, eseguita in modalità automatica con l'utilizzo del software Iramuteq. L'approccio utilizzato, di tipo quali-quantitativo, è volto all'attuazione di una rilevazione della *customer satisfaction* allo scopo di valutare la performance delle Gallerie degli Uffizi relativamente alla percezione degli utenti del servizio offerto dal museo tramite il social network Facebook.

L'utilizzo, per l'analisi, di testi provenienti da un social network ha comportato la necessità di un'impegnativa fase di *pre-processing* del corpus testuale. Rispetto al totale dei commenti presenti nella pagina alla data del 18 agosto 2021, pari a 64.489, l'analisi testuale è stata applicata a 41.225 commenti, con l'esclusione di quelli che non contenevano testo, dei commenti in lingua non italiana, dei commenti scritti da parte dell'*account* delle Gallerie e di quelli rilasciati in risposta ad altri commenti. In

aggiunta, oltre all'esclusione delle tipologie di commenti sopra esposte, sono stati necessari interventi sui testi di tipo prevalentemente ortografico (che hanno riguardato, per esempio, azioni sulla punteggiatura o la trasformazione delle parole dal modo di scrivere '*social*' – più simile alla lingua parlata, dove spesso le parole appaiono allungate o abbreviate – a uno in linea con le regole della lingua scritta), in modo da consentire al software di riconoscere i significanti delle forme. Sono stati inoltre rimossi da ogni commento *emoji*, *emoticon*, *URL* e *hashtag*, sono state create alcune *multi-word expressions* ed è stata operata, tramite il software, una lemmatizzazione automatica di tutte le forme attive. Infine, il corpus dei commenti è stato classificato in base ad una variabile temporale riferita al mese di pubblicazione dei singoli contenuti digitali commentati, al fine di consentire analisi delle tendenze e, nel lungo periodo, confrontabilità periodiche dei risultati.

Fig. 2 Relazioni tra le forme nei commenti: connessioni e comunità di parole.



Per esaminare il contenuto dei commenti rilasciati dagli utenti e sondare dunque le percezioni di questi ultimi in merito alla fruizione del patrimonio culturale digitale delle Gallerie degli Uffizi, è stata innanzi-

tutto eseguita, con il software Iramuteq, una *network analysis* delle forme del corpus, volta ad individuare le relazioni tra le parole sulla base delle co-occorrenze presenti nei testi esaminati. La Fig. 2 mostra i risultati dell'analisi. Tramite una lettura consequenziale delle parole che nel grafico appaiono tra loro collegate, è possibile ricostruire i discorsi degli utenti sui contenuti digitali ed individuare, in base alle relazioni tra le parole e alla loro grandezza (nel grafico la grandezza delle parole è proporzionale alla loro frequenza nel corpus), quelle che contraddistinguono il maggior numero di commenti. Una piena comprensione del contenuto di questi ultimi non può in ogni caso prescindere da una successiva lettura qualitativa dei testi, che tuttavia può essere eseguita sulla base delle risultanti dell'analisi automatica, assumendo più che altro valore di supporto.

Nel grafico, le reti di relazione tra i termini sono rappresentate dalle linee che collegano le parole l'una all'altra e dalle diverse colorazioni delle nuvole all'interno delle quali le stesse sono contenute, che raffigurano le comunità e le sotto-comunità di parole.

Come è facile notare dalla Fig. 2, la maggior parte delle parole sono in relazione con la forma "grazie", che è anche la parola più frequente in assoluto nell'intero corpus.[24] Le parole a questa più collegate sono "bellissimo", "presentazione", "meraviglioso", "iniziativa", "cultura", "splendido", "direttore", "opera", "infinito", "lezione", "bellissimi", "dipinto", "spiegazione", "stupendo", "diretta", "bravissimo", "bellezza", "piacere", "interessantissimo", "bravo", "meraviglia", "ottimo", "storia", "meraviglioso".

La quasi totalità delle parole rappresentate nel grafico si trova posizionata all'interno della comunità di parole che include la forma "grazie", includendo tra queste anche i termini ricompresi in sotto-comunità che assumono, rispetto alla nuvola principale, la forma di sottoinsiemi. Tali sottoinsiemi sono quasi sempre interamente racchiusi nella comunità principale, con poche eccezioni (di cui si dirà a breve), connesse a precise scelte contenutistiche da parte del museo, perché concernenti determinate tematiche o di celebrazione di ricorrenze ed eventi particolari, o in quanto relative a riflessioni degli utenti a carattere fortemente specifico. Sono esempi le forme "latino", riguardante i video pubblicati dal museo in lingua latina; "fiore", in relazione ai video registrati nel Giardino di Boboli; "sala|sale", con riguardo alla serie di video intitolata la "#miasala" e dedicata a racconti-lezioni a cura dello *staff* del museo relativamente alle proprie sale preferite; "donna", concernente contenuti illustrati da o relativi a una figura femminile; "pasqua", con riferimento alle festività

[24] 18.429 frequenze.

pasquali; "anno", riferita al periodo pandemico; "dettaglio", riguardante i numerosi dettagli osservabili grazie alla fruizione digitale; "vedere vicino", in relazione alla possibilità offerta dalla fruizione digitale di vedere le opere più da vicino, o alle richieste agli operatori di mostrare qualcosa più da vicino; "continuare", attinente alle richieste di continuare ad offrire contenuti culturali digitali anche al termine della pandemia Covid-19; "tornare", con riferimento alla speranza o all'intenzione di tornare al museo, utilizzata soprattutto in relazione al contesto pandemico; "visitare", relativa alla speranza o all'intenzione di visitare il museo.

Alla luce dell'analisi eseguita, è possibile affermare che il lessico utilizzato dagli utenti nella manifestazione di interesse per la fruizione del patrimonio culturale digitale del museo si sostanzi nell'impiego di un linguaggio comune espressione del favore degli utenti nei confronti dell'offerta digitale, esprimendo prevalentemente opinioni positive e dimostrando, cioè, forte gradimento del servizio ricevuto.

Sebbene, in via generale, gli utenti percepiscano e valutino positivamente la qualità dei servizi digitali del museo, non mancano sfumature ed eccezioni. Nonostante nel nostro caso queste ultime si configurino in quanto numericamente limitate, una buona analisi di *customer satisfaction* dovrebbe ugualmente evidenziare aspetti positivi e negativi dei giudizi degli utilizzatori, ai fini di garantire loro adeguata considerazione nelle successive revisioni della *policy* digitale. Il grafico in Fig. 2 consente di individuare tali eccezioni senza troppa difficoltà, posizionandole in comunità di parole esterne o parzialmente esterne a quella principale. Tra esse troviamo le forme "inglese" e "sottotitolo", che si riferiscono a proteste connesse alla condivisione, da parte del museo, di alcuni contenuti in lingua inglese sprovvisti di sottotitoli in lingua italiana; le forme "audio" e "manco", che si riferiscono alla segnalazione di malfunzionamenti del suono nel corso delle dirette; la forma "musica", di cui talvolta si lamenta l'eccessivo volume; la forma "mascherina", che si riferisce alle critiche nei confronti di partecipanti ad un evento del dicembre 2020, che nonostante il momento delicato dovuto alla pandemia Covid-19, indossavano in maniera non appropriata la mascherina protettiva. La forma "morire", infine, necessita di una spiegazione a parte. Essa non figura infatti una critica nei confronti dei contenuti digitali, ma è fortemente connessa all'anno di pandemia e posizionata all'esterno della comunità di parole principale per via della sua accezione negativa.

In conclusione, il servizio culturale digitale offerto dal museo è generalmente percepito positivamente dagli utenti, i quali adoperano vocaboli che delineano opinioni in prevalenza positive, soffermandosi solo talvolta

su aspetti di critica o disappunto. Come risulta evidente dalla Fig. 2, il lessico che esprime tali opinioni negative è infatti in netta minoranza rispetto al vocabolario prevalente, oltre che scarsamente significativo.

Tendenze della *customer satisfaction*

A partire dall'ipotesi che la valutazione generalmente positiva degli utenti dei contenuti digitali data nel paragrafo precedente sia contraddistinta da stabilità nel tempo, in questa sezione è misurata la tendenza delle percezioni degli utenti rispetto al servizio digitale offerto dal museo. Per esaminare il trend delle percezioni degli utenti e confermare l'ipotesi della solidità della soddisfazione nel tempo, o piuttosto la stabilità di interesse nei confronti dell'offerta digitale nel periodo considerato – contraddistinto dall'alternanza di chiusure e riaperture del museo – è stato selezionato il lessico maggiormente adoperato nel complesso dei commenti sottoposti ad analisi automatica e rappresentato, in Fig. 3, il suo utilizzo medio rispetto al numero di commenti su base mensile nel periodo compreso tra marzo 2020 e giugno 2021. Il nucleo di parole impiegate per l'elaborazione del grafico è stato ricavato dall'analisi statistica eseguita dal software ed è altamente rappresentativo del corpus nel suo complesso. Sebbene, infatti, esso includa solo lo 0,26% delle forme grafiche attive presenti nei testi, le parole selezionate rappresentano oltre il 50% del totale delle frequenze dell'intero corpus.[25] Per tale ragione, si può sostenere che siffatto vocabolario sia una buona approssimazione del quadro delle percezioni espresse dagli utenti nei loro commenti e, quindi, della loro valutazione positiva dei contenuti digitali.

Le parole chiave impiegate per la costruzione del grafico in Fig. 3 sono "grazie"[26], "bellissimo"[27], "interessante"[28], "complimenti"[29], "bello"[30], "bravo"[31], "bravissimo"[32], "meraviglia"[33], "meraviglioso"[34], "grazie mille"[35],

[25] Per l'elaborazione del grafico sono state selezionate le forme con frequenza pari o superiore a 600.
[26] 18.429 frequenze.
[27] 3.380 frequenze.
[28] 3.168 frequenze.
[29] 2.947 frequenze.
[30] 2.505 frequenze.
[31] 1.900 frequenze.
[32] 1.802 frequenze.
[33] 1.252 frequenze.
[34] 1.490 frequenze.
[35] 1.581 frequenze.

"vedere"[36], "potere"[37], "splendido"[38], "stupendo"[39], "bellezza"[40], "spiegazione"[41], "iniziativa"[42], "grande"[43], "opera"[44], "uffizi"[45], "arte"[46], "piacere"[47], "storia"[48], "video"[49], "buongiorno"[50], "dottore"[51], "presentazione"[52], "buonasera"[53]. Tali parole rappresentano il nucleo principale delle opinioni degli utenti ed esprimono le tematiche ad esse sottese, a partire dalle emozioni e dalle conoscenze e informazioni acquisite, fino a giungere alle relazioni connesse all'esperienza del patrimonio culturale digitale, tematiche che saranno approfondite nel paragrafo sulle determinanti della *customer satisfaction*.

Come si può osservare nella Fig. 3, le linee che delineano il trend mensile dell'impiego nei commenti delle parole chiave selezionate e riportate sull'asse delle ascisse non presentano un'importante varianza dai valori medi caratterizzanti i singoli termini, ma mantengono nei mesi un andamento abbastanza lineare.

Allo scopo di svelare eventuali irregolarità significative nell'andamento della soddisfazione non rilevate dall'analisi delle parole chiave appena esposta in quanto dovute all'utilizzo di un lessico diverso rispetto a quello atteso, e quindi scostamenti più o meno significativi dal vocabolario prevalente mostrato in Fig. 3, è mostrata in Fig. 4 un'analisi delle corrispondenze delle forme del corpus rispetto alla variabile temporale "mese" nelle sue modalità, eseguita tramite il software Iramuteq, per esplorare non più il lessico prevalente e le tendenze nel suo impiego, bensì la specificità dei vocabolari alla base dei commenti rispetto ai singoli mesi nel periodo compreso tra marzo 2020 e giugno 2021.

[36] 1.124 frequenze.
[37] 1.314 frequenze.
[38] 714 frequenze.
[39] 669 frequenze.
[40] 1.007 frequenze.
[41] 2.715 frequenze.
[42] 778 frequenze.
[43] 1.300 frequenze.
[44] 1.653 frequenze.
[45] 1.129 frequenze.
[46] 1.260 frequenze.
[47] 785 frequenze.
[48] 697 frequenze.
[49] 602 frequenze.
[50] 3.100 frequenze.
[51] 680 frequenze.
[52] 690 frequenze.
[53] 795 frequenze.

Fig. 3 Regolarità nel lessico maggiormente utilizzato dagli utenti (periodo marzo 2020-giugno 2021).

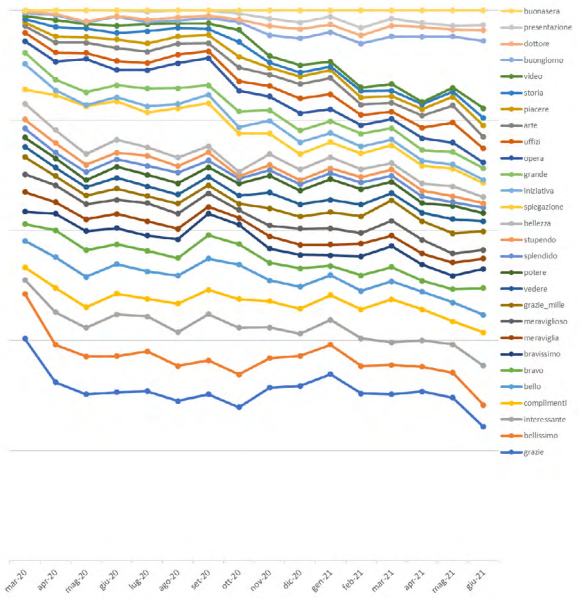
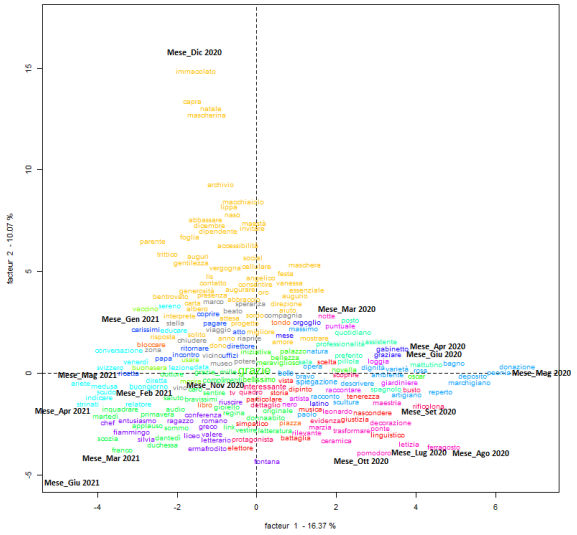


Fig. 4 Analisi delle corrispondenze: specificità dei vocabolari alla base dei commenti rispetto al mese di pubblicazione dei contenuti commentati (periodo marzo 2020-giugno 2021).



La Fig. 4 mostra la distribuzione delle forme del corpus su un piano cartesiano, dove l'approssimarsi o il distanziarsi delle parole rispetto al baricentro rappresentano, rispettivamente, la crescita e la decrescita della somiglianza tra i vocabolari rispetto alla variabile mese di riferimento. Se le parole concentrate nel punto di incontro tra i due assi sono quelle che figurano maggiori affinità tra i diversi periodi considerati, corrispondendo in parte al nucleo di parole chiave rappresentate in Fig. 3, le modalità della variabile temporale e i rispettivi termini dislocati verso l'esterno dei quadranti e lontano dagli altri mesi considerati consentono di cogliere la presenza di relazioni, la cui intensità è data dal valore χ^2 di associazione delle parole con la variabile temporale, che potrebbero segnalare incongruenze rispetto alla generale valutazione positiva da parte degli utenti dei contenuti digitali condivisi dal museo data nel precedente paragrafo.

La concentrazione della maggior parte delle parole verso il centro del piano in Fig. 4 mostra chiaramente l'assenza di tematiche salienti estranee a quelle rappresentate dal nucleo di parole chiave di cui alla precedente Fig. 3, e come ad allontanarsi dal baricentro siano piuttosto parole che, al contrario di quelle ricomprese nel nucleo chiave, si distanziano dalle tematiche comuni alla generalità dei commenti, delineando connessioni tra il corpus dei commenti e argomenti trattati in specifici contenuti digitali del museo, ed esprimendo quella che può essere definita come la capacità del patrimonio culturale digitale di coniugare l'esperienza digitale con quella reale. Si citano in tal senso la celebrazione delle ricorrenze (si vedano per esempio le forme "auguri", "natale", "rificolona", "ferragosto" e "dantedì"), e la condivisione, da parte del museo, di determinati contenuti che generano il sovrautilizzo di un certo lessico (si vedano le forme "capra" e "mascherina", relative all'evento che nel dicembre 2020 ha suscitato numerosi commenti "negativi", di cui si è parlato nel paragrafo sulla *network analysis*, o le forme "vaccino", "dicembre", connesse all'esperienza collettiva della pandemia). Infine, si citano le forme "deposito", "marchigiano", "trittico", "pomodoro", "rosa", anch'esse fortemente specifiche in quanto riferite a contenuti digitali specializzati.

Per un esame approfondito delle evidenze dell'analisi delle corrispondenze, sono riportati in Tab. 1 i profili lessicali più significativi per ciascun mese, ossia gli elenchi delle parole in ordine decrescente rispetto al valore di associazione tra le parole e i mesi dato dal χ^2. Dalla tabella, dove sono evidenziate in rosso alcune parole rappresentative di eventi e contenuti specifici per mese di riferimento, emergono i termini che generano le specificità dei periodi analizzati.

Tab.1 Profili lessicali per mese di riferimento. Vocabolari specifici che coniugano l'esperienza digitale con determinati eventi ed esperienze collettive o riguardano contenuti specializzati.

mar-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
grazie	37,9965	direttore	4,5149	attuale	2,3802
iniziativa	33,5381	momento	4,0145	splendore	2,3499
cucina	22,9266	meraviglioso	3,2675	splendido	2,3247
bellissimo	10,5659	virus	3,1895	mitologico	2,3237
pazienza	10,1319	messaggio	3,1832	consolare	2,3237
cuore	8,6793	casa	2,9371	grazie_mille	2,271
meraviglia	7,6964	stupendo	2,787	suggestivo	2,2169
bellezza	6,5479	collegare	2,7127	wow	2,2009
idea	6,5366	bello	2,707	spettacolo	2,1703
novella	6,0791	stare	2,6811	emozionante	2,1137
lodevole	5,6454	respirare	2,5236	fuoco	2,106
salvare	5,5338	acqua	2,5045	meravigliare	2,106
palazzo	5,3967	oscar	2,5045	vista	2,0619
prato	5,1311	preghiera	2,4513	ideale	2,0119
apr-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
pasqua	13,0252	fortuna	4,2055	occhio	2,5862
biblioteca	11,0564	bellissimo	4,2028	impressione	2,5635
spagnolo	10,8223	visitare	3,9153	imperdibile	2,5178
sala	10,4622	stupendo	3,8237	porre	2,499
bravo	9,743	stampa	3,7636	tesoro	2,4984
appuntamento	9,0681	esistenza	3,7247	mattina|mattino	2,4227
spiegazione	8,2394	bellezza	3,6967	affascinante	2,4165
quotidiano	7,9169	aprile	3,6369	serie	2,4064
giornata	6,2864	passeggiare	3,6369	meraviglioso	2,3911
chicca	5,9682	assistente	3,4858	scultura	2,3505
bello	5,8845	storia	3,3493	ringraziare	2,3039
professionalità	5,4814	tornare	3,3104	guida	2,2584
pillola	5,4814	bellissimi	3,0733	appassionare	2,241
direttore	5,1794	presto	3,0486	eleganza	2,2348

giorno	5,0626	cappella	2,9521	opera	2,1754
volume	4,9198	immagine	2,876	interessante	2,1635
maestoso	4,8533	corridoio	2,7948	semplice	2,0984
statua	4,8052	giornaliero	2,768	momento	2,0571
mattino	4,5744	scelta	2,7447	conoscere	2,048
mattutino	4,5304	limonaia	2,7393	acquisizione	2,0292
camelia	4,3877	tenero	2,7393	regale	2,0292
grazie	4,2663	struggente	2,7393	arricchire	2,0219
mag-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
spiegazione	40,4008	bellissimo	4,546	dedicare	2,5724
opera	39,7072	signor	4,2295	fruibile	2,5471
rosa	37,9539	descrivere	4,1794	varietà	2,4701
maggio	22,6682	scelta	3,7418	piacevole	2,4082
peonia	18,2378	antico	3,6453	autoritratto	2,4075
deposito	17,7624	spagnolo	3,6301	illustrare	2,3928
bellissimi	14,6067	artigiano	3,6133	ricco	2,3764
fiore	13,95	vasca	3,6103	appassionare	2,3748
mamma	13,0712	marmo	3,4049	gusto	2,3691
donazione	12,9572	delicatezza	3,2612	storia	2,3587
bravo	11,3552	scoperta	3,0895	sensibilità	2,3326
scultura	11,2281	paolo	3,0877	passeggiata	2,2705
salvatore	7,6404	limone	3,0787	sala|sale	2,2552
presentazione	7,4249	raffinato	3,0564	partecipazione	2,2399
bagno	7,1213	professionalità	3,0318	lottare	2,1831
giardino	6,9023	giorno	2,9462	cura	2,1804
conoscere	6,8991	sala	2,913	racconto	2,1798
marchigiano	6,5388	incantevole	2,8823	attento	2,1677
ambiente	6,2245	tesoro	2,8186	pianta	2,1466
bello	6,0512	mito	2,8016	maestria	2,1242
vanessa	5,999	orgoglioso	2,754	raffinatezza	2,1242
massimo	5,9935	commuovere	2,7379	sorprendere	2,1242
lavoratore	5,7129	gesso	2,7274	fattore	2,1075
ritratto	5,6884	profumo	2,7271	perfezione	2,1075
artista	5,1257	natura	2,7084	verde	2,0239
ottimo	4,8852	mattina|mattino	2,6972	ala	2,016
marca	4,882	dignità	2,6036	lode	2,016

collezione	4,7168	acconciatura	2,6036	decorazione	2,016
psiche	4,5685	descrizione	2,5918	commento	2,0049
giu-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
latino	32,6185	riaprire	4,4582	isola	2,7201
repubblica	28,4639	mese	4,4462	dizione	2,7201
sorpresa	12,1861	vivo	4,3906	novella	2,6642
continuare	11,8773	evviva	4,264	speciale	2,6278
festa	11,7441	esistenza	3,9592	scoprire	2,5737
graziare	10,4001	appartamento	3,9592	sentimento	2,5241
rosa	9,4726	ripresa	3,6924	appuntamento	2,4191
fotografico	8,0623	distanza	3,5924	antico	2,4
ritrarre	7,3237	ripartire	3,3539	campo	2,3863
spiegazione	7,2624	commuovere	3,2602	oscar	2,2919
lingua	7,2578	imperatore	3,2588	raffinato	2,2213
emozionante	5,7653	inizio	3,1909	difficile	2,1987
interessante	5,6833	mattutino	2,9976	bello	2,1693
pillola	5,4233	riascoltare	2,9586	geniale	2,1601
orgoglio	5,2025	giornaliero	2,9586	giorno	2,1176
profumo	5,0818	dettagliare	2,9112	efficace	2,0707
gabinetto	4,6621	assembramento	2,8979	bentornato	2,0457
giugno	4,5361	arma	2,7201	bellezza	2,0308
				esauriente	2,0086
lug-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
loggia	15,7255	ambiente	3,2341	dipingere	2,462
artista	9,0758	uomo	3,2095	lingua	2,4546
gnocchi	8,4816	traduzione	3,2058	passione	2,452
scarpa	7,9605	fronte	3,1795	cercare	2,3931
artemisia	7,6826	storia	3,1717	correggere	2,3647
piazza	7,016	stagione	3,1374	estetico	2,3647
sottotitolo	6,4911	arredo	3,1374	introduzione	2,3644
maestro	6,0797	inglese	3,0978	dolcissimo	2,3644
raccontare	5,8342	esprimere	3,0361	sguardo	2,3514
comprendere	4,9258	stupore	2,9778	effetto	2,2681
statua	4,9167	musica	2,8478	rilevante	2,1812

nero	4,7576	maddalena	2,8302	mille	2,1616
giardiniere	4,3682	rendere	2,8174	tela	2,0998
novella	4,2892	rappresentare	2,7266	allievo	2,0562
dipinto	4,0301	professor	2,7024	sindrome_di_stendhal	2,0562
sublime	3,9128	lupo	2,6983	pieno	2,0459
latino	3,82	paura	2,6983	personaggio	2,0365
donna	3,6266	animo	2,5989	ottimo	2,0273
impegnare	3,3194	attuale	2,5479	piangere	2,0206
spirituale	3,3194	importante	2,5331	onorare	2,0206
pittore	3,3072	conoscere	2,5238	sbaglio	2,0206
				commuovere	2,0228

ago-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
ferragosto	25,2307	donna	3,8212	superbo	2,3634
letizia	20,8211	giardiniere	3,5742	insieme	2,3115
leonardo	15,0808	ombra	3,5256	splendido	2,2873
spiegazione	14,4786	disegno	3,4935	vestire	2,2777
pomodoro	14,2083	fontana	3,3663	informazione	2,2229
marzia	9,4774	rilevante	3,1437	rosario	2,2155
ponte	9,3897	pietra	3,0734	collezione	2,1449
ritrarre	7,8793	esaustivo	3,0218	eccellenza	2,1412
ceramica	7,7758	musica	2,933	architettura	2,1412
trasformare	6,5032	maestria	2,8318	risultato	2,1412
accurato	6,3683	magnifico	2,7468	esposizione	2,1353
genio	5,7838	consiglio	2,5699	decorazione	2,0902
capolavoro	5,6444	commento	2,5341	scultoreo	2,0902
notte	5,2963	preferire	2,568	trasparire	2,0631
volume	4,4555	palazzo	2,4648	italiano	2,0518
opera	4,3288	bianco	2,4624	mezzo	2,0338
giocondo	4,2093	pianta	2,4605	giorno	2,0325
inglese	3,865	giovane	2,3832	tecnica	2,0097
				dettaglio	2,0008

set-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
rificolona	20,5977	dettaglio	3,5116	scheda	2,5745

autunno	14,3332	acqua	3,275	bello	2,5462
spiegazione	8,9502	adulto	3,1356	burbero	2,4108
cardellino	8,2566	elemento	3,1356	bimbo	2,3815
bambino	8,1813	finire	3,1004	precedente	2,2835
sfumatura	7,463	iconografia	2,8783	delizioso	2,2653
interessante	7,2786	dettagliare	2,8106	romano	2,2652
opera	6,4489	estate	2,7945	approfondire	2,18
colore	5,9468	viso	2,7887	misura	2,1696
rosso	5,2464	lingua	2,7858	proseguire	2,1696
pubblicare	4,8484	sfondo	2,7651	fiorentino	2,0831
cecilia	4,3437	minuzioso	2,7178	piacevolmente	2,0666
pala	4,286	amare	2,684	mostra	2,0638
descrizione	4,2632	busto	2,6604	bravo	2,0514
evidenza	3,9249	fede	2,6604	descrivere	2,0378
latino	3,9228	foto	2,6299	mettere	2,0273
figura	3,6162	scoprire	2,5771		

ott-20

Lessico	Valore χ^2 di associazione	Lessico	Valore χ^2 di associazione	Lessico	Valore χ^2 di associazione
tenerezza	14,287	paesaggio	3,7818	simpaticissimo	2,4712
leone	8,5666	descrizione	3,7065	raccontare	2,4406
giustizia	8,5346	diverso	3,5337	cultura	2,4045
dipinto	7,1033	punto	3,4436	abito	2,4007
donna	7,0808	vista	3,1858	parola	2,3976
voce	6,0535	spiegazione	3,152	divulgare	2,3688
battaglia	6,048	macchiaiolo	2,983	famiglia	2,3324
conferenza	5,9836	restauratore	2,983	sogno	2,303
storia	5,7867	risultato	2,983	nipote	2,1852
leonardo	5,7522	ricordare	2,8948	linguistico	2,1844
simpatico	5,5463	cominciare	2,845	burbero	2,1844
segno	5,5237	abbandonare	2,7802	soldo	2,1844
figura	4,9363	adatto	2,7025	novità	2,1844
musica	4,8827	particolare	2,6586	postare	2,1844
interessante	4,6993	ascoltare	2,6581	informazione	2,1335
scelta	4,695	evidenza	2,6039	rendere	2,1124
favola	4,4169	bello	2,5923	linguaggio	2,1045
elettore	4,3753	signore	2,5918	veloce	2,1023
video	4,35	palatino	2,5723	opificio	2,1023

restauro	4,1333	cogliere	2,5472	tema	2,0856
nascondere	4,0955	ricerca	2,5082	comprensione	2,0594
collega	3,9084	fattore	2,4712	chiave	2,0246
narrare	3,8245	carino	2,4712	grandissimo	2,0123
nov-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
violenza	19,3533	sezione	4,0258	insegnante	2,4733
parcheggiare	17,3154	educazione	3,9868	curatore	2,4108
ricordo	15,9592	alunno	3,9627	foto	2,3961
tribuna	13,4264	entrare	3,8696	raro	2,3855
grottesco	12,2176	stanza	3,8391	stare	2,3359
venere	10,139	pavimento	3,8082	esaustivo	2,3065
piazza	9,7376	bianco	3,8064	parola	2,2054
centro	8,5792	potere	3,4016	ultimo	2,2869
iniziativa	8,0556	pronto	3,1667	contesto	2,2703
didattico	7,501	peccato	3,0744	battaglia	2,2398
macchina	5,7126	mediceo	3,0545	turista	2,2398
girare	5,6459	messaggio	3,0393	differire	2,2398
artemisia	5,4884	padre	2,9919	toccante	2,2312
libro	5,4775	privilegio	2,9507	acquistare	2,2312
compleanno	5,3988	elementare	2,9307	esperimento	2,183
scuola	5,239	bloccare	2,8207	male	2,183
vedere	4,9926	dottore	2,8065	apprezzabile	2,183
problema	4,661	effettuare	2,7272	felicità	2,183
civiltà	4,5682	risolvere	2,7272	incisione	2,183
biblioteca	4,5389	impossibile	2,7019	esistere	2,1369
brutto	4,3118	esporre	2,6722	città	2,1366
compagno	4,2975	finestra	2,6106	tardo	2,124
acquisizione	4,2943	riprendere	2,5564	bisognare	2,0843
visita	4,2942	connessione	2,5469	prossimo	2,068
motivo	4,1234	duomo	2,5469	basso	2,0049
disegno	4,1074	momento	2,4917		
dic-20					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
natale	79,087	vergogna	4,9429	dottore	2,8576
auguri	70,2626	maestà	4,916	interprete	2,7653

mascherina	49,0972	lis	4,7673	albero	2,6967
festa	18,8355	oro	4,6323	notte	2,6967
immacolato	15,9629	cellulare	4,5586	tenere	2,6042
invitare	12,5512	social	4,3653	consentire	2,4905
lippa	11,7052	foglia	4,041	compagnia	2,4775
anno	10,9043	sereno	3,9292	gentilezza	2,3923
capra	10,7091	massimo	3,7631	direttore	2,3883
abbraccio	9,0718	angelico	3,7131	attesa	2,364
macchiaiaolo	8,8216	uffizi	3,6058	spazio	2,3603
naso	7,5842	speranza	3,517	cuore	2,2705
parente	7,3553	accessibilità	3,4334	beato	2,2333
augurare	7,2326	dipendente	3,4334	occupare	2,2295
archivio	7,1056	progetto	3,3952	dono	2,1226
vanessa	6,7034	augurio	3,3514	critico	2,1145
trittico	6,1086	maschera	3,2132	natura	2,0715
amore	5,9758	dicembre	3,1969	contatto	2,0708
buonasera	5,8557	differenza	3,0418	ineguagliabile	2,0708
abbassare	5,8116	migliore	2,9838	aiuto	2,0266
presenza	5,4599	pittura	2,8902	direzione	2,0082
iniziativa	5,2645	segno	2,8722		

gen-21

Lessico	Valore χ^2 di associazione	Lessico	Valore χ^2 di associazione	Lessico	Valore χ^2 di associazione
viaggio	13,7022	mangiare	4,0689	lezione	2,7105
ministra	12,1727	carne	3,9495	compagnia	2,6476
speranza	11,0917	stella	3,9495	cecilia	2,6459
riaprire	10,4215	manco	3,8942	viso	2,6403
potere	9,5821	scala	3,8417	interessantissimo	2,6355
conforto	9,1812	tornare	3,6772	leonardo	2,48
marco	8,322	madonna	3,6651	sentire	2,4704
evviva	7,041	favore	3,6201	seguire	2,4425
virtù	6,9806	vino	3,5664	tessuto	2,3357
anno	5,7842	uffizi	3,4219	creare	2,3015
diretta	5,3597	arte	3,4218	felice	2,2981
architettura	5,0638	sinistra	3,3898	virtuale	2,2949
museo	5,0345	continuare	3,3422	frutta	2,2632
regione	4,5723	zona	3,3316	sperare	2,2475
restauro	4,5544	beato	3,3301	muovere	2,2322

audio	4,5203	ministro	3,267	espressivo	2,2322
mano	4,3255	assembramento	3,267	memoria	2,2089
uscire	4,1838	segno	2,9962	teatro	2,2018
cibo	4,1363	vicino	2,9651	rivedere	2,1297
bentornato	4,1363	presto	2,9577	dicembre	2,0946
conservazione	4,1127	nostalgia	2,8166	cena	2,0946
domanda	4,1015	narrazione	2,7634	cammino	2,0946
online	4,0738	normale	2,7123	chiuso	2,0748
				viaggiare	2,0349
feb-21					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
buonasera	27,1997	statua	3,3305	provare	2,3404
buongiorno	19,9806	bianco	3,2904	sensibilità	2,3253
agrume	14,1522	potere	3,211	possibilità	2,2878
scienza	9,0224	pianta	3,1881	dottore	2,272
resilienza	8,762	venerdì	3,1777	magico	2,2686
vaso	8,1366	piatto	3,1739	apertura	2,241
osso	7,4869	strada	3,099	scultore	2,2348
cappello	7,0087	limone	3,0723	indossare	2,2348
moglie	5,7204	accordo	3,0723	opportunità	2,2052
drago	5,3688	interessantissimo	3,0625	venire	2,1935
grotta	5,1784	intervento	3,008	disponibilità	2,1707
vaccino	5,0709	metodo	2,9253	ragione	2,1606
saluto	4,7565	squadra	2,9253	buona_serata	2,1215
morire	4,6584	provare	2,8715	bravissimo	2,1189
frutto	4,4674	pomeriggio	2,8577	scientifico	2,1025
forca	3,9803	base	2,5975	accadere	2,1025
anima	3,9123	cordiale	2,3804	simpaticissimo	2,1025
psiche	3,6028	cardinale	2,3804	nutrimento	2,0118
dire	3,5513	infinito	2,361	basare	2,0118
mar-21					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
buongiorno	45,8183	dantedì	5,884	greco	2,9687
complimenti	35,4032	italiano	5,8605	vestire	2,9221
franco	23,4564	opportunità	5,7635	realizzare	2,8846

audio	21,4382	argentino	5,624	link	2,8732
primavera	18,8365	infinito	5,27	fantastico	2,8277
sentire	17,2631	scozia	4,9625	termine	2,6925
donna	17,0289	lingua	4,774	letteratura	2,6925
saluto	15,7302	diretta	4,4554	attore	2,4715
commedia	12,3597	relatore	4,3809	calice	2,4715
dottore	12,0646	funzionare	4,2909	bere	2,4715
caro	10,6724	pomeriggio	4,2772	regina	2,4607
grazie_mille	10,447	botanico	4,1706	ispirazione	2,3676
bravissimi	9,9107	inquadrare	3,7347	applauso	2,3448
marzo	9,0432	abito	3,6138	culturale	2,2954
poeta	8,4161	vetro	3,5578	età	2,2264
gioiello	8,0487	bronzare	3,4604	accattivante	2,2264
divino	7,5803	omaggio	3,353	piatto	2,2103
sommo	6,9272	interessantissimi	3,353	silvia	2,194
bravissimo	6,7131	sorridere	3,1456	venere	2,0932
duchessa	6,1747	lezione	3,0665	vittoria	2,062
martedì	6,1415	entusiasmo	3,0346	cotanto	2,062
				bimbo	2,0219

apr-21					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
buongiorno	106,7616	relatore	5,7879	malta	2,6502
pasqua	103,3679	grazie_mille	4,7311	santo	2,6164
medusa	20,3757	pomeriggio	4,1607	narrare	2,5904
diretta	14,2989	venerdì	4,0517	diffondere	2,5788
cristo	13,0749	saluto	4,0465	gesso	2,4901
auguri	12,0658	camelia	3,9646	caro	2,486
lezione	12,0066	resistenza	3,689	villa	2,419
buonasera	11,249	complimenti	3,4866	raggiungere	2,335
compleanno	8,8271	annunciazione	3,4239	gita	2,335
scudo	8,751	svizzero	3,4239	vaso	2,3026
strinati	8,2979	seguire	2,9695	guardare	2,294
ariete	8,2979	parente	2,8311	ammirazione	2,2851
professore	7,4743	staff	2,7747	scultura	2,2298
croce	7,0869	collaboratore	2,7339	indicare	2,1565
sereno	6,4458	mediceo	2,7082	dio	2,0782
dottore	6,2869	voce	2,7032	argentino	2,0625

				interessantissimo	2,0147
mag-21					
Lessico	**Valore X^2 di associazione**	**Lessico**	**Valore X^2 di associazione**	**Lessico**	**Valore X^2 di associazione**
buongiorno	20,0419	ritornare	4,1937	allestimento	2,5855
russo	12,7925	vasariano	4,1162	riproduzione	2,5697
uffizi	10,2665	museo	3,8884	visibilità	2,5697
direttore	9,9733	granduca	3,7765	incontro	2,4763
buonasera	9,754	autoritratto	3,4928	venire	2,4151
maggio	8,0537	arte	3,4357	orario	2,414
diretta	7,0093	pubblicità	3,4322	contribuire	2,4057
madre	6,3218	maniera	3,3567	figlio	2,382
galleria	6,0165	diritto	3,2426	vedere	2,3514
rosso	5,8087	culturale	3,224	papa	2,3099
notizia	5,6812	grazie_mille	2,9922	mantenere	2,3099
digitale	5,2761	iniziare	2,9584	dedizione	2,2746
fiorentino	5,2649	comprare	2,7756	valorizzare	2,244
scorso	5,2019	primo	2,7011	brillante	2,2128
strozzare	4,8631	pelle	2,6455	settimana	2,125
sala|sale	4,6	nascere	2,6086	sabato	2,1218
ultimo	4,527	lunedì	2,5855	coprire	2,0342
ricetta	4,2215	fonte	2,5855	arricchire	2,0325
giu-21					
Lessico	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**	**Lessico**	**Valore χ^2 di associazione**
buongiorno	32,8714	porta	4,4446	letterario	2,7427
ragazzo	19,0655	bravo	4,2479	piacere	2,7127
fiammingo	12,8602	eco	4,1067	chef	2,5531
liceo	11,4296	passione	4,0613	napoleone	2,4744
ermafrodito	9,7685	romano	3,8701	santo	2,4107
silvia	9,5394	studente	3,6864	mitico	2,3866
andrea	8,9386	onore	3,6601	audio	2,3813
pomeriggio	8,8759	interessante	3,6024	generale	2,2321
bravissimi	8,7656	greco	3,5036	creativo	2,2321
conferenza	6,829	togliere	3,4525	foto	2,2177
entusiasmo	6,5536	galleria	3,1588	maschile	2,1286
sentire	5,9825	programma	3,1504	insegnante	2,1058

buonasera	5,4458	povero	3,0605	riuscire	2,0898
fontana	5,1243	video	2,8885	valere	2,0443
archeologico	5,1243	dottore	2,8044	nonno	2,0342
trittico	5,0009	congratulazione	2,784		

Le determinanti della *customer satisfaction*: l'esperienza digitale

Dopo aver individuato il lessico prevalente espressione della soddisfazione degli utenti e i vocabolari specifici connessi all'esperienza digitale, in questa sezione sono presentati alcuni esempi di commenti scelti a partire da quello che è stato definito essere il nucleo del lessico maggiormente impiegato nella fruizione digitale del caso studio in esame (cfr. Fig. 3). A partire dal nucleo di parole chiave e con il supporto degli esempi di commenti, sono determinate le tematiche alla base dei discorsi degli utenti che identificano le aspettative sul servizio museale e, dunque, i connotati dell'offerta digitale determinanti la soddisfazione.

Il primo tema risultante dall'analisi dei commenti e delle parole chiave è legato alle emozioni. L'offerta digitale del museo, che si sostanzia in una varietà di stimoli, dalla mera visione delle opere d'arte riprodotte virtualmente a momenti più dinamici che implicano una maggiore partecipazione dell'utente, invitato ad ascoltare spiegazioni ad opera di esperti e, talvolta, ad intervenire con domande, produce contenuti culturali digitali che commuovono, appassionano e coinvolgono, suscitando ricordi e la voglia di vedere dal vivo il museo, Firenze e i capolavori che insieme custodiscono. Si riportano alcuni esempi di commenti.

> **Splendido** per chiarezza espositiva e per l'accostamento alle fonti, una **spiegazione** eccellente che mi ha commosso (commento #32175, contenuto del 25 luglio 2020).

> **Bravissimo**! Che bella **spiegazione**! Non noiosa ma coinvolgente e appassionante. **Grazie_mille** (commento #4148, contenuto del 30 aprile 2021).

> Spero di **potere vedere** dal vero questo capolavoro e grazie per avere dato **spiegazioni** dettagliate e chiare (commento #31020, contenuto del 23 agosto 2020).

> Pazzesco è stato emozionante un approfondimento straordinario infinitamente **grazie grazie** (commento #22538, contenuto del 23 dicembre 2020).

> **Grazie** di cuore **spiegazione** molto chiara e coinvolgente (commento #32807, contenuto del 11 luglio 2020).

> Fiore **meraviglioso** mi ricorda tantissimo la mia infanzia (commento #7378, contenuto del 10 aprile 2021).
>
> **Grazie** per ciò che ci illustri tutto questo però fa crescere dentro di me l'amore per la toscana terra **meravigliosa** dove cultura e **bellezza** fanno da padrona (commento #16467, contenuto del 11 febbraio 2021).
>
> Buongiorno che **bello grazie_mille** (commento #58, contenuto del 28 giugno 2021).
>
> **Bellissima** iniziativa che invoglia a **venire** di persona per gustare dal vivo queste **meraviglie** (commento #25046, contenuto del 27 novembre 2020).
>
> Grazie è stato **bellissimo** ascoltare la **spiegazione** d'arte con la sua voce **complimenti** (commento #28780, contenuto del 15 ottobre 2020).
>
> Amo molto l'entusiasmo con cui la **dottoressa** spiega e fa venire la voglia di **venire** subito a firenze (commento #4872, contenuto del 27 aprile 2021).
>
> **Grazie** vi seguo con passione (commento #19992, contenuto del 18 gennaio 2021).

Il secondo tema riguarda l'identificazione culturale e la generazione di connessioni tra passato e presente derivanti dalla visione virtuale dei capolavori del passato – che hanno regalato, all'Italia in generale e alla Toscana in particolare, un immenso patrimonio artistico e culturale – e dall'ascolto delle storie raccontate con le voci degli esperti. A seguire alcuni esempi di commenti.

> Queste sono le dirette che dobbiamo **vedere** e condividere con i nostri figli e nipoti **arte** mito leggenda poesia **storia** e natura si fondono in una **meravigliosa** armonia con un tocco di mistero un'**opera** che non smette mai di stupirci (commento #10371, contenuto del 23 marzo 2021).
>
> **Spiegazione** ineccepibile e di grande impatto emotivo per il legame che hai restituito tra storia del passato e realtà attuale! **Grazie** Vanessa! **Bellissimo** quadro oggi lo guardiamo con occhi diversi. **Grazie** a voi e a tutti gli infermieri (commento #37352, contenuto del 3 maggio 2020).
>
> **Grazie** di cuore per la **bellezza** la cultura e la speranza che diffondete a tutti noi l'**arte** la cultura umanistica e scientifica dovrebbero essere il pane e la speranza di un popolo civile e progredito buon anno (commento #21679, contenuto del 2 gennaio 2021).
>
> **Brava** commento entusiasmante ci fa ricordare la nascita della nostra repubblica il diritto di voto alle donne bravissima (commento #35063, contenuto del 2 giugno 2020).

> **Grazie** per le **spiegazioni** sempre utili per conoscere la nostra **storia** buona giornata (commento #27309, contenuto del 8 novembre 2020).

> Stupefatta! **Interessanti** i collegamenti con il mondo di oggi. Complimenti per la fluida **spiegazione** in lingua latina che così torna sorprendentemente a vivere (commento #30168, contenuto del 12 settembre 2020).

Il terzo tema è connesso all'esperienza del patrimonio culturale digitale in quanto momento di intrattenimento e svago, un appuntamento che suscita interesse e che permette di evadere dalla quotidianità. Si riportano a seguire alcuni esempi di commenti.

> **Stupendo**. Relax tra arte e natura (commento #33218, contenuto del 2 luglio 2020).

> **Grazie** per la bellissima **iniziativa** (commento #32858, contenuto del 9 luglio 2020).

> **Grazie_mille** molto **interessante** che gioia questi appuntamenti con voi (commento #21601, contenuto del 3 gennaio 2021).

> **Grazie** per questo appuntamento al quale non posso più rinunciare (commento #29732, contenuto del 24 settembre 2020).

> **Grazie** il commento è **bellissimo** ed esauriente la stampa è splendida spero che continuerete ad offrirci queste pillole di conoscenza dei tesori degli **uffizi** anche dopo l'emergenza (commento #37970, contenuto del 27 aprile 2020).

> **Grazie** un bel viaggio raccontato con passione (commento #5726, contenuto del 21 aprile 2021).

> È un gran **piacere** la mattina ascoltare con una tazzina di caffè a casa tante cose **interessanti grazie** (commento #2400, contenuto del 23 maggio 2021).

> **Grazie** a te adoro ascoltarvi mentre faccio colazione (commento #29029, contenuto del 10 ottobre 2020).

Il quarto tema è relativo all'esperienza culturale in quanto momento di accrescimento delle proprie conoscenze e acquisizione di informazioni. Le guide virtuali non solo permettono di ascoltare – e riascoltare – spiegazioni ad un livello di minuzia che difficilmente accompagna la fruizione fisica, ma consentono anche di scoprire dettagli prima impensabili. A seguire alcuni esempi di commenti.

> **Bellissima presentazione** e **spiegazione** di tanti dettagli che arricchiscono la cultura di chi vede ed ascolta **grazie** anche per oggi (commento #37271, contenuto del 5 maggio 2020).

> La sua **spiegazione** molto dettagliata ci aiuta a comprendere il lavoro preparatorio eseguito da leonardo e successivamente modificato (commento #30993, contenuto del 23 agosto 2020).
>
> Che dire **grazie** per questa perla invisibile degli **uffizi** disegni scientifici ma pieni di poesia come sempre la relatrice nella **presentazione** unisce alla conoscenza l'entusiasmo e l'orgoglio quasi da padrona di casa (commento #2681, contenuto del 17 maggio 2021).
>
> In effetti siamo dei privilegiati perché possiamo scrutare i minimi dettagli ed abbiamo delle guide eccezionali **grazie** (commento #15410, contenuto del 19 febbraio 2021).
>
> Eccellente interpretazione è un'**opera bellissima** delle mie preferite (commento 32801, contenuto del 11 luglio 2020).
>
> Bellissimo **video interessante** anche l'aspetto didattico le gallerie degli **uffizi** stanno facendo un'azione divulgativa incredibile ed unica **complimenti** (commento #28551, contenuto del 22 ottobre 2020).
>
> Un racconto visivo coinvolgente e quante informazioni **interessanti grazie** (commento #21073, contenuto del 9 gennaio 2021).
>
> **Grazie** per questa panoramica luminosa ed esaustiva suggerirò questo intervento ai miei studenti (commento #1545, contenuto del 2 giugno 2021).
>
> Spero che queste magnifiche **spiegazioni** che sono vere lezioni d'**arte** continueranno nel tempo futuro (commento #2302, contenuto del 24 maggio 2021).

Il quinto e ultimo tema riguarda le relazioni sociali alla base della fruizione digitale. Sebbene quest'ultima sia un'esperienza per lo più individuale, gli utenti creano legami con le proprie vicende private e, soprattutto, riconoscono il carattere collettivo della fruizione, costituendosi in quanto social *community*. Si riportano a seguire alcuni esempi di commenti.

> **Buongiorno**, **Uffizi**! **Buongiorno**, colleghi di "**lezione**" (commento #1909, contenuto del 31 maggio 2021).
>
> **Grazie**, avevo dedicato una foto di questa statua a mio figlio papà per la seconda volta in una mia visita agli **Uffizi** ma non ne sapevo la **storia** (commento #37804, contenuto del 29 aprile 2020).
>
> Visto che non abbiamo fatto quest'anno la 100 del Passatore causa Covid questa **iniziativa** rinsalda i nostri legami (commento #32861, contenuto del 9 luglio 2020).

> Buona_serata a tutti gli amici della **Galleria** (commento #29133, contenuto del 7 ottobre 2020).
>
> **Buongiorno** a tutti modo appassionato quello che usa la relatrice quindi un grande **piacere** essere con voi e questi volumi sono una **meraviglia** (commento #82, contenuto del 28 giugno 2021).
>
> **Grazie** per questi mesi in vostra compagnia davvero tutti eccezionali **grazie grazie grazie** e continuate mi raccomando (commento #34828, contenuto del 3 giugno 2020).
>
> **Buonasera** sempre un **piacere** essere con voi per queste lezioni di **arte grazie** di cuore (commento #4579, contenuto del 28 aprile 2021).
>
> **Grazie bellissima lezione** che **meraviglia** questa ora passata insieme (commento #9542, contenuto del 24 marzo 2021).
>
> **Grazie spiegazione** molto **interessante** la riascolterò stasera insieme ai miei figli buon natale a lei e a tutti i suoi collaboratori (commento #22420, contenuto del 24 dicembre 2020).

In definitiva, possiamo concludere che la *customer satisfaction*, nel caso studio considerato, sia determinata dall'esperienza, nella fruizione digitale del patrimonio culturale, di emozioni, identificazione culturale, intrattenimento, nuove informazioni, conoscenze e relazioni sociali.

Considerazioni conclusive

Oggetto del capitolo è stata un'analisi di *customer satisfaction* realizzata mediante un approccio quali-quantitativo, che ha in parte visto l'applicazione di tecniche di text mining, allo scopo di valutare la performance della policy digitale dell'istituto culturale delle Gallerie degli Uffizi relativamente alle percezioni degli utenti rispetto al servizio online ricevuto e all'invariabilità o mutevolezza nel tempo di tali percezioni.

Secondo la principale normativa nazionale in materia di performance organizzativa, infatti, la rilevazione del grado di soddisfazione dei destinatari delle attività e dei servizi assume rilevanza affinché siano realizzate valutazioni della performance più partecipate e relative anche all'esperienza digitale, dove le percezioni dei cittadini siano una misura dell'efficacia qualitativa delle policy.

Sulla base del medesimo orientamento, gli Istituti culturali sono stati chiamati, a partire dal 2020, a predisporre appositi strumenti di rilevazione del grado di soddisfazione degli utenti, che consentano una confrontabilità almeno annuale dei risultati.

Il presente lavoro vuole, in primo luogo, mettere in luce il valore pubblico prodotto dall'esperienza digitale degli utenti del museo degli Uffizi e, in secondo luogo, incentivare l'analisi dei contenuti di tale esperienza per valutare e, conseguentemente, migliorare l'accesso e la fruizione dei servizi online tramite strategie innovative in grado di sfruttare l'immenso – ma scarsamente valorizzato – patrimonio di dati custoditi dai social network.

In considerazione della rilevanza acquisita dal tema della digitalizzazione del patrimonio culturale tra le priorità politiche del Ministero della cultura, la previsione per il prossimo futuro è infatti quella di un incremento delle policy di digitalizzazione dei musei, che comporterà maggiore necessità di realizzare indagini di *customer* orientate anche alla fruizione digitale.

Nel capitolo è stata presentata un'analisi di *customer satisfaction*, replicabile in altri contesti, a partire dai commenti rilasciati dagli utenti ai contenuti digitali condivisi dalle Gallerie degli Uffizi nella propria pagina sul social network Facebook, al fine di valutare la performance dell'ente museale relativamente alla percezione degli utenti di un servizio digitale.

Nello specifico, tramite un'analisi lessico-testuale automatica del contenuto dei commenti e una successiva lettura dei vocabolari e di alcuni esempi dei testi, sono state prodotte evidenze in merito alla soddisfazione degli utenti del servizio ricevuto e a come questa sia rimasta costante nel periodo considerato, nonostante le variazioni del contesto relative all'evoluzione della pandemia.

In conclusione, l'implementazione di uno strumento permanente di *digital customer satisfaction analysis* quale quello qui proposto potrebbe essere esteso a tutte le altre piattaforme online del museo, sia nell'ottica di favorire il raggiungimento dei propri obiettivi di performance contenuti nei Piani triennali, sia nel tentativo di supportare un miglioramento dell'efficacia della policy. L'impiego delle opinioni degli utenti nei processi di adeguamento delle decisioni andrebbe quindi inteso come strumento ordinario per integrare la *customer* nell'erogazione di servizi, nelle strutture fisiche come in ambiente digitale. L'analisi automatica del contenuto dei commenti presentata nel capitolo ha infatti mostrato come, a costi bassi – pur in considerazione della necessità di una più o meno onerosa fase di *pre-processing* dei dati – sia possibile analizzare, a livelli variabili di dettaglio a seconda delle immediate esigenze, e confrontare nel tempo, ma anche in tempo reale, i feedback degli utenti, raccogliendo evidenze circa le percezioni, cioè l'efficacia delle politiche, e indicazioni in merito alle aspettative, cioè il miglioramento dell'efficacia delle policy.

La valutazione di progetti da finanziare. Uno scenario possibile

Antonio A. Aggio[1]

Progettazione finanziata, Formazione professionale, Valutazione, Salienza, Regione Veneto.

Introduzione

Possiamo usare il text mining nelle fasi istruttorie di procedure concorsuali che prevedono la presentazione di documenti testuali? E affidarci all'analisi automatica dei testi per conoscere e classificare i contenuti, ad esempio, di centinaia di progetti? E magari spingerci fino alla stessa valutazione dei progetti, attribuendo un punteggio sulla misura della salienza di alcune parole chiave nei testi esaminati? E se addirittura volessimo indagare se gli elaborati pervenuti sono più o meno simili tra loro, per scartare i testi 'fotocopia' e premiare quelli più originali? Se aggiungiamo

[1] Funzionario della Regione del Veneto dal 1997, è stato capo segreteria degli assessorati alla Formazione professionale e alle Politiche sociali e programmazione sociosanitaria, responsabile dell'Ufficio informatizzazione presso la Segreteria della Giunta e dell'Ufficio coordinamento internet, coordinatore tecnico dell'Osservatorio regionale per le politiche sociali. Attualmente è responsabile dell'Ufficio gestione e innovazione presso la Direzione Formazione e Istruzione. Da alcuni anni insegna al laboratorio di integrazione socio-sanitaria del Master EMAS dell'Università Ca' Foscari di Venezia. Ha collaborato nella stesura del capitolo con Paola Bolzonello, che si è occupata delle applicazioni del pacchetto Stylo di R per l'analisi dell'originalità dei progetti.

poi la necessità di poter condurre queste operazioni in tempi rapidi e di ridurre la soggettività del valutatore, il text mining è lo strumento che può fare al caso nostro.

In questo capitolo è descritta la conduzione di un'esperienza che ha visto la realizzazione di un'analisi del contenuto tramite il software Iramuteq su un corpus di indagine composto dai progetti presentati a valere su un bando della Formazione professionale della Regione del Veneto emanato nel 2018. È doveroso premettere che la tecnica qui di seguito descritta è del tutto sperimentale e che il suo utilizzo non è né previsto né disciplinato dalle vigenti normative. La scelta del corpus deriva dal fatto che il database contenente i testi e le variabili da analizzare era già nella disponibilità dell'Amministrazione regionale, configurandosi di fatto come una miniera pronta per l'estrazione.

Su questo insieme di progetti ci siamo posti l'obiettivo di eseguire in maniera automatica una parte del processo di valutazione utilizzando alcune delle tecniche descritte in questo volume.

Il capitolo si articola in tre parti principali. A seguito della descrizione del corpus in esame, saranno presentate una *topic detection* con metodo Reinert del corpus dei progetti per l'individuazione di classi semantiche, e una successiva analisi delle corrispondenze applicata ai *cluster*, al fine di comprendere la vicinanza o lontananza dei *topics* precedentemente determinati. Queste operazioni non sono strettamente finalizzate alla valutazione dei progetti, ma sono utili per raggiungere una visione d'insieme dei *topics* caratterizzanti le varie progettualità. Successivamente, passando dal generale al particolare, saranno illustrate analisi delle corrispondenze sul corpus, per verificare il livello di associazione tra il contenuto dei progetti finanziati e alcune variabili significative dei progetti, quali linea progettuale, provincia di provenienza dell'ente presentatore, e tipologia di attività economica. Infine, tramite l'individuazione di alcune parole chiave (*keyword*), scelte tra quelle che più rappresentano gli obiettivi posti dal bando e dal *Goal* 8 'Incentivare una crescita economica duratura, inclusiva e sostenibile, un'occupazione piena e produttiva e un lavoro dignitoso per tutti' dell'Agenda 2030 (al quale il bando è connesso), e tramite la misura del grado di correlazione tra le parole chiave e i progetti, sarà presentata l'analisi della salienza – in termini statistici – dei temi individuati dalle *keyword* nei progetti. La correlazione tra *keyword* e progetti, che dovrebbe permettere di valutare la coerenza dei progetti con il bando, è espressa con un valore numerico (il χ^2), a partire dal quale si è tentato di costruire un'ipotesi di punteggio attribuibile in corso di istrut-

toria. In ultimo, prima delle considerazioni conclusive, sarà brevemente presentato un tentativo di individuazione di progetti fotocopia.

I progetti presentati

Il corpus è stato ricavato dai progetti presentati da organismi di formazione a valere sul bando della Regione Veneto 'Protagonisti del cambiamento. Strumenti per le persone e le organizzazioni', dedicato alla formazione professionale. Si tratta di un'iniziativa particolarmente significativa e complessa emanata nel 2018 che, a partire dall'obiettivo tematico 8 del Programma Operativo POR FSE 2014-2020, ha definito i criteri per la presentazione di progetti di formazione rivolti alla promozione di una occupazione sostenibile e di qualità e a sostenere la mobilità dei lavoratori. L'obiettivo del bando è peraltro in linea con il *Goal* 8 di Agenda 2030 'Incentivare una crescita economica duratura, inclusiva e sostenibile, un'occupazione piena e produttiva e un lavoro dignitoso per tutti'.

In generale, il processo di presentazione dei progetti avviene tramite una procedura *web-based*, che prevede la presenza di una piattaforma web nella quale gli organismi di formazione devono caricare tutti gli elementi costitutivi delle proprie proposte di progetti di formazione quali, ad esempio: presentatore, titolo, descrizione, obiettivi formativi, partenariato, metodologie utilizzate per la formazione, piano economico, ecc. Ogni record del database così costruito contiene decine di informazioni, che possono essere espresse in numeri, testi brevi selezionati sulla base di un nomenclatore predefinito, o testi lunghi liberi.

La procedura concorsuale di cui sopra, che ha raccolto e valutato al termine 474 progetti di formazione, ammettendone a finanziamento 208, prevedeva l'indicazione, per ciascun progetto, di alcuni elementi testuali, strutturati secondo numerose variabili, tra le quali, ai fini del presente lavoro, sono state selezionate ed estratte le seguenti:

- un numero identificativo anonimizzato del progetto;
- un codice anonimizzato dell'ente presentatore;
- la provincia in cui è localizzato l'intervento formativo;
- il codice Attività economica riclassificato secondo la classificazione IGRUE[2] relativo all'ambito di appartenenza del soggetto proponente;
- la linea del progetto (il bando prevedeva la presentazione di

[2] IGRUE è l'acronimo di Ispettorato Generale per i Rapporti finanziari con l'Unione Europea.

progetti all'interno di una delle quattro direttrici specifiche che saranno descritte nel seguito del capitolo);
- l'unione dei campi contenenti la descrizione del progetto, le motivazioni dell'intervento, gli obiettivi formativi, e la descrizione dei destinatari (dati testuali che hanno costituito la base per la creazione del corpus).

Il metodo e le analisi che saranno presentati nel capitolo mirano ad inserirsi all'interno del processo di valutazione delle progettualità condotto dalla Regione, puntando ad automatizzare e a rendere maggiormente oggettive alcune fasi del processo. I progetti presentati nell'ambito di un bando sono infatti soggetti a procedure di valutazione condotte secondo i criteri definiti dal bando stesso: esperito un primo controllo formale, i progetti sono sottoposti alla valutazione di merito da parte di un nucleo di esperti appositamente costituito. La valutazione di merito è condotta sulla base dei parametri illustrati in una griglia di valutazione e dà luogo alla formulazione di una graduatoria basata sui punteggi attribuiti. Nel bando utilizzato ai fini del presente lavoro, i parametri di valutazione prendevano in considerazione:

- la finalità della proposta (coerenza con le esigenze specifiche del territorio, analisi delle necessità di sviluppo di competenze dei destinatari, grado di incidenza del progetto nella soluzione di problemi occupazionali, descrizione dell'impatto del progetto nel tessuto economico, descrizione dei fabbisogni cui il progetto intende rispondere. Il punteggio attribuibile va da 0 a 10 punti);
- gli obiettivi progettuali (grado di coerenza della proposta con il Programma operativo e l'Obiettivo specifico del bando. Punteggio da 0 a 10);
- la qualità della proposta (in termini di chiarezza espositiva, completezza ed esaustività, dettaglio delle singole fasi. Punteggio da 0 a 10);
- innovatività delle metodologie didattiche e qualità delle metodologie di monitoraggio degli esiti (punteggio da 0 a 10);
- i *partner* coinvolti (punteggio da 0 a 10);
- il grado di realizzazione di attività pregresse (punteggio da 0 a 5).

Il metodo innovativo di valutazione qui proposto si pone, in particolare, in ausilio al parametro b relativo alla valutazione della coerenza delle

proposte progettuali con gli obiettivi del bando.

Nel seguito del capitolo saranno dettagliatamente descritte le fasi del processo sperimentale.

La predisposizione dei dati

I progetti sono stati raccolti tramite il software gestionale *web-based* descritto nel precedente paragrafo, che restituisce un'estrazione delle variabili in formato foglio di lavoro. Il nostro file contiene 474 righe, una per progetto, ed una colonna per ciascuna variabile, o campo. I medesimi dati sono stati successivamente combinati e restituiti nella forma richiesta da Iramuteq, il software utilizzato per l'analisi testuale. Come anticipato, il corpus dei progetti è stato classificato secondo una serie di variabili (che sono state utilizzate per la realizzazione di alcune delle analisi che saranno presentate nel corso del capitolo), quali un numero identificativo dei progetti ("idcorpus"), il codice dell'ente presentatore[3] ("ente"), la linea di intervento su cui si basa il progetto tra le quattro contenute nel bando in esame ("linea"), la provincia in cui ha sede il progetto ("provincia") e il codice Ateco dell'ente di formazione ("codice Ateco").

I contenuti dei progetti presentati

Al corpus dei progetti è stata innanzitutto applicata una *topic detection* tramite metodo Reinert, che ha individuato la presenza di 3 classi semantiche, o mondi lessicali.

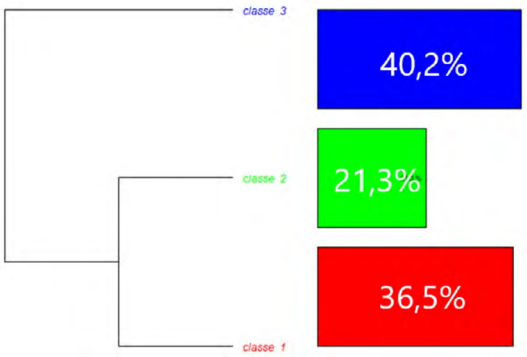
> «Un mondo lessicale consiste in uno specifico vocabolario che eredita le sue proprietà dai contenuti dei progetti, ed è estratto dal corpus senza alcuna informazione a priori» (Sbalchiero et al. 2016, pag. 1338).

La classificazione del corpus in classi semantiche ha permesso di investigare con rapidità l'intero corpus dei progetti e classificare i contenuti all'interno di un limitato numero di *cluster* di significato prevalente, realizzando così una prima valutazione delle caratteristiche generali del corpus selezionato. E, ricordiamolo, senza la minima influenza da parte dell'analista.

Per ognuno dei tre *cluster* prevalenti di contenuto individuati dal software (Fig. 1) siamo subito in grado di capire l'ampiezza in termini di percentuali di testi (progetti) che vi si possono ascrivere.

[3] Per questo lavoro, tutti i codici identificativi degli enti sono stati anonimizzati.

Fig. 1 Dendrogramma a cascata che illustra l'ampiezza delle classi semantiche.



Il dendrogramma in Fig. 2 illustra i principali contenuti dei *cluster*, ossia dei progetti, e la relazione gerarchica tra gli stessi. La dimensione delle parole nel raggruppamento ad albero è proporzionale alla significatività delle parole nel *cluster*. Due sono gli ambiti contenutistici di fondo che caratterizzano i progetti: il mondo dell'impresa (suddiviso a sua volta tra produzione e *marketing*) e quello dei lavoratori.

Un'analisi delle corrispondenze applicata ai *cluster* (Fig. 3) mostra su un piano cartesiano la distanza, calcolata con il metodo statistico del χ^2, tra le 3 classi tematiche sopra individuate.

Come si può osservare dal grafico, le classi appaiono distanti tra loro e ben identificabili, confermando la presenza nel corpus di due *topics* ben distinti, uno dei quali coinvolge, con le dovute differenze che saranno a breve illustrate, due delle tre classi contenute nel corpus.

Dev'essere chiaro che la posizione delle classi nel piano cartesiano non rappresenta in nessun caso un giudizio di merito, ma denota la composizione delle singole classi e la significatività al loro interno di determinate parole, che tanto più rappresentano il *cluster* di riferimento e i contenuti dei progetti che vi fanno parte quanto più si allontanano dagli assi (Fig. 3).

Fig. 2 Dendrogramma a cascata con le parole che caratterizzano maggiormente i mondi semantici.

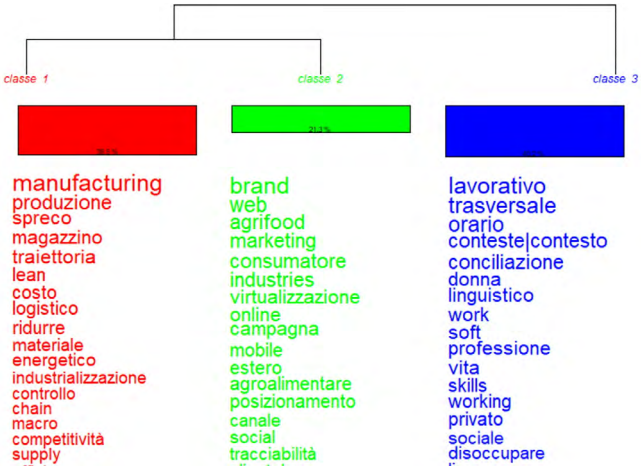
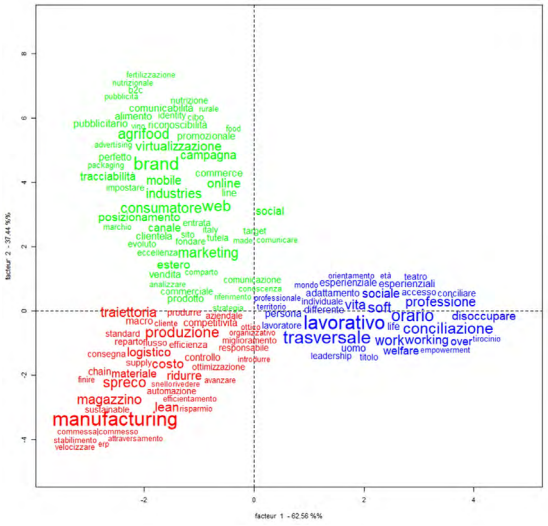


Fig. 3 Posizionamento dei mondi lessicali sul piano cartesiano.



La dimensione delle parole è proporzionale alla loro significatività nella classe.

Nel grafico in Fig. 3, la classe 1 (di colore rosso) e la classe 2 (di colore verde) riguardano progetti che hanno a che fare con il mondo dell'impresa, si trovano sul lato sinistro degli assi cartesiani e sono più complementari tra loro. Se la classe 1 tratta i temi della produzione, la classe 2 è imperniata sul *marketing* e sul consumatore. La classe 3 (di colore blu), che si trova invece sulla parte destra del piano, riguarda progetti rivolti direttamente ai lavoratori.

Nel seguito, tramite l'utilizzo delle parole più significative riportate in Fig. 2, sono brevemente descritti i contenuti dei tre *cluster*.

Il primo *cluster* (classe 1) è prevalentemente legato al mondo produttivo aziendale. Il suo contenuto (che caratterizza il 38,5% dei progetti analizzati) è lessicalmente più vicino a temi quali:

> "manifattura", "produzione", "ridurre" (lo) "spreco", "*lean*", "logistica", "industrializzazione", "controllo", "*supply chain*", "costo" (del) "materiale", "competitività".

Il secondo *cluster* (classe 2), più piccolo, si riferisce prevalentemente a temi legati all'immagine dell'azienda, alla comunicazione, al posizionamento sul web e alla qualità e tracciabilità dei prodotti. Esso caratterizza il 21,3% dei progetti, e contiene molteplici riferimenti a concetti quali:

> "web" ("canale", "social", "online", "*mobile*", "posizionamento", "virtualizzazione"), "*agrifood*" ("*industries*"), "*marketing*", "consumatore", "virtualizzazione", "campagna", "tracciabilità".

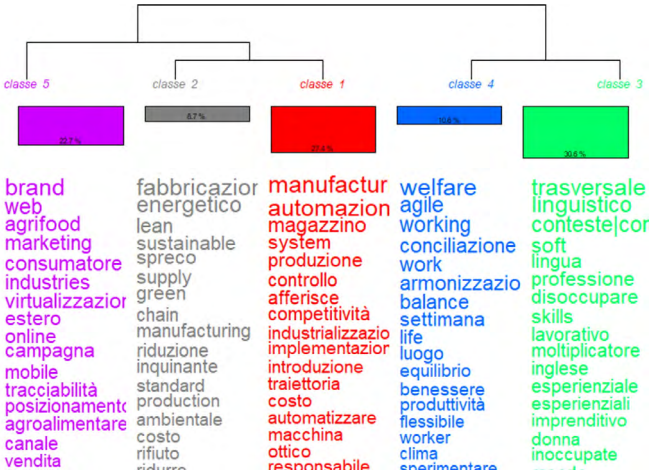
I *cluster* 1 e 2 sono rappresentati nei dendrogrammi in Fig. 1 e 2 come "figli" dello stesso ramo, e in effetti si differenziano entrambi dalla classe 3, che non presenta parole collegate al mondo dell'impresa, ma a quello dei lavoratori. Il 40,2% dei progetti, pertanto, contiene parole più collegate alle esperienze delle persone quali:

> "orario" ("lavorativo"), "contesto" ("lavorativo"), "conciliazione", "professione", "vita", "*working skills*" (competenze), "sociale", "privato", "disoccupazione".

Allo scopo di esplorare ulteriormente il corpus, la *topic detection* è stata ripetuta impostando una diversa configurazione dei parametri di Iramuteq, che hanno prodotto una classificazione discendente del corpus in cinque *cluster*, come rappresentati nella Fig. 4.

Nonostante i *cluster* siano passati da tre a cinque, vi riconosciamo comunque l'iniziale tripartizione (cfr. Fig. 2) e la successiva scissione della classe legata al mondo produttivo aziendale e di quella connessa al mondo dei lavoratori in due classi. Il *cluster* della produttività si scinde in una classe riguardante "tecniche di produzione" e "competitività" e in una concernente la "sostenibilità ambientale". Il *cluster* riguardante i lavoratori si dipana a sua volta in due sottoclassi: una prevalentemente dedicata al benessere e alla conciliazione dei tempi di vita e di lavoro, e l'altra all'acquisizione di competenze, lavorative e linguistiche, con riferimento anche alla parità di genere e allo *status* occupazionale.

Fig. 4 Dendrogramma ottenuto con una differente impostazione dei parametri del software, che ha esteso a cascata la prima classificazione.



Similarità e differenze dei progetti presentati

Se le analisi Reinert hanno fornito informazioni utili per meglio comprendere il contenuto delle progettualità analizzate, in questo paragrafo sono illustrate tre analisi delle corrispondenze, applicate questa volta al corpus nella sua interezza e non ai *cluster*. Tali analisi sono state realizzate a partire da alcune delle variabili e relative modalità con le quali sono stati classificati i testi, descritte nel paragrafo 2 del presente capitolo e relative alla linea di intervento, alla provincia di svolgimento del corso e al codice Ateco dell'organismo di formazione, al fine di individuare, rispetto alle suddette variabili, le specificità dei progetti, ossia ricercare similarità e differenze tra i testi.

I grafici presentati nel seguito sono stati elaborati a partire da una tabella di contingenza, predisposta automaticamente dal software, di cui si riporta un estratto in Fig. 5.

Fig. 5 Esempio di tabella di contingenza.

Forms | Banal forms | POS | **Forms frequencies** | POS frequencies | Forms relative

formes	*linea...	*linea_2	*linea_3	*linea_4
più	1251	195	520	176
produzione	1219	24	59	11
intervento	1167	206	345	153
gestione	1122	130	135	47
essere|sonare	1119	158	385	161
innovazione	1042	106	139	42
anche	1017	206	273	136
ad	1008	170	259	165
sistema	1001	93	114	21
marketing	956	8	58	37
modello	954	231	96	18
strumento	945	131	184	51
attività	929	161	207	102
cambiamento	905	200	506	82
produttivo	895	60	137	31
tutto	886	145	169	69
come	884	122	251	132
attraverso	879	85	167	97

Per ogni variabile (linea di intervento) è rappresentato il numero di volte in cui compare la forma (parola) indicata a destra. Sono presenti anche le forme supplementari.

Analisi delle corrispondenze applicata alla variabile "linea di intervento". Dalle quattro linee di intervento emergono le peculiarità dei progetti

La prima analisi delle corrispondenze (Fig. 6) è stata elaborata utilizzando la variabile "linea di intervento". Nel grafico in Fig. 6, le parole appaiono posizionate all'interno di quattro gruppi, le cui coordinate, che ne determinano la posizione nel piano cartesiano, sono calcolate dal software in ragione della distanza statistica data dal valore χ^2 di associazione dei singoli termini componenti il corpus con la variabile di riferimento.

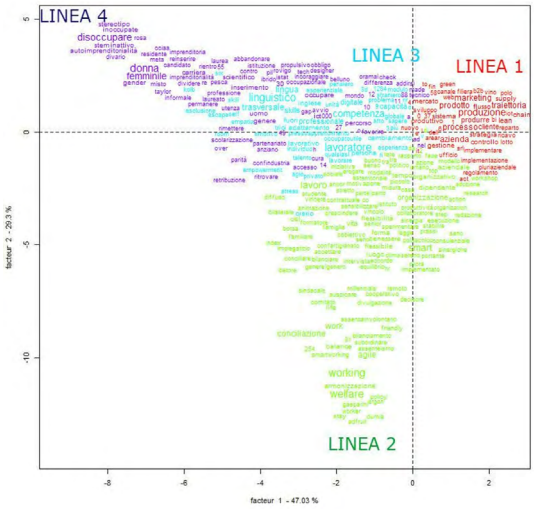
La Fig. 6 mostra le parole più significative all'interno di ciascuna linea progettuale. Le parole che più si posizionano all'esterno del grafico sono quelle che meglio rappresentano le differenze tra i gruppi. Al contrario, quelle posizionate centralmente figurano le analogie tra di essi.

Riconosciamo, nei vocabolari dei quattro gruppi, le quattro linee di intervento previste dal bando:

- linea 1 (imprese che cambiano i modelli di *business*): "produzione", "processo", "filiera", "*marketing*", "mercato", "controllo", "cliente", ecc.;
- linea 2 (imprese che introducono modelli più flessibili di lavori): "*welfare*", "*smart|smart working*", "agile", "conciliazione",

"equilibrio", "genere", "sindacati", "benessere", "armonizzazione", ecc.;
- linea 3 (sostegno ai lavoratori per l'adattamento e il cambiamento): "lingua", "competenza" (professionale), "straniero", "sapere", "trasversale", "empatia", "apprendimento", ecc.;
- linea 4 (sostegno alle donne attraverso nuove opportunità di occupazione): "autoimprenditorialità", "disoccupazione"/"inoccupazione", "scolarizzazione"/"laurea", "abbandonare", "reinserire", "incoraggiare", "retribuzione", ecc.

Fig. 6 Analisi delle corrispondenze per la variabile "linea di intervento".



Nel piano cartesiano sono riportate le parole più significative per ciascuna modalità della variabile.

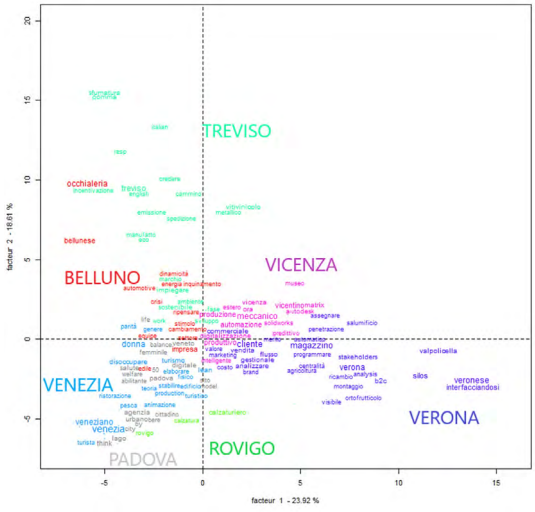
Le parole più distanti tra di loro sono "produzione", "*welfare*" e "disoccupazione", più tipiche delle rispettive linee 1, 2 e 4, contrariamente ad "azienda", "lavoro" e "lavoratore", che si posizionano nell'area centrale del grafico e si possono pertanto trovare con maggiore frequenza in tutte le linee progettuali.

Analisi delle corrispondenze applicata alla variabile "provincia". Ogni provincia veneta esprime le sue peculiarità

La Fig. 7 mostra le corrispondenze testuali dei progetti per la variabile "provincia". Le parole più significative in base alla provincia di rife-

rimento sono “occhialeria” a Belluno, “calzaturiero” a Rovigo, “turismo” ma anche “disoccupazione” a Venezia, “meccanica”, “digitalizzazione” e “automazione” a Vicenza, “viticoltura” a Treviso, “agricoltura”, “ortofrutticolo” e “magazzino” a Verona.

Fig. 7 Analisi delle corrispondenze per la variabile “provincia”.



Nel piano cartesiano sono riportate le parole più significative per ciascuna modalità della variabile.

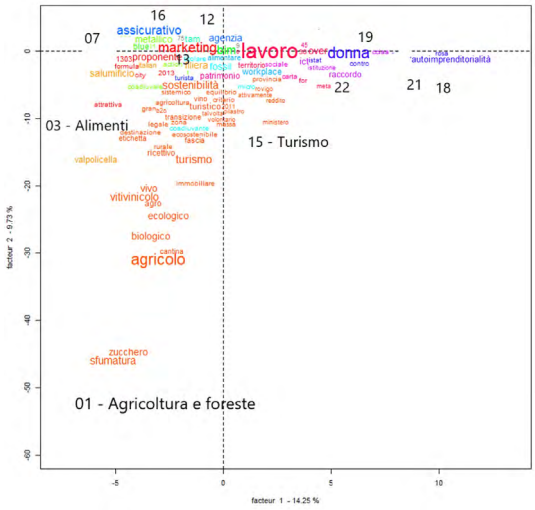
Analisi delle corrispondenze applicata alla variabile “codice Ateco”. Agricoltura, turismo e alimentazione: i progetti più originali del corpus

La Fig. 8 mostra i risultati dell’analisi delle corrispondenze rispetto alla variabile “codice Ateco” (riclassificato secondo la classificazione IGRUE), il cui grafico, che si appiattisce verso l’alto, rappresenta la distribuzione delle attività economiche di riferimento degli organismi di formazione. Le parole contenute nei quadranti in basso, che maggiormente si distanziano dalle altre – configurando la presenza, in questi settori, di progettualità particolarmente originali – si riferiscono prevalentemente alle modalità “Agricoltura e foreste” (codice Ateco 01), “Turismo, servizi di alloggio e di ristorazione” (codice Ateco 15) e “Industrie alimentari e delle bevande” (codice Ateco 03). Questi tre gruppi, appartenenti a una sorta di area “agrituristica”, presentano una variabilità di parole superiore al resto del corpus.

Il quadrante in alto a sinistra concentra invece le parole tra loro più vicine, e vi riconosciamo i codici Ateco corrispondenti ad attività industriali, manifatturiere e del terziario: "Fabbricazione di computer e prodotti di elettronica e ottica" (codice Ateco 06), "Altre industrie manifatturiere non specificate" (codice Ateco 07), "Energia elettrica, gas, vapore, acqua calda e aria condizionata" (codice Ateco 10), "Trasporti e stoccaggio (codice Ateco 12), "Azioni di informazione e comunicazione, comprese le telecomunicazioni, le attività dei servizi d'informazione, la programmazione informatica, la consulenza e le attività connesse" (codice Ateco 13), "Attività finanziarie e assicurative" (codice Ateco 16).

In alto a destra troviamo, infine, le attività istituzionali e sociali: "Pubblica amministrazione" (Ateco 18), "Istruzione" (Ateco 19), "Attività di assistenza sociale, servizi pubblici, sociali e personali" (Ateco 21) e infine le "Attività connesse all'ambiente e ai cambiamenti climatici" (Ateco 22).

Fig. 8 Analisi delle corrispondenze per la variabile "codice Ateco".



Nel piano cartesiano sono riportate le parole più significative per alcune modalità della variabile. Per esigenze tipografiche, non è stato possibile includere nel grafico tutte le modalità ricomprese nella variabile.

La coerenza dei progetti rispetto al bando e all'Agenda 2030

Alla luce delle evidenze prodotte dalle analisi delle corrispondenze, che hanno visto l'emergere di peculiarità riferite alle variabili di volta in volta esaminate, in questa sezione è misurata la salienza statistica nel

corpus di alcune *keyword*, selezionate in quanto rappresentative degli obiettivi posti dal bando sul quale sono stati presentati i progetti analizzati e dal *Goal* 8 dell'Agenza 2030. Misurando il valore χ^2 di associazione tra i gruppi di progetti (dove il gruppo è rappresentato dai progetti che afferiscono alle medesime modalità delle variabili considerate, ad esempio gruppo dei progetti di Belluno, nel caso di progetti presentati nella provincia di Belluno in base alla variabile "provincia"), e le *keyword*, è stato definito il grado di correlazione (e coerenza) tra progettualità da una parte e obiettivi del bando e *Goal* 8 dell'Agenda 2030 dall'altra.

Successivamente, utilizzando il valore χ^2 di associazione, in questo caso non più tra gruppi di progetti, ma tra singole progettualità, e *keyword*, è stata costruita l'ipotesi di punteggio attribuibile in corso di istruttoria.

La scelta delle parole chiave è un'azione totalmente rimessa alla discrezionalità del ricercatore.

Ai fini del presente studio, sono stati individuati due gruppi di parole chiave che fanno riferimento:

- al *Goal* 8 di Agenda 2030 'Incentivare una crescita economica duratura, inclusiva e sostenibile, un'occupazione piena e produttiva e un lavoro dignitoso per tutti' e ad alcuni dei *target* in cui lo stesso è suddiviso;
- al bando sul quale sono stati presentati i progetti analizzati nel corso del capitolo e, in particolare, all'Obiettivo tematico 8 del Fondo sociale europeo 2014-2020 'Promuovere un'occupazione sostenibile e di qualità e sostenere la mobilità dei lavoratori', al quale il bando fa riferimento.

Le *keyword* selezionate sono illustrate nella Tab. 1.

Tab. 1 Elenco delle *keyword* selezionate per l'analisi della salienza dei contenuti dei progetti con il *Goal* 8 di Agenda 2030 e le finalità del bando.

***KEYWORD* AGENDA 2030**		***KEYWORD* BANDO**	
Competenze	Sostenibilità	Strategie innovative	Adattamento
Tecniche	Piena occupazione	Nuovi modelli	Potenziamento
Professionali	Lavoro dignitoso	Innovazione prodotto	Competenze trasversali
Capacità imprenditoriale	Uomini	Innovazione processo	Competenze digitali
Vulnerabili	Donne	Cambiamento	Competenze linguistiche
Uguaglianza di genere	Disabili	Sperimentare	Approccio al lavoro globale
Qualità tecnica	Disabilità	Flessibili	Genere
Crescita economica	Parità	Armonizzazione	Progressioni carriera
Inclusione	Retribuzione	Conciliazione	Donne
Lavoro dignitoso	Giovani	Generi	Valore
Imprenditorialità	Disoccupati	Generazioni	
Creatività	Studi	Smart work	
Crescita economica	Formazione		
Degrado ambientale	Turismo sostenibile		

I grafici in Fig. 9, 10 e 11 mostrano il grado di associazione calcolato in base al valore del χ^2 tra i due insiemi di parole chiave e le variabili con cui è stato classificato il corpus – rispettivamente, "linea di intervento", "Provincia" e "codice Ateco"– nelle loro modalità. In altri termini: quanto i progetti, considerati per linea, per provincia o per codice Ateco, sono correlati con le parole chiave?

Analizzando i grafici generati, desta particolare curiosità il risultato mostrato in Fig. 9: i progetti della linea 1 (che, ricordiamo, promuovono il cambiamento dei modelli di *business* delle imprese) sembrano trovare poca corrispondenza con le *keyword*. Se possono aver poco a che fare con il *Goal* 8 di Agenda 2030, li troviamo anche 'distanti' dai temi chiave del bando. Al contrario, per i progetti delle linee 3 e 4 (sostegno ai lavoratori e sostegno alle donne) il grado di salienza è netto.

La Fig. 10 mostra come la salienza delle *keyword* nei progetti risulti maggiore nelle attività che sono state presentate da soggetti appartenenti alle aree "Pubblica amministrazione" (codice Ateco 18), "Istruzione" (codice Ateco 19), "Attività di assistenza sociale, servizi pubblici, sociali e personali" (codice Ateco 21), oltre che alla tipologia residuale 24.

L'analisi per provincia (Fig. 11) è parimenti curiosa: i progetti di Verona e Vicenza sono distanti sia da Agenda 2030 che dalle *keyword* del bando, diversamente da quanto accade per quelli di Venezia e di Padova.

Fig. 9 Grado di salienza delle *keyword* per variabile "linea di intervento".

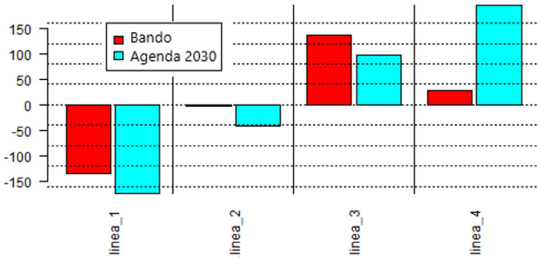


Fig. 10 Grado di salienza delle *keyword* per variabile "codice Ateco".

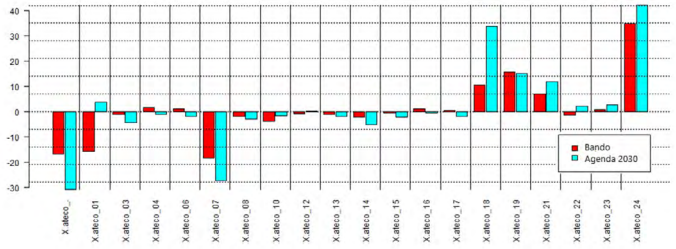
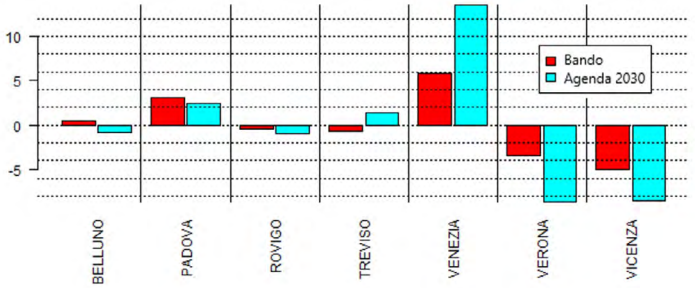


Fig. 11 Grado di salienza delle *keyword* per variabile "provincia".



Un punteggio per ciascun progetto

Da ultimo, la stessa analisi di salienza delle parole chiave è stata eseguita sui singoli progetti (anziché sui progetti raggruppati nelle variabili linea di intervento, Ateco e provincia). Tramite questa operazione si è tentato di attribuire un punteggio ai progetti, basato sul valore χ^2 della correlazione tra gli stessi e le *keyword*. Considerato l'alto numero dei progetti componenti il corpus, pari a 474, non è possibile una rappresen-

tazione grafica analoga alle figure nel precedente paragrafo. Eseguite le analisi sui singoli progetti con il software, in questa sede sono presentati pertanto i soli risultati in forma aggregata (Fig. 12).

Come illustrato in figura, i valori del χ^2, distribuiti in percentili, generati dal calcolo della salienza delle parole chiave di Agenda 2030 sui singoli progetti sono compresi tra -64.201 e 162.954, con una mediana di -4.065, mentre quelli generati dal calcolo della salienza delle parole chiave del bando sui singoli progetti vanno da -126.692 a 121.941, con un valore mediano di -2.879. Metà dei progetti, quindi, hanno una misura di correlazione negativa rispetto alle *keyword* selezionate.

Fig. 12 Distribuzione in percentili dei valori χ^2 relativi alla salienza nei singoli progetti delle *keyword* di Agenda 2030 e del bando.

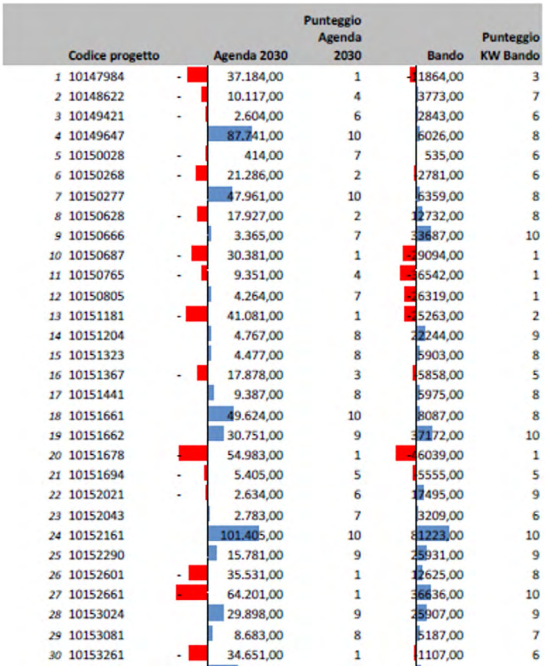
	VALORI CHI2 PER SALIENZA	
PERCENTILI	**AGENDA 2030**	**BANDO**
minimo	-64.201,00	-126.692,00
percentile 10	-25.667,20	-26.254,20
percentile 20	-17.907,40	-16.705,20
percentile 30	-11.594,40	-9.763,00
percentile 40	-7.089,00	-6.074,40
mediana	-4.065,00	-2.879,00
percentile 60	-788,60	3.263,00
percentile 70	4.472,20	5.686,00
percentile 80	9.685,00	12.829,40
percentile 90	33.208,40	31.074,80
max	162.954,00	121.941,00

È evidente che, per utilizzare lo strumento di analisi qui presentato ai fini di un supporto alla fase di valutazione di merito delle proposte progettuali, tali valori non possano essere utilizzati per costruire punteggi di valutazione dei progetti. Una soluzione possibile, che mutua la scala 0-10 abbinata ai parametri presenti negli attuali bandi (e che non farebbe scartare *tout court* i progetti con un valore χ^2 negativo), potrebbe essere quella di suddividere in decili i punteggi ottenuti, e attribuire un punto al primo decile, due punti al secondo, tre al terzo, e così via, fino al valore massimo.

L'attribuzione del punteggio per decili forzerebbe la distribuzione dei progetti in 10 gruppi di eguali dimensioni, ottenendo comunque lo scopo di valorizzare i progetti nei quali la salienza delle *keyword* è più elevata.

La Fig. 13 rappresenta un esempio di come potrebbero essere illustrate sia le risultanze ottenute, sia la funzione di calcolo della salienza sulle parole chiave del bando e di Agenda 2030.

Fig. 13. Esempio di attribuzione di punteggi in base al percentile di appartenenza dei valori χ^2 di salienza dei progetti.



	Codice progetto	Agenda 2030	Punteggio Agenda 2030	Bando	Punteggio KW Bando
1	10147984	37.184,00	1	1864,00	3
2	10148622	10.117,00	4	3773,00	7
3	10149421	2.604,00	6	2843,00	6
4	10149647	87.741,00	10	6026,00	8
5	10150028	414,00	7	535,00	6
6	10150268	21.286,00	2	2781,00	6
7	10150277	47.961,00	10	6359,00	8
8	10150628	17.927,00	2	2732,00	8
9	10150666	3.365,00	7	3687,00	10
10	10150687	30.381,00	1	9094,00	1
11	10150765	9.351,00	4	6542,00	1
12	10150805	4.264,00	7	6319,00	1
13	10151181	41.081,00	1	5263,00	2
14	10151204	4.767,00	8	2244,00	9
15	10151323	4.477,00	8	5903,00	8
16	10151367	17.878,00	3	5858,00	5
17	10151441	9.387,00	8	5975,00	8
18	10151661	49.624,00	10	8087,00	8
19	10151662	30.751,00	9	371,72,00	10
20	10151678	54.983,00	1	6039,00	1
21	10151694	5.405,00	5	5555,00	5
22	10152021	2.634,00	6	7495,00	9
23	10152043	2.783,00	7	3209,00	6
24	10152161	101.405,00	10	81223,00	10
25	10152290	15.781,00	9	5931,00	9
26	10152601	35.531,00	1	2625,00	8
27	10152661	64.201,00	1	6836,00	10
28	10153024	29.898,00	9	5907,00	9
29	10153081	8.683,00	8	5187,00	7
30	10153261	34.651,00	1	1107,00	6

La figura si ferma alla riga 30 delle 474 righe totali generate.

La modellizzazione del processo

In una prospettiva di modellizzazione del processo adottato ai fini di un suo possibile utilizzo all'interno della pubblica amministrazione, si descrivono in Tab. 2 le principali fasi delle sequenze operative fin qui descritte, con una stima della tempistica necessaria per ciascuna operazione.

Tab. 2 Modellizzazione del metodo di valutazione dei progetti

FASE	DESCRIZIONE	TEMPO
I	Esportazione del database in formato foglio di calcolo dal sistema gestionale SIU. Il corpus dei testi viene predisposto per il caricamento e, una volta caricato, si verifica che non presenti errori e sia trattabile dal software.	1 ora

II	Caricamento in Iramuteq. Analisi statistica del testo: conteggio delle occorrenze (totale delle parole presenti, delle parole differenti e delle parole utilizzate per la prima volta). Lemmatizzazione del testo.	1 minuto
III	Applicazione del metodo Reinert per la classificazione dei testi in classi semantiche, o mondi lessicali. Produzione di matrici e grafici.	da 1 a 5 minuti
IV	Primo commento sui risultati: descrizione delle aree semantiche.	30 minuti
V	Analisi delle corrispondenze: mappatura delle similarità e delle differenze all'interno di gruppi di progetti (ad esempio per linea di intervento) e distribuzione in un grafico cartesiano delle parole con riferimento ai gruppi.	da 1 a 5 minuti
VI	Definizione delle *keyword* e misurazione della distanza intertestuale dei testi con le *keyword.*	30 minuti
VII	Esportazione dei risultati in formato CSV, elaborazione, produzione di punteggi ed *attachment* dei punteggi al file CSV originale per la prosecuzione delle fasi istruttorie.	<30 minuti
VIII	Eventuale creazione del *subcorpus* per linea progettuale e ripetizione delle fasi da II a VI.	<30 minuti per fase

Quanto si assomigliano i progetti tra loro? Due software a confronto

In quest'ultimo paragrafo è presentato un tentativo di valutazione delle progettualità volto all'individuazione di progetti fotocopia. L'obiettivo è mostrare come l'analisi automatica del testo possa essere di ausilio ai processi di valutazione qualora fosse necessaria la selezione di progetti in base all'originalità o meno dei contenuti.

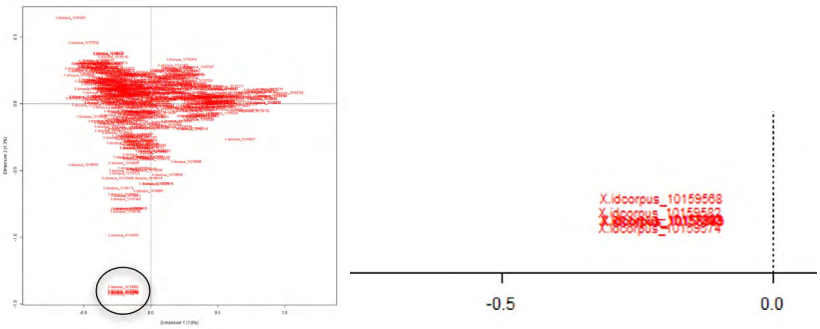
Tramite un'analisi delle corrispondenze con Iramuteq applicata alla variabile "codice del progetto" (cfr. paragrafo 2), sono stati ricercati i progetti più o meno "simili" tra di loro. L'esito è rappresentato in Fig. 14, dove, come già ampiamente spiegato nei paragrafi precedenti, le similarità possono essere individuate grazie al posizionamento delle parole nel grafico.

I progetti evidenziati con il cerchio, periferici ma sovrapposti tra loro, sono quelli per i quali ci attendiamo una somiglianza, da verificare empiricamente.

Per farlo, è sufficiente aprire il corpus alla fonte per ricercarvi materialmente i progetti con i codici anonimizzati (…)9568, (…)9582 e (…)9574, che riconosciamo dall'immagine. Si tratta di tre progetti presentati sulla

linea 1, da tre organismi di formazione distinti, con sede di svolgimento in tre province diverse.

Fig. 14 Posizionamento geometrico della variabile “codice progetto”. A destra: ingrandimento dell’area evidenziata dal cerchio, che permette la lettura dei codici dei progetti selezionati.



In Tab. 3 sono riportati gli *incipit* di alcuni dei campi descrittivi dei progetti (motivazione, obiettivi formativi e azioni complementari), confermando come l’algoritmo abbia riconosciuto e posizionato vicini tra loro tre progetti scritti con le stesse sequenze di parole, probabilmente dalla mano del medesimo progettista e presentati da tre richiedenti diversi.

Nonostante le forti evidenze e le implicazioni sul piano della valutazione dei progetti che tale risultato comporta, il limite dell’esperimento sta nell’impossibilità di intervenire sul grafico generato dal software al fine di evitare le sovrapposizioni, che ne impediscono di fatto una lettura più approfondita.

Un risultato analogo, ma di più agevole interpretazione, è stato ottenuto utilizzando il pacchetto Stylo in ambiente R. Le funzionalità di tale pacchetto consentono di effettuare un’analisi stilometrica, traducendo informazioni testuali in dati numerici su cui poter effettuare elaborazioni statistiche al fine dell’attribuzione di una 'paternità' dei testi.

Tab. 3 Quadro sinottico delle componenti descrittive di tre progetti in posizione limitrofa nel piano cartesiano.

Progetto (…)9568	**Progetto (…)9582**	**Progetto (…)9574**
Descrizione della motivazione		

La visione del Piano di sviluppo Industria 4.0 cambia completamente la pianificazione del ciclo di vita del prodotto/servizio e il modo in cui l'azienda lo gestisce, lo segue e lo controlla e pone le proprie basi nella connessione e nell'integrazione, permettendo di avere una visione di insieme di tutte le fasi del ciclo di vita del prodotto, anche quando lo stesso esce dall'azienda di produzione ed entra nel circuito della distribuzione per entrare nelle fabbriche dei clienti. Ma grazie alle digitalizzazione (…)	La visione del Piano di sviluppo Industria 4.0 cambia completamente la pianificazione del ciclo di vita del prodotto/servizio e il modo in cui l'azienda lo gestisce, lo segue e lo controlla e pone le proprie basi nella connessione e nell'integrazione, permettendo di avere una visione di insieme di tutte le fasi del ciclo di vita del prodotto, anche quando lo stesso esce dall'azienda di produzione ed entra nel circuito della distribuzione per entrare nelle fabbriche dei clienti. Ma grazie alle digitalizzazione (…)	La visione del Piano di sviluppo Industria 4.0 cambia completamente la pianificazione del ciclo di vita del prodotto/servizio e il modo in cui l'azienda lo gestisce, lo segue e lo controlla e pone le proprie basi nella connessione e nell'integrazione, permettendo di avere una visione di insieme di tutte le fasi del ciclo di vita del prodotto, anche quando lo stesso esce dall'azienda di produzione ed entra nel circuito della distribuzione per entrare nelle fabbriche dei clienti. Ma grazie alle digitalizzazione (…)
Descrizione degli obiettivi formativi		
Il presente progetto si propone di fornire ai partecipanti le competenze necessarie per rinnovare le strategie, le tecniche e gli strumenti di comunicazione e marketing in ottica 4.0, allo scopo di massimizzare l'impatto commerciale aziendale. Attraverso la realizzazione di 3 interventi + 1 di accompagnamento, i partecipanti saranno in grado di raggiungere i seguenti obiettivi formativi: (…)	Il presente progetto si propone di fornire ai partecipanti le competenze necessarie per rinnovare le strategie, le tecniche e gli strumenti di comunicazione e marketing in ottica 4.0, allo scopo di massimizzare l'impatto commerciale aziendale. Attraverso la realizzazione di 7 interventi (6 formativi di cui uno esperienziale + 1 di accompagnamento), i partecipanti saranno in grado di raggiungere i seguenti obiettivi formativi: (…)	Il presente progetto si propone di fornire ai partecipanti le competenze necessarie per rinnovare le strategie, le tecniche e gli strumenti di comunicazione e marketing in ottica 4.0, allo scopo di massimizzare l'impatto commerciale aziendale. Attraverso la realizzazione di 4 interventi (3 formativi + 1 di accompagnamento), i partecipanti saranno in grado di raggiungere i seguenti obiettivi formativi: (…)
Descrizione delle azioni complementari		

Non previste: il progetto è scaturito da incontri preliminari tra il management dell'azienda e il gruppo di lavoro (progettisti, coordinatori ed equipe di docenti e consulenti) oltre a sessioni in cui i collaboratori si sono incontrati per trasferirsi reciprocamente esigenze, fabbisogni e aspettative rispetto al percorso da avviare. (…)	Non previste: il progetto è scaturito da incontri preliminari tra il management dell'azienda e il gruppo di lavoro (progettisti, coordinatori ed equipe di docenti e consulenti) oltre a sessioni in cui i collaboratori si sono incontrati per trasferirsi reciprocamente esigenze, fabbisogni e aspettative rispetto al percorso da avviare. (…)	Non previste: il progetto è scaturito da incontri preliminari tra il management dell'azienda e il gruppo di lavoro (progettisti, coordinatori ed equipe di docenti e consulenti) oltre a sessioni in cui i collaboratori si sono incontrati per trasferirsi reciprocamente esigenze, fabbisogni e aspettative rispetto al percorso da avviare. (…)

Stylo classifica i testi in base alla distanza intertestuale tra i documenti. L'output dell'analisi è costituito da una matrice quadrata di distanze che può essere elaborata con un foglio di lavoro in cui ogni singolo testo viene confrontato con se stesso e con tutti gli altri. Nella diagonale della matrice, dove ciascun testo è confrontato con se stesso, il valore risulta zero (perfetta corrispondenza); nel resto della matrice viene calcolato quanto ciascun progetto è più o meno distante da tutti gli altri con una misura di correlazione statistica.

Più basso è il valore, più i progetti sono simili tra loro. La Fig. 15 rappresenta una porzione del foglio di calcolo che, per il nostro corpus, contiene 474 righe ed altrettante colonne. Nella figura sono state formattate (con colore rosa e testo rosso) le caselle contenenti una misura di correlazione appartenente al primo percentile (da 0 a 0,769681302) del totale dei valori ottenuti (che vanno da 0 a 1,889574968). Tra le celle formattate si possono distinguere i progetti già evidenziati nella Tab. 3.

La soglia del primo percentile è stata scelta arbitrariamente, nell'assunto di fondo per cui i progetti così selezionati siano quelli più simili tra loro.

I progetti con la misura di distanza più bassa hanno matematiche possibilità di non essere del tutto originali. Mentre è certo che quelli con i valori più alti sono del tutto originali e privi di copia. Un'ipotesi per implementare questo strumento ai fini valutativi può essere quella di contare, per ciascun progetto, letto per riga o per colonna, il numero di caselle con valore di distanza inferiore alla soglia fissata: più alto è il valore ottenuto con il conteggio, più è alta la probabilità che quel progetto sia meno originale degli altri.

In conclusione, i risultati proposti non intendono, ribadiamo, attribuire un giudizio alla qualità dei progetti. Anche applicando i metodi e le

tecniche proposte in questo lavoro, la valutazione rimarrebbe infatti in capo agli esperti, che potranno sondare le singole progettualità mettendo in campo le proprie specifiche competenze, ma con il vantaggio, dato dall'analisi automatica del testo, di disporre di più tempo per svolgere questo tipo di attività.

Fig. 15 Porzione della matrice quadrata di rappresentazione delle distanze tra i progetti del corpus creata con il pacchetto Stylo.

	A	PU	PV	PW	PX	PY	PZ
1		9515	9519	9568	9574	9582	9584
11	9450	0,854664172	0,881700155	0,915402946	0,919977752	0,890732658	0,813334106
12	7148	0,860877718	0,829410046	0,845574007	0,883482768	0,897397371	0,921153394
13	9738	0,85855075	0,846730996	0,873256056	0,87410438	0,86012867	0,726788227
14	3624	0,936909748	0,875491139	0,898966897	0,88911551	0,93251337	0,782638066
15	8264	0,896299791	0,866229574	0,949456078	0,944545932	0,960136056	0,862868187
16	9421	0,95958067	0,961291722	0,764459628	0,755873962	0,767286512	0,730066332
17	9751	1,027693367	0,901510675	0,919250889	0,932793799	0,890168517	0,840058466
18	5173	1,034134747	0,945258247	0,814492275	0,79610478	0,826656773	0,820262019
19	8590	1,020383108	0,860504612	0,663553414	0,650644776	0,669335413	0,80895723
20	7667	1,010989408	0,896279144	0,907585435	0,898867316	0,892792567	0,842647358
21	8287	0,943798535	0,907182299	0,852526357	0,830854187	0,838535754	0,814444097
22	9498	0,973172627	0,791650125	0,927025667	0,909341389	0,927549728	0,729950629
23	9568	0,998859427	0,954145612	0	0,140766812	0,116920541	0,835349064
24	9574	0,989547629	0,934926548	0,140766812	0	0,148811257	0,80252204
25	5769	1,002529281	0,9446268	0,075409309	0,142287314	0,085628377	0,819070051
26	8681	1,002529281	0,9446268	0,075409309	0,142287314	0,085628377	0,819070051
27	8845	1,002529281	0,9446268	0,075409309	0,142287314	0,085628377	0,819070051
28	8986	0,991928733	0,847415315	0,77565685	0,797332152	0,762793054	0,794050725
29	9582	0,991982593	0,923163851	0,116920541	0,148811257	0	0,836443322
30	1181	1,028076459	0,842429283	0,928642907	0,90961378	0,920877861	0,834975525
31	5323	0,953563884	0,945069157	0,970072874	0,967396358	0,981721209	0,876114875
32	5422	1,015278069	0,936202085	0,860549497	0,849283357	0,833831657	0,787152065

Conclusioni

Nella realizzazione dello studio presentato nel capitolo e nel vedere gli strumenti all'opera sono due gli aspetti che più hanno dato fiato e fiducia al lavoro.

Il primo è una sorta di 'effetto sorpresa'. C'è stato nel momento in cui, assieme ai colleghi di Regione Veneto, abbiamo visto ricostruire mondi lessicali contenuti in migliaia di progetti per mezzo di un software basato su calcoli matematici e statistici. E questi mondi lessicali si sono rivelati in linea con le parole chiave e le linee di intervento del bando preso in

esame. Abbiamo lavorato con progetti della formazione professionale, ma sarebbero potuti appartenere a qualsiasi altra materia.

Il secondo è stato quello della 'velocità'. Trentotto secondi: è il tempo con cui Iramuteq ha elaborato un corpus di mezzo milione di parole e ne ha tirato fuori parole chiave e mondi semantici. Poche ore ulteriori di lavoro ci hanno permesso di condurre analisi più approfondite sul corpus dei progetti presentati. La velocità non è un *must*, perché poi tutto può dipendere dalla domanda di ricerca e di valutazione che viene posta. Ma in un mondo che va veloce, questo aspetto, se non altro, non può non affascinare.

In un'epoca in cui si moltiplicano i bandi e il numero di progetti da valutare ed esaminare, l'idea di poter trovare un'applicazione delle tecniche di text mining nell'ambito dell'attività amministrativa, qual è ad esempio un'istruttoria concorsuale, è immediata quanto sfidante. La sfida è quella di impiegare questo approccio misto quali-quantitativo in ausilio a procedure che prevedono la lettura e la valutazione qualitativa di centinaia di testi, supportando ed efficientando il sistema, riducendo il più possibile il sovraccarico. Il vantaggio auspicato risiede sia nell'opportunità di ridurre i *bias* legati alla soggettività del valutatore, sia nel superamento delle limitazioni insite nelle valutazioni di tipo esclusivamente qualitativo, consentendo di investigare, se non tutti, almeno alcuni aspetti salienti di ampie quantità di testi progettuali in tempi brevi; di esplorare sentieri nascosti; di fornire più robustezza alle valutazioni (Sbalchiero et al. 2016).

È chiaro che gli approcci quali-quantitativi, per quanto condotti con sofisticate applicazioni, non sono in grado di sostituire totalmente i metodi tradizionali di valutazione, e che l'eventuale implementazione e applicazione delle tecniche descritte in questo capitolo nei processi di valutazione richiederebbe sicuramente un *training* sperimentale fatto di prove e tentativi. Tuttavia, vale la pena non lasciare nessuna strada intentata, e il presente capitolo rappresenta un primo tentativo di applicazione. Come diceva Edison, l'innovazione è fatta per l'1% da *inspiration*. Il resto è tutto *perspiration*, fatica, lavoro, ripetizione di tentativi, affinamento di tecniche, consolidamento di procedure, apertura di nuove strade.

Valutare e migliorare la qualità delle decisioni. L'analisi delle istruttorie del Settore Sismica della Regione Toscana

Stefano Acciaioli[1]

Soccorso Istruttorio, Istruttorie, Supporto alle decisioni, Settore Sismica, DPR 380/01, Costruzioni, Uniformità, Discrezionalità, Regione Toscana.

Introduzione

La pubblica amministrazione in generale, e gli uffici che svolgono attività di controllo in particolare, devono tendere all'applicazione dei principi cardine dell'imparzialità, di parità di trattamento, di equilibrio, di uniformità e di lealtà che regolano il procedimento amministrativo.

Oggetto del presente capitolo è la proposta di uno strumento innovativo per la valutazione del procedimento amministrativo del soccorso istruttorio relativo al controllo dei progetti strutturali delle costruzioni edilizie depositati presso il Settore Sismica della Regione Toscana, nell'ambito del quale l'azione amministrativa, spesso e per diversi fattori concorrenti, tende ad allontanarsi da tali principi, determinando di fatto un aggravamento del procedimento. La fase istruttoria del procedimen-

[1] Ingegnere civile strutturista, si è occupato in particolare di progettazione di strutture in carpenteria metallica e di opere da ponte, di prevenzione del rischio sismico, prima come libero professionista e dal 2008 ad oggi per il Settore Sismica di Regione Toscana. In questo ambito ha partecipato alle attività di protezione civile di gestione dell'emergenza post-sisma come valutatore del danno e agibilità degli edifici, nei crateri dei principali eventi sismici degli ultimi anni.

to di controllo del rispetto della normativa tecnica nelle costruzioni, nel corso della quale la pubblica amministrazione può attivare il cosiddetto soccorso istruttorio – ossia la richiesta di integrazione sui contenuti dei progetti – è una fase essenziale del procedimento, che anticipa la decisione amministrativa formale che produce effetti verso l'esterno.

L'azione della macchina amministrativa è spesso rappresentata come poco incline allo snellimento e al miglioramento dei processi, all'applicazione delle nuove tecnologie ed in generale al cambiamento per incrementare funzionalità ed efficienza. Una delle più recenti innovazioni che sono state attuate dalla pubblica amministrazione è il passaggio dall'utilizzo dei documenti cartacei alla versione informatizzata. La dematerializzazione dei documenti, nella maggior parte dei casi, determina un effetto indotto molto sottovalutato: la disponibilità di una grande quantità di dati, più o meno organizzati, per essere analizzati

È a partire da queste considerazioni che nasce l'idea (e la necessità) di introdurre un'analisi testuale automatica dei contenuti delle richieste d'integrazione prodotte dai tecnici istruttori del Settore Sismica della Regione Toscana, volta ad indagare la fase istruttoria del procedimento di controllo dei progetti strutturali delle costruzioni edilizie che sono depositati presso il Settore Sismica e assoggettati per legge all'attività di controllo. L'analisi automatica delle istruttorie ha il duplice scopo di fornire uno strumento di valutazione della fase istruttoria e di migliorare le decisioni in un ambito importante e delicato come quello della garanzia della pubblica incolumità, della salvaguardia della vita umana e della sicurezza delle costruzioni. Si tratta quindi di individuare se vi siano meccanismi distorsivi del buon andamento del procedimento amministrativo, nello specifico legati alla non uniformità di trattamento ed in generale ad aspetti relativi all'uso della discrezionalità nell'azione amministrativa.

Il capitolo è strutturato in tre parti principali. Dopo aver chiarito il contesto dell'analisi testuale (primo e secondo paragrafo), a partire dal terzo paragrafo sarà presentata l'analisi del contenuto applicata alle richieste di integrazione ai progetti edilizi formulate dai tecnici istruttori del Settore Sismica, che costituiscono una fonte di informazioni fino ad oggi mai analizzata in modo strutturato.

Nello specifico, l'analisi delle istruttorie, attuata con l'utilizzo del software Iramuteq, ha visto l'applicazione di diverse tecniche di analisi testuale (*topic detection* con il metodo Reinert per la classificazione delle istruttorie in *topics*; analisi delle corrispondenze applicata ai *cluster*, per ricercare le distanze relative tra le classi; determinazione di parole chiave

e della loro salienza nelle diverse classi), allo scopo di cogliere comportamenti non uniformi nelle istruttorie, evidenziando le distanze testuali sia negli aspetti qualitativi che quantitativi e dal punto di vista semantico, mettendo in luce le peculiarità che determinano le non uniformità delle istruttorie trattate ed i motivi delle distanze in funzione delle variabili definite. Il lavoro è stato sviluppato in due successive fasi:

1) in una prima fase è stato sottoposto alla popolazione dei tecnici istruttori regionali (52 in totale) un progetto 'test', unico per tutti e scelto tra quelli depositati presso il Settore Sismica della Regione Toscana, con la richiesta di condurre una consueta attività di controllo individuale e di produrre l'eventuale richiesta di integrazione; ciò allo scopo di confrontare tali istruttorie e di misurare il grado di difformità delle stesse, sia nei modi di conduzione dell'attività di controllo, sia nel merito dei contenuti delle richieste d'integrazione;
2) in una seconda fase è stata analizzata l'intera banca dati delle richieste di integrazione formulate dai tecnici istruttori ed emesse nell'ordinario svolgimento delle attività di controllo (6092 richieste di integrazione in fase istruttoria e 67 istruttori in totale) sui progetti depositati presso il medesimo Settore tra il 2015 e il 2020, con lo scopo, oltre che di valutare il grado di difformità delle richieste per istruttore, di misurare la variabilità territoriale e tematica dei contenuti, ricercando quali aspetti contribuiscano maggiormente alla discrezionalità dell'azione amministrativa.

La discrezionalità del procedimento amministrativo

L'uso della discrezionalità nel procedimento amministrativo può confliggere con il principio dell'imparzialità a cui la pubblica amministrazione dovrebbe ispirarsi sia verso i soggetti esterni, sia nei confronti dei propri dipendenti. Un uso distorto della discrezionalità amministrativa e tecnica interna, infatti, può non garantire un corretto supporto alle decisioni esterne che la p.a. deve assumere con l'obiettivo di perseguire il bene pubblico.

Le analisi delle distanze tra i contributi istruttori a supporto delle decisioni che formano l'esito del procedimento amministrativo che saranno presentate nel corso del capitolo non hanno lo scopo di valutare nel merito la correttezza o inesattezza delle richieste di supplemento istruttorio, ma guardano alla garanzia dell'equo trattamento del destinatario della decisione, alla luce dell'applicazione di una normativa tecnica. L'esercizio

svolto dovrebbe favorire, come effetto indotto, anche un miglioramento della correttezza e del coordinamento nel merito delle istruttorie condotte, favorendo il confronto tra i tecnici istruttori ed individuando i temi più critici e soggetti alla discrezionalità tecnica, sui quali attivare eventuali meccanismi correttivi.

Lo scopo del lavoro è pertanto duplice: da un lato, come obiettivo specifico, favorire il rispetto dei principi del procedimento amministrativo ed il miglioramento dell'azione amministrativa; dall'altro, come obiettivo generale, favorire l'aumento della sicurezza delle costruzioni (*safety*) e della garanzia della privata e pubblica incolumità. Tale incolumità è affidata al rispetto delle condizioni di esercizio ordinario delle strutture, ma anche alla difesa ed al contenimento degli effetti prodotti da eventi eccezionali quali catastrofi e calamità naturali – e, quindi, degli obiettivi definiti dal *Goal* 11 dell'Agenda 2030[2] e dall'Agenda di Sendai[3].

Se l'obiettivo specifico del miglioramento del procedimento amministrativo, della garanzia di parità ed uniformità di trattamento nell'azione amministrativa, dell'aumento dell'efficacia e dell'efficienza del controllo operato dal Settore Sismica comporta benefici diretti per i soggetti committenti le opere (privati e pubblici), oltre che per i soggetti tecnici (progettisti, direttori lavori, collaudatori) incaricati dalle committenze e che si rapportano direttamente con il Settore nelle procedure di deposito dei progetti, e produce effetti più facilmente e direttamente misurabili, altrettanto non si può dire della garanzia della sicurezza delle costruzioni. Questa, infatti, coinvolge effetti di più lungo periodo, difficilmente misurabili direttamente, oltreché poco auspicabili nella misura stessa, in quanto da riferirsi alla contabilizzazione delle perdite occorse in termini economici e di vite umane, a seguito dell'evento che determina il fallimento della costruzione.

Il Settore Sismica di Regione Toscana e il procedimento amministrativo

Storicamente, si registra a livello regionale una forte variabilità della risposta del Settore in fase istruttoria, con particolare riferimento ai con-

[2] Agenda 2030 per lo sviluppo sostenibile è un programma d'azione per le persone, il pianeta e la prosperità sottoscritto nel settembre 2015 dai governi dei 193 Paesi membri dell'ONU.

[3] *Sendai Framework for Disaster Risk Reduction* 2015-2030. Il Quadro di riferimento di Sendai ha inaugurato il passaggio dalla sola 'gestione delle catastrofi', all'attuale interpretazione estesa di 'gestione del rischio di catastrofi'.

tenuti delle richieste di integrazione, dovuta alle diverse competenze tecniche degli istruttori, alle realtà territoriali dei vari uffici, alle prassi consolidate nei vari presidi ed alla gestione individuale dell'istruttoria. Ad ulteriore complicazione, si sovrappone un apparato normativo tecnico di riferimento molto complesso e dettagliato, che si presta spesso ad interpretazioni e letture contrastanti, per non dire talvolta contraddittorie.

Tutto ciò ha esposto da sempre il Settore Sismica della Regione Toscana a critiche da parte dei tecnici professionisti esterni e degli ordini professionali che li rappresentano, dei committenti e di altri soggetti che hanno interesse diretto o indiretto nel procedimento.

Il Settore Sismica della Regione Toscana si occupa dei controlli e della vigilanza sulle costruzioni, in riferimento al rispetto delle Norme Tecniche sulle Costruzioni, secondo quanto disposto dagli artt. 83 e 93 del Decreto del Presidente della Repubblica 6 giugno 2001, n. 380 (D.P.R. 380/01), ai fini di garantire, sul territorio di competenza, i livelli di sicurezza non derogabili stabiliti a livello nazionale, a garanzia della privata e pubblica incolumità.

Sono oggetto di controllo i progetti strutturali delle costruzioni, sviluppati da tecnici professionisti, su incarico di committenti pubblici e privati, costituiti da elaborati aventi natura prettamente tecnica, prodotti in forma di grafici esecutivi e di relazioni. Lo scopo dei progetti è quello di relazionare sul dimensionamento e sulla verifica degli elementi costituenti la struttura, in ragione delle azioni sollecitanti, nel rispetto delle Norme Tecniche per le Costruzioni attualmente vigenti (Decreto Ministeriale 17 gennaio 2018). Il deposito dei progetti avviene per mezzo del portale web PO.R.TO.S.[4] (Regione Toscana 2020).

Il Settore Sismica, unico su tutta la Regione Toscana, è articolato in uffici con competenza territoriale su base provinciale. Il controllo è effettuato da personale tecnico (geometri, architetti, ingegneri, geologi), che svolge le proprie istruttorie in forma individuale.

Il procedimento amministrativo, nel quale si sostanzia l'attività di controllo svolta dal Settore, comprende diverse fasi: l'avvio del procedimento, la fase istruttoria, la fase decisoria/costitutiva ed eventualmente la fase integrativa dell'efficacia.

La fase istruttoria, che, come abbiamo già detto, costituisce oggetto di indagine nel presente lavoro, è spesso sottovalutata, ma essenziale per la

[4] PO.R.TO.S è l'acronimo di POrtale della Regione TOscana per la Sismica ed è il portale web attraverso il quale vengono gestiti i procedimenti relativi ai progetti delle strutture in zona sismica ai sensi del DPR 380/01.

corretta formazione della decisione finale. Essa è infatti dedicata all'accertamento dei fatti e dei presupposti ed all'acquisizione delle notizie ed informazioni necessarie per poter pervenire ad una decisione corretta, e la sua attivazione avviene tramite l'emissione, da parte dei tecnici del Settore, delle richieste di integrazione.

In generale, qualunque progetto che interessi le parti strutturali delle costruzioni (edifici nuovi o esistenti, di qualunque tipologia costruttiva, con qualsiasi destinazione/funzione svolta, ecc.) dev'essere depositato presso il Settore, e può essere assoggettato ad un controllo di merito che verifica il rispetto delle norme tecniche per le costruzioni emanate dal Ministero dei lavori pubblici, periodicamente aggiornate sulla base delle nuove conoscenze e dell'evoluzione tecnico-scientifica delle materie dell'ingegneria civile in generale e delle costruzioni in zona sismica in particolare.

I progetti sono assoggettati a procedimenti diversi (deposito o autorizzazione)[5], con livelli differenziati di controllo, in funzione del grado di sismicità del luogo dove saranno realizzati ed in funzione dell'importanza e della tipologia di opera.

In generale, interventi che ricadono in zone a più alta sismicità, di particolare rilevanza (ad es. scuole), o di particolare complessità strutturale (ad es. ponti), sono tutti sottoposti a controllo. Interventi che non presentano le precedenti caratteristiche sono istruiti solo se sorteggiati sulla base di un campione estratto a cadenza mensile. Il tenore del controllo verte sugli stessi principi in entrambi i casi, pertanto, nel merito, l'attività istruttoria non si distingue per principio in funzione del procedimento, dovendo in generale controllare la conformità alla norma tecnica per le costruzioni, unica per tutte le tipologie di intervento e dovendo, in linea di principio, garantire livelli minimi di sicurezza uniformi su tutto il territorio nazionale.

Se l'esito dell'attività di controllo è positivo, essa dà luogo ad una fase decisoria che si concretizza nel rilascio di un atto autorizzativo per l'esecuzione dei lavori o con un'espressione di conformità. Viceversa, l'attività istruttoria si concretizza nell'emissione di una o più richieste di integrazioni da parte del tecnico istruttore, a cui viene assegnato il controllo della pratica: il procedimento qui richiamato è di fatto l'attivazione del

[5] Il procedimento di autorizzazione consente l'esecuzione dei lavori solo dopo l'emissione da parte del settore dell'atto autorizzativo, il procedimento di deposito viceversa dà titolo all'esecuzione dei lavori dal momento della protocollazione del progetto.

principio del soccorso istruttorio[6] previsto dall'art. 6, comma 1, lett. b della Legge 7 agosto 1990, n. 241 (L. 241/90).

L'attività istruttoria è quindi centrale e sostanziale per la corretta adozione del provvedimento finale, in quanto determina l'acquisizione di conoscenza della realtà e di fatti necessari alla corretta formazione della decisione.

Gli aspetti più critici che riguardano questa fase del procedimento sono quelli legati ai profili di responsabilità e alla delega di responsabilità, alla forte variabilità dei comportamenti e delle richieste di integrazioni prodotte, e alla risposta del Settore verso l'esterno.

In questa fase, la correttezza dell'attività svolta dagli istruttori concorre alla correttezza delle attività svolte da altri soggetti che ricoprono altri ruoli. L'istruttore svolge un'attività cruciale, di cui però non è direttamente responsabile verso l'esterno, essendo questa coperta da profili di responsabilità di altri soggetti sovraordinati gerarchicamente e protagonisti della fase decisoria, quali il responsabile del procedimento o il dirigente. Circostanza che può determinare una sottovalutazione dell'importanza dell'attività da parte degli stessi istruttori. Dall'altra parte, si nota spesso come i soggetti che hanno la responsabilità dell'assunzione della decisione finale non operino alcuna valutazione sull'istruttoria loro rimessa, ma piuttosto la facciano avanzare o meno sulla base di un criterio fiduciario nei riguardi dell'istruttore. Infine, l'attività istruttoria è un'attività tipicamente individuale, o comunque circoscritta a pochi soggetti interagenti, elemento che la rende esposta a forte variabilità dei contenuti e delle modalità con cui viene condotta. A ciò si aggiungono ulteriori condizioni quali variabili ambientali, peculiarità del territorio, prassi consolidate, competenze e così via, senza voler arrivare a citare anche le possibili distorsioni del procedimento stesso. Sono queste tutte variabili che contribuiscono alla non uniformità della risposta nel procedimento amministrativo, che contrasta con il buon andamento del procedimento stesso, con particolare riferimento alla L. 241/90, che all'art. 1 individua, tra i principi a cui deve conformarsi l'azione amministrativa, quelli di efficacia, di imparzialità e di non aggravamento del procedimento, e nei disposti dell'art. 6 inquadra l'attività dei tecnici istruttori, nell'ambito della quale il già citato soccorso istruttorio sottende il dovere di lealtà e di imparzialità, che impone alla pubblica amministrazione di trattare situazioni analoghe allo stesso modo, oltre ai principi del buon andamento e della trasparenza.

[6] Legge 7 agosto 1990, n. 241, art. 6, c. 1, lett. b "Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativo".

Sarebbe tuttavia riduttivo considerare il problema solo dal punto di vista del mero rispetto della norma giuridica che regola l'attività della pubblica amministrazione o del rispetto della regolazione tecnica, e non anche come parte del processo di implementazione della public policy, essendo la risposta dell'ufficio rivolta verso l'esterno.

È evidente, in primo luogo, che il tema in oggetto concerne anche l'efficienza e l'efficacia dell'azione amministrativa nel perseguimento dei compiti attribuiti di garanzia e salvaguardia della privata e pubblica incolumità e, in definitiva, del bene pubblico.

In secondo luogo, pur essendo la fase istruttoria volta alla formazione di una decisione che si costituisce internamente alla p.a., essa presenta una forte interazione con l'esterno, andando ad incidere anche su come viene percepita la p.a. da chi è coinvolto nel procedimento ed in generale dalla collettività.

Il metodo e i dati

Per la realizzazione delle analisi presentate nei paragrafi successivi sono stati utilizzati due corpora testuali, ognuno per ciascuna fase della ricerca. In linea generale, i testi analizzati non hanno una forma prestabilita. La richiesta d'integrazione è infatti definita autonomamente dal tecnico istruttore, che la redige secondo le necessità relative alla specifica attività di controllo sul singolo progetto. Tuttavia, i testi presentano una lunghezza tendenzialmente breve nella maggior parte dei casi.

Per la fase 1 dedicata al progetto test, il corpus è stato costruito con le richieste di integrazione predisposte dai tecnici istruttori al di fuori del portale PO.R.TO.S., a seguito della conduzione di un'istruttoria in relazione ad un progetto, scelto *ad hoc* tra i vari depositati presso il Settore Sismica unico per tutti gli istruttori, in quanto presentava alcune criticità relativamente all'inquadramento degli interventi nell'ambito del disposto normativo, prestandosi ad interpretazioni diverse. Complessivamente, sono stati coinvolti nel progetto 'test' 52 tecnici istruttori del Settore Sismica di Regione Toscana, e di questi, 28 hanno prodotto una richiesta di integrazione. I rimanenti sono da annoverare tra chi non ha risposto per scelta e chi ha considerato il progetto sottoposto al controllo coerente con il dettato della normativa tecnica per le costruzioni, non ritenendo quindi di dover richiedere alcuna integrazione.

Per la fase 2, dedicata all'analisi delle richieste istruttorie prodotte tra marzo 2015 e giugno 2020, è stata utilizzata la globalità delle richieste di integrazione (6.092 in totale) prodotte dai 67 tecnici istruttori sui vari pre-

sidi che hanno operato le attività di controllo sulle pratiche dei progetti nell'arco di tempo preso in considerazione. Le richieste sono state estratte dal database di PO.R.TO.S.

Per ogni richiesta, sono state inoltre individuate alcune variabili e relative modalità di riferimento, utili per classificare entrambi i corpora ai fini delle elaborazioni successive, che, alla luce delle considerazioni fin qui operate, ricomprendono: il tecnico istruttore associato, il titolo di studio del tecnico istruttore, la tipologia di pratica (autorizzazione o deposito), il numero di richiesta di integrazione (nel caso di più richieste di integrazione), la tipologia di intervento (nuova costruzione o le categorie degli interventi sulle costruzioni esistenti, quali locale, miglioramento o adeguamento), la provincia di esecuzione dell'intervento.

Considerati gli obiettivi dello studio, si è scelto di non associare direttamente il tecnico istruttore all'istruttoria, per evitare possibili condizionamenti o innescare tensioni non volute nella fase di un eventuale successivo riallineamento dell'attività istruttoria. L'associazione tra l'istruttore ed il testo della richiesta di integrazioni è stata pertanto mantenuta, in quanto necessaria anche per gli scopi dell'analisi, ma anonimizzata, attribuendo in modo del tutto casuale un indice numerico ad ogni singolo tecnico istruttore, e ottenendo così un'individuazione univoca, ma non esplicita.

L'analisi testuale dei corpora è stata preceduta da una prima operazione di filtraggio tesa ad eliminare dati spuri ed a correggere errori di battitura o grammaticali ripetuti. Sono quindi stati riconosciuti ed associati alla stessa forma testuale abbreviazioni ed acronimi, spesso per riferimenti a testi di legge, e sono state definite le cosiddette *multi-word expressions*.

Vista la notevole quantità di dati oggetto di analisi, sia per la procedura di anonimizzazione, sia per il lavoro di *pre-processing* in generale, si è proceduto alle sostituzioni con un procedimento automatizzato implementato con la scrittura di poche righe di codice di linguaggio VBA[7], definendo liste di sostituzione ed applicandole ai corpora. Infine, questi ultimi sono stati ulteriormente trattati direttamente all'interno del software Iramuteq con il procedimento di lemmatizzazione per ricondurre la forma flessa della parola contenuta nel testo alla forma canonica.

[7] VBA acronimo di *Visual Basic for Application*, è un linguaggio di programmazione ad alto livello derivato dal *Visual Basic* ed utilizzato per controllare ed automatizzare le operazioni caratteristiche di un determinato applicativo.

Istruttorie Progetto test

La scelta di sottoporre un unico progetto alla popolazione degli istruttori e di analizzare successivamente le loro richieste di integrazione prodotte ha lo scopo di individuare i temi sui quali le richieste si distanzino maggiormente le une dalle altre, in funzione del comportamento del singolo istruttore. Il progetto è stato scelto, tra tutti quelli disponibili depositati presso il Settore Sismica sui quali non fosse stato attivato un procedimento di controllo, in quanto caratterizzato da un inquadramento normativo al limite[8], intendendo con ciò stimolare la risposta degli istruttori ed avere anche una maggior ampiezza dei temi trattati dalle possibili richieste, in modo tale da testare la risposta istruttoria dei vari tecnici. A questi ultimi è stato chiesto di condurre un'istruttoria usuale, ma al di fuori del normale procedimento amministrativo.

L'analisi con il metodo Reinert del primo corpus ha individuato la presenza nello stesso di 4 *cluster* semantici (Fig. 1), che possono essere così descritti:

- la classe 1 riguarda la realizzazione di un ampliamento in muratura (definiamo la classe "prescrizioni tecniche esecutive dell'intervento");
- la classe 2 fa riferimento al cambiamento di destinazione d'uso del sottotetto (definiamo la classe "inquadramento dell'intervento");
- la classe 3 riguarda i contenuti più generali di ambito normativo e regolatorio e di impostazione dell'intervento (definiamo la classe "ambito normativo");
- la classe 4 attiene alla richiesta di dettagli di migliori specifiche di esecuzione contenute negli elaborati (definiamo la classe "dettagli esecutivi").

La sovrapposizione o, di converso, la distanza tra le 4 aree semantiche riconosciute nel *corpus* è evidenziata dal posizionamento delle classi su un piano fattoriale relativo ad un'analisi delle corrispondenze applicata ai *cluster* (Fig. 2, parte sinistra). La Fig. 2 parte destra mostra la distribu-

[8] Il progetto scelto riguarda un intervento su una costruzione esistente. La normativa tecnica per le costruzioni definisce tre possibili inquadramenti che definiscono anche gli obblighi conseguenti in termini di valutazione della sicurezza da verificare e produrre, ma il limite tra le tre tipologie non è un confine netto ed è oggetto di interpretazioni diversificate della norma stessa.

zione, sul medesimo piano fattoriale, della variabile relativa al tecnico istruttore, evidenziando, in relazione ai 4 *cluster*, la sovrapposizione o la distanza tra i tecnici istruttori.

Fig. 1 *Topics* delle istruttorie del Progetto test.

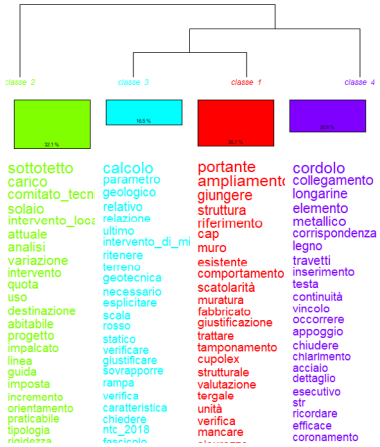
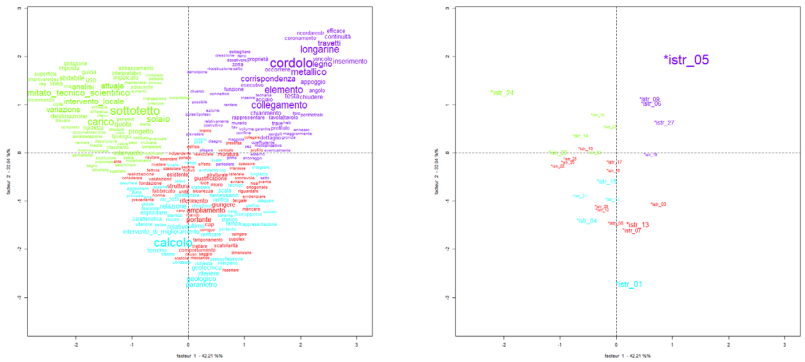


Fig. 2 Distribuzione dei *topics* trattati dagli istruttori su piano fattoriale.



Analisi dei risultati del progetto test

Partendo dal presupposto che il progetto scelto come test non fosse correttamente allineato ai dettami della norma tecnica, l'aspetto che si evidenzia in questa sede è che il focus sul quale si sarebbero dovute concentrare tutte le istruttorie è quello della conformità alle prescrizioni normative in generale, siano esse di inquadramento degli interventi o di rispetto di prescrizioni tecniche esecutive.

Tali argomenti sono effettivamente presenti nelle classi 1, 2 e 3. Dalla Fig. 2 si nota inoltre come, per le classi 1 e 3, vi sia una parziale sovrapposizione sul piano fattoriale. Viceversa, la classe 2 si presenta nettamente separata dalle altre due. L'interpretazione che si può dare è legata al fatto che tutte e tre le classi colgono il non allineamento alla normativa, ma per motivi differenti: la classe 3, di livello più generale, richiama il calcolo e la normativa tecnica per le costruzioni; la classe 1 concentra il problema sull'intervento di realizzazione di ampliamento in muratura, che è carente rispetto alle prescrizioni tecniche esecutive di norma; la classe 2, infine, riguarda l'intervento di cambio di destinazione d'uso al sottotetto, non allineato alla norma per l'inquadramento dello stesso. La sovrapposizione tra le classi 1 e 3 evidenzia, in particolare, come il problema che caratterizza le richieste della classe 1 sia maggiormente supportato dal disposto normativo. La classe 2 sembra invece trovare supporto, per la contestazione dell'inquadramento dell'intervento, non tanto sulla norma tecnica per le costruzioni, ma piuttosto su linee guida emanate dal Settore, in quanto si trova nella stessa classe il riferimento al Comitato Tecnico Scientifico[9].

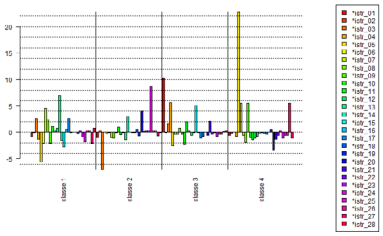
Molto diverso è il caso della classe 4, che si concentra sulle richieste di dettaglio dei particolari esecutivi, non cogliendo le carenze di inquadramento generale del progetto rispetto alla norma.

Sia dal grafico in Fig. 2 parte destra, sia dalla successiva Fig. 3, che rappresenta il livello di associazione tra i *cluster* e i singoli istruttori determinato dal valore del χ^2, è evidente come sia in particolare l'istruttore n. 5 a concentrare la propria richiesta istruttoria sull'area delle richieste di "dettaglio" dei particolari esecutivi.

Entrambe le figure mostrano inoltre la presenza di altri istruttori 'poco baricentrici', ossia quei tecnici che utilizzano forme testuali chiaramente discoste rispetto alla normale distribuzione dei contenuti istruttori, in particolare gli istruttori n. 24 e n. 1. In questo caso, è interessante notare come, sulla base dell'interpretazione dei 4 *cluster* individuati, entrambi gli istruttori si concentrino sul non allineamento ai dettami della norma tecnica, ma su presupposti differenti: il n. 24 richiama l'intervento di cambio di destinazione d'uso del solaio di sottotetto, mentre il n. 1 temi più generali.

[9] Il comitato Tecnico Scientifico (CTS) è un organo consultivo della Giunta Regionale Toscana istituito nel 2010 per l'espressione di pareri in ambito tecnico e di interpretazione normativa.

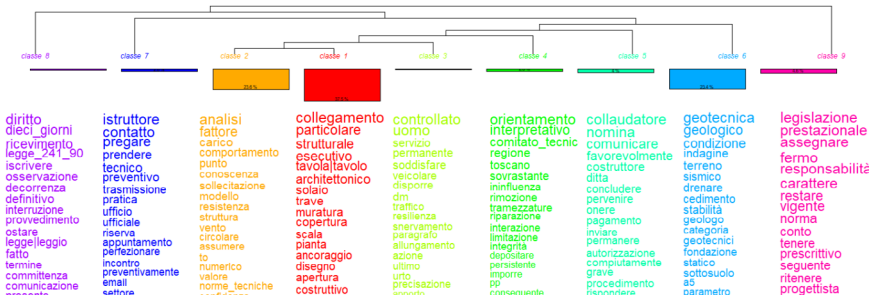
Fig. 3 Associazione della variabile relativa all'istruttore con le 4 classi.



Istruttorie 2015-2020

L'analisi del corpus delle Istruttorie per il periodo 2015-2020 con il metodo Reinert ha individuato la presenza di 9 *cluster*, 5 dei quali da interpretare come 'deboli', in quanto non mostrano masse testuali rilevanti rispetto al corpus nel suo complesso (si vedano a tal proposito le percentuali di segmenti di testo per ciascuna classe riportate in Fig. 4).

Fig. 4 *Topics* delle Istruttorie periodo 2015-2020.



Nello specifico, le classi 'deboli' derivano dalla notevole ripetitività caratterizzante il testo delle istruttorie. Alcune ripetizioni sono fisiologiche all'attività dei singoli istruttori, in quanto è usuale preparare parti di testo preimpostate per agevolare il lavoro di redazione, da adattare poi al caso specifico (esempi nella classe 3), altre sono viceversa porzioni di testo standardizzate riguardanti temi specifici (classi 5, 7, 8, 9).

Partendo da destra, la classe 9 riguarda elementi di testo preformati, non di merito tecnico, ma di carattere più generale o di premessa, come ad esempio:

> «Tenuto conto del carattere prestazionale e non prescrittivo della norma, e ferme restando le responsabilità che la legislazione vigente assegna al progettista, si ritengono comunque necessarie almeno le seguenti integrazioni/chiarimenti [...]».

Il tema, anche se non di particolare interesse tecnico, solleva comunque la riflessione se tale premessa o altre del genere siano da includere o meno nel testo base della richiesta di integrazione, cioè in quella parte narrativa preformata e identica per tutte le istruttorie che contiene riferimenti generali e normativi sul procedimento, alla quale viene appesa la parte testuale della richiesta di integrazione.

Prescindendo dalla sostanzialità e correttezza del contenuto, la scelta da prendere è se tale porzione testuale debba confluire in modo stabile nella parte del testo base dell'istruttoria, o se viceversa il testo base sia da ritenersi già completo ed allora, per uniformità, la premessa non debba comparire nelle richieste.

La classe 5 si riferisce a istruttorie che richiedono integrazioni in merito alle nomine dei soggetti da individuare, obbligatoriamente per legge, al momento del deposito del progetto per l'esecuzione dell'intervento, e che risultano mancanti. Sono correlate a porzioni di testo del tipo: «[…] produrre le nomine del Direttore dei Lavori, della Ditta Costruttrice e del Collaudatore […]», evidenziando peraltro che, trattandosi di aspetti non di merito tecnico ma procedurale, questi potrebbero essere demandati più efficacemente al controllo automatizzato da parte del sistema software durante la fase di deposito dei progetti tramite il portale web.

La classe 7 individua invece forme relative alla gestione dei rapporti tra tecnico istruttore e professionista esterno, riferendosi principalmente all'invito a prendere contatti con il tecnico istruttore al fine di definire in modo interlocutorio i contenuti dell'istruttoria prima della trasmissione ufficiale. Se da un lato quest'attività potrebbe dimostrare apertura e proiezione dell'attività dell'ufficio verso l'esterno, virtuosismo che sarebbe da incentivare, dall'altro nasconde un'eccessiva sinteticità delle richieste di integrazione. Si noti infatti come le porzioni di testo componenti la classe 7 corrispondano all'intero testo contenuto nelle istruttorie, confermando la forte sinteticità delle stesse, che spesso coincidono con la sola richiesta di contatto per discutere *de visu* gli aspetti del progetto da integrare.

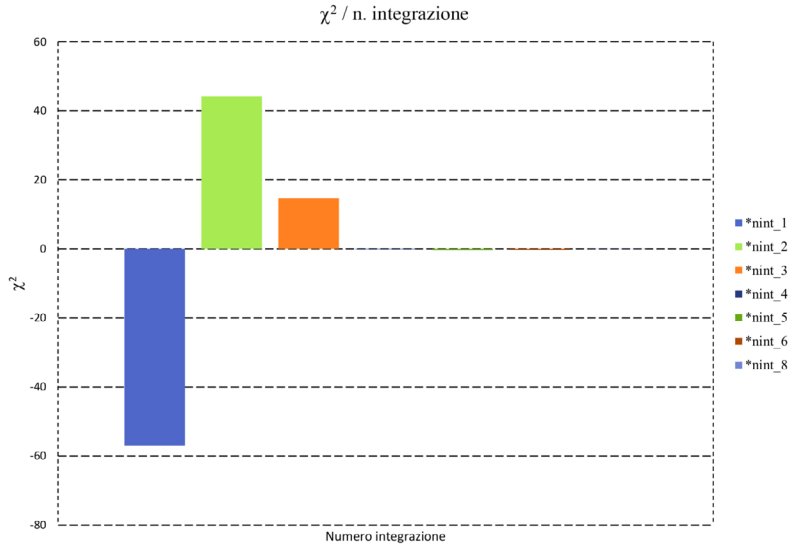
Di interesse, ma per motivi diversi, risulta essere anche la classe 8, che riguarda comunicazioni di preavviso di diniego ai sensi della L. 241/90, con testi del tipo:

> «Viste le risultanze istruttorie di cui sopra la presente è da intendersi come preavviso di diniego ai sensi dell'art.10-bis della legge n.241/1990. Entro 10 giorni dal ricevimento della presente comunicazione la S.V. ha diritto di presentare, memorie o osservazioni per dimostrare il superamento delle mancanze sopra evidenziate. Trascorsi dieci giorni dal ricevimento della presente richiesta, in permanenza dei motivi ostativi di cui sopra, sarà emesso il provvedimento definitivo di diniego».

Vista la portata delle conseguenze che tale comunicazione potrebbe generare sul proseguimento dell'istruttoria e sul procedimento in generale, è stata eseguita un'analisi del χ^2 volta ad individuare i valori di associazione tra il *cluster* 8 e alcune delle variabili con cui è stato classificato il corpus, nello specifico "numero di richiesta di integrazione", "provincia di esecuzione dell'intervento" e "tecnico istruttore associato".

La Fig. 5 (elaborata rispetto alla variabile "numero di richiesta di integrazione") mostra come la classe 8 sia sovrarappresentata soprattutto nelle seconde e nelle terze richieste di integrazione.

Fig. 5 Valori di associazione del *cluster* 8 rispetto alla variabile "numero di richiesta di integrazione".



La Fig. 6 (elaborata rispetto alla variabile "provincia di esecuzione dell'intervento") mostra il primato della Provincia di Arezzo relativamente all'emissione di preavvisi di diniego.

Infine, la Fig. 7 (elaborata rispetto alla variabile "tecnico istruttore associato") evidenzia una forte associazione positiva tra la classe 8 e cinque istruttori, mentre la restante popolazione degli istruttori ha valori di associazione tendenzialmente neutrali.

Fig. 6 Valori di associazione del *cluster* 8 rispetto alla variabile "provincia di esecuzione dell'intervento".

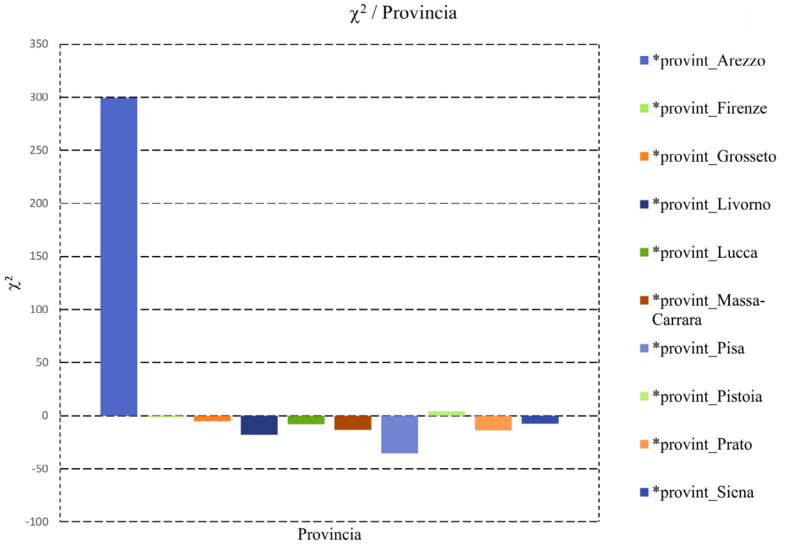
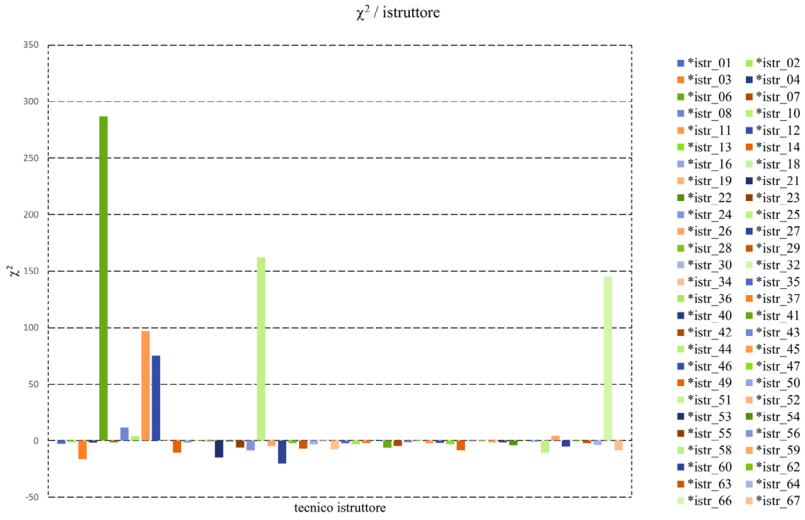


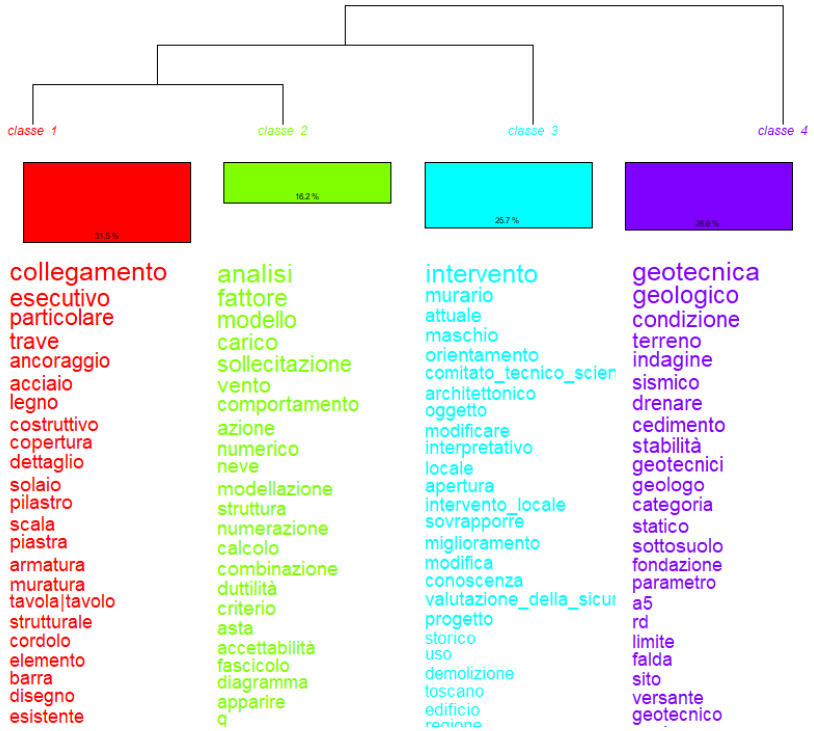
Fig. 7 Valori di associazione del cluster 8 rispetto alla variabile "tecnico istruttore associato".



Allo scopo di concentrare l'analisi sui soli *cluster* 'forti', così definiti in quanto comprendenti percentuali notevolmente maggiori di testi, la *topic detection* con metodo Reinert è stata ripetuta al fine di ottenere una rappresentazione dei soli *cluster* indicativi di una certa rappresentatività rispetto alla massa testuale del corpus. La Fig. 8 mostra il dendrogramma dei 4 *cluster* individuati dall'analisi, i quali possono essere così descritti:

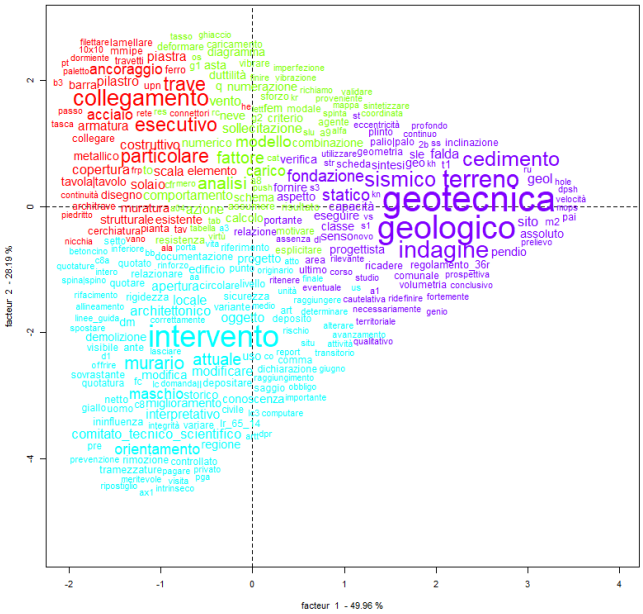
- la classe 1 fa riferimento alla richiesta di dettagli, di migliori specifiche di esecuzione contenute negli elaborati (definiamo la classe "dettagli esecutivi");
- la classe 2 attiene al tema generale del calcolo di verifica e dimensionamento delle strutture (definiamo la classe "calcolo strutturale");
- la classe 3 fa riferimento al tema di corretto inquadramento dell'intervento nell'ambito normativo di riferimento (definiamo la classe "inquadramento dell'intervento");
- la classe 4 riguarda gli aspetti specifici geologici e geotecnici e di relazione opera-terreno (definiamo la classe "geotecnica").

Fig. 8 *Topics* delle istruttorie periodo 2015-2020. Rappresentazione delle sole classi 'forti'.



Dall'analisi delle corrispondenze applicata ai *cluster,* rappresentata dal piano fattoriale in Fig. 9, si nota come le aree relative alle classi appaiano ben separate e poco sovrapposte. Il *cluster* 2, riguardante l'analisi ed il calcolo strutturale, appare come il più baricentrico e di collegamento tra gli altri temi.

Fig. 9 *Topics* delle istruttorie periodo 2015-2020. Rappresentazione su piano fattoriale delle sole classi 'forti'.



I vocabolari associati alle quattro classi evidenziano macro-ambiti di carattere piuttosto generale, non riconducibili a richieste di dettaglio, quanto piuttosto a interi settori della disciplina tecnica. In particolare, si distinguono le classi 1, 2 e 3, che riguardano in generale il 'mondo' delle strutture per la parte in elevazione, dalla classe 4, che evidenzia invece forte corrispondenza con l'area tematica geologico-geotecnica, ed in generale tratta l'interazione delle strutture con il sottosuolo.

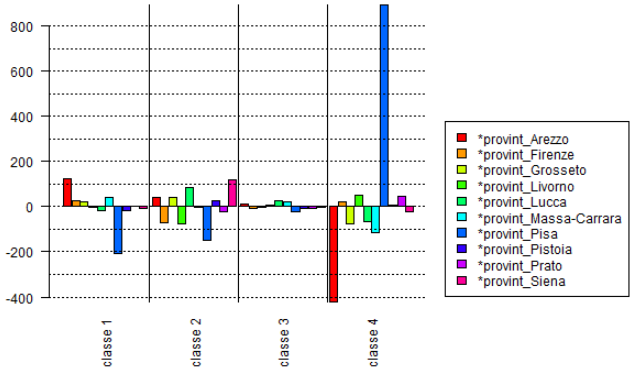
Il risultato ottenuto, pur potendosi ritenere una macro-classificazione del corpus, pone tuttavia aspetti interessanti di lettura. Delle 3 classi relative al tema strutture, si nota come la classe 1 sia riferibile alle richieste relative a dettagli esecutivi, la classe 2 all'ambito del calcolo strutturale e dell'analisi del comportamento delle strutture, e la classe 3 a forme relative all'inquadramento degli interventi. Trattandosi di classi abbastanza

generali, stupisce come, tra le forme più frequenti, non compaiano le forme riconducibili alle normative tecniche per le costruzioni di emanazione ministeriale. Tale evidenza per il *cluster* 1, che appare poco connesso al tema normativo, potrebbe essere giustificata dal fatto che esso sia più vicino a temi e prassi del buon costruire, che meno riscontrano nel disposto normativo, anche in virtù dell'evoluzione della norma nel tempo, che si caratterizza per essere sempre più di carattere prestazionale e meno prescrittivo. Il tema assume invece rilievo nella classe 4, relativa ai temi geologico-geotecnici, suggerendo che per gli istruttori in questo ambito sia più forte il richiamo prescrittivo della norma.

Nei successivi grafici sono mostrate analisi del χ^2 volte ad individuare i valori di associazione tra i 4 *cluster* 'forti' e alcune delle variabili e relative modalità nelle quali è classificato il corpus, nello specifico le variabili "provincia di esecuzione dell'intervento" (Fig. 10) e "titolo di studio del tecnico istruttore" (Fig. 11).

La Fig. 10 mostra come il tema geologico-geotecnico sia al contempo sovrautilizzato nel territorio di Pisa e sottoutilizzato in quello di Arezzo. Occorre precisare che non vi sono particolari condizioni tecniche riguardanti la natura dei terreni che giustifichino questa particolare distanza tra le due province.

Fig. 10 Valori di associazione dei 4 *cluster* rispetto alla variabile "provincia di esecuzione dell'intervento".

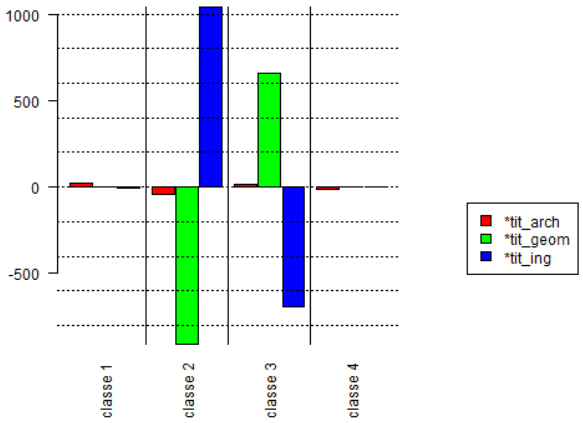


Interessante è anche l'andamento delle classi in funzione delle competenze tecniche dell'istruttore che ha redatto la richiesta (Fig. 11).

In particolare, si nota come il *cluster* 2 (analisi e calcolo) sia fortemente associato alle istruttorie svolte dagli ingegneri e viceversa sottou-

tilizzato dagli istruttori geometri: questa evidenza non può certo definirsi una sorpresa, trovando corrispondenza fisiologica nei diversi percorsi di formazione dell'istruttore, ma nell'ottica di trovare un equilibrio occorrerebbe dare maggior centralità al contenuto del calcolo per le istruttorie condotte dai geometri e, forse, spostare le attenzioni degli ingegneri su altri temi. Di lettura opposta la distribuzione della classe 3, relativa più all'inquadramento dell'intervento, che vede l'invertirsi dei due ruoli del geometra e dell'ingegnere. In questo scenario rimangono invece neutrali gli istruttori architetti.

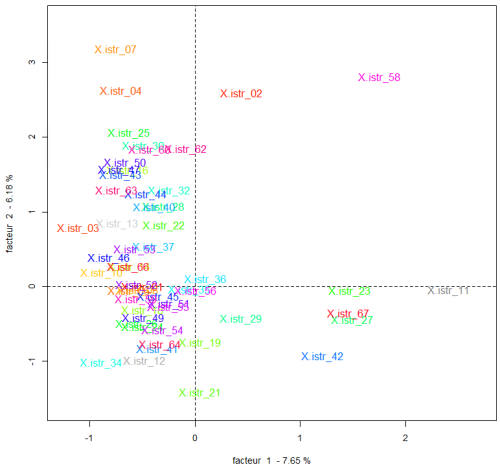
Fig. 11 Valori di associazione dei 4 *cluster* rispetto alla variabile "titolo di studio del tecnico istruttore".



Analisi delle corrispondenze e parole chiave nel corpus delle Istruttorie 2015-2020: uniformità e distanze

Al fine di cogliere l'uniformità o difformità dei comportamenti dei singoli istruttori nelle richieste di integrazione, è stata eseguita sull'intero corpus delle istruttorie 2015-2020 un'analisi delle corrispondenze, per evidenziare similarità o difformità dei vocabolari associati ai singoli istruttori, il cui risultato è rappresentato in Fig. 12. L'analisi esprime, nello specifico, le relazioni tra la variabile "tecnico istruttore associato" e relative modalità e il vocabolario alla base delle istruttorie, mostrando sul piano cartesiano il posizionamento dei singoli istruttori rispetto agli altri.

Fig. 12 Analisi delle corrispondenze corpus Istruttorie 2015-2020 in base alla variabile “tecnico istruttore associato”.



Dalla Fig. 12 è subito evidente la distanza di alcuni istruttori (in particolare gli istruttori 7, 11, 21, 42, 58) rispetto a tutti gli altri. Nello specifico, l’istruttore n. 11 contribuisce a costruire l’asse fattoriale 1, opponendosi alla maggioranza degli istruttori, mentre il n. 7 ed il n. 21 contribuiscono, opponendosi tra loro e distanziandosi dagli altri, alla costruzione dell’asse fattoriale 2. In entrambi i casi le distanze rivelano come gli istruttori periferici tendano ad utilizzare forme lessicali che non sono presenti nelle altre istruttorie. L’istruttore n. 58 è quello che più si distanzia da tutti gli altri.

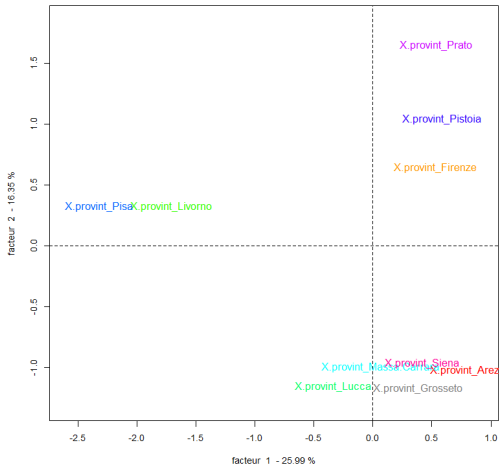
La medesima analisi è stata ripetuta aggregando il dato degli istruttori su base territoriale, utilizzando la variabile “provincia di esecuzione dell’intervento” (Fig. 13).

Ciò che salta all’occhio osservando la Fig. 13 è un’aggregazione delle province sul piano cartesiano per aree territoriali: una possibile spiegazione è data dal fatto che i tecnici progettisti che si interfacciano con l’ufficio operano mediamente su un determinato territorio, limitando di fatto la ‘popolazione’ che dall’esterno ha rapporti con l’ufficio. Per tale ragione, nelle medesime aree territoriali ci si attende una maggiore uniformità delle richieste di integrazione, in quanto certe tematiche sono già condivise, mentre altre continuano ad essere sottovalutate nella redazione dei progetti, sia per convinta interpretazione della norma che per prassi progettuale. Inoltre, anche le sedi provinciali che condividono lo stes-

so responsabile del procedimento (Lucca-Massa Carrara e Prato-Pistoia) sono nel grafico abbastanza vicine, evidenziando che l'attività svolta dai responsabili ha un certo effetto locale sulle uniformità istruttorie. Ciò potrebbe indicare una necessità di maggior confronto tra i responsabili dei vari uffici territoriali.

Infine, un'ulteriore lettura della Fig. 13 riguarda la distanza intercorrente tra territori con più elevata sismicità e maggiore storia sismica e territori a sismicità più bassa. In questo senso, la distribuzione evidenzia, nella parte destra del piano fattoriale, tutte le province della fascia appenninica, con l'aggiunta di Siena e Grosseto, che pur avendo territori prevalentemente a bassa sismicità condividono, sul confine, il territorio vicino al monte Amiata, zona classificata a maggior pericolosità. Si distanziano viceversa i territori delle province di Pisa e Livorno, che oggi presentano esclusivamente zone a bassa sismicità (zone 3 e 4)[10].

Fig. 13 Analisi delle corrispondenze corpus Istruttorie 2015-2020 in base alla variabile "provincia di esecuzione dell'intervento".



In conclusione, al fine di comprendere non solo le distanze relative tra gli istruttori, ma anche lo scostamento delle loro istruttorie rispetto ad alcuni temi definibili chiave nell'ambito del soccorso istruttorio, l'analisi del corpus delle istruttorie è stata completata con la selezione di alcune

[10] L'attuale classificazione sismica della Regione Toscana è stabilita dalla Delibera di Giunta Regionale Toscana n. 421 del 26/05/2014. La classificazione suddivide il territorio in 4 zone con pericolosità sismica decrescente dalla 1 alla 4.

parole chiave, scelte tra quelle maggiormente frequenti nel corpus e con la misura del valore χ^2 di associazione delle parole con i 5 istruttori (n. 7, 11, 21, 42, 58) che sono stati precedentemente definiti 'periferici'.

Nella Tab. 1 sono riportati i gruppi di parole chiave definiti per l'analisi.

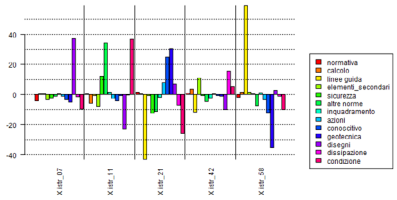
Tab. 1 Gruppi di parole chiave per la misurazione dell'associazione con gli istruttori "periferici".

Gruppi delle *keyword*							
	Calcolo	**Condizione**	**Linee guida**	**Elementi secondari**	**Geotecnica**	**Sicurezza**	**Normativa**
Keyword	Calcolo, calcolazioni, a08, a09, a8, a9, fascicolo, tabulato, verifica	Statico, sismico, condizione	Comitato tecnico-scientifico, linee-guida, orientamento, interpretativo	Secondario, tamponamento, tamponatura, tamponature, tramezzature, tramezzo, espulsione, ribaltamento	Geologico, geotecnica, geotecnico, a05, a06, a07, a5, a6, a7	Sicurezza, valutazione della sicurezza	Ntc_2018, ntc, ntc_2008, normativa, norma, norme_ tecniche

Gruppi delle *keyword*						
	Azioni	**Inquadramento**	**Disegni**	**Dissipazione**	**Conoscitiva**	**Altre norme**
Keyword	Vento, neve, accidentale, sovraccarico	Intervento locale, intervento adeguamento, intervento miglioramento, nuova costruzione	Quotatura, disegno, a10, a02, graficamente, grafico, rappresentazione, rappresentare, architettonico	Gerarchia, duttilità, plastico, q, dissipazione, dissipativo, fragile, comportamento	Fc, lc, lc1, lc2, lc3, fotografico, indagine, saggio, conoscenza, confidenza	DPR_380_01, LR_65_14

Dal grafico in Fig. 14 risulta evidente come l'istruttore n. 7 si concentri su richieste che riguardano la rappresentazione grafica, quindi relative alla completezza degli elaborati grafici esecutivi degli interventi. L'istruttore n. 11 viceversa non presenta ricorrenze del tema grafico nelle proprie istruttorie, ma supporta le sue richieste con forte utilizzo delle linee guida emanate dal Comitato Tecnico Scientifico, le quali appaiono essere invece molto poco centrali per l'istruttore n. 21, che predilige il tema geotecnico. Per l'istruttore n. 42 emergono moderatamente i temi della dissipazione nelle strutture dell'energia sismica e quelli relativi alle verifiche degli elementi non strutturali o secondari, mentre appare evidente come per il n. 58 assuma salienza il riferimento alla regolamentazione, sia essa da norma tecnica che da altre disposizioni, e risulti al contrario sottorappresentato il tema geologico-geotecnico.

Fig. 14 Associazioni tra le *keyword* e gli istruttori 'periferici'.



Analisi istruttorie 2015-2020: estrazione di 3 *subcorpora* tematici

L'analisi delle istruttorie 2015-2020 trattata nel precedente paragrafo presenta una vista sul problema dell'uniformità delle richieste di integrazione a scala macroscopica, avendo preso in considerazione, per la costruzione del corpus, la totalità delle richieste, ricomprendenti quindi tutte le tipologie possibili di intervento sulle strutture. Per analizzare con maggior dettaglio le peculiarità delle richieste d'integrazione in relazione ad alcuni specifici ambiti tematici, sono stati estratti 3 *subcorpora* dal corpus delle Istruttorie 2015-2020, sulla base delle tipologie di inquadramento dell'intervento come previste dalle norme tecniche per le costruzioni (interventi locali, miglioramento e adeguamento, nuova costruzione). I 3 *subcorpora* riguardano quindi richieste di integrazione relative a specifici interventi eseguiti sulle costruzioni esistenti e su quelle nuove.

Nello specifico, intervento locale, miglioramento e adeguamento sono le tre tipologie di intervento previste sulle costruzioni esistenti, in ordine crescente di invasività sulle strutture ed in ordine decrescente come numero di interventi depositati. Nuova costruzione è invece la tipologia di intervento relativa alla realizzazione di strutture *ex novo*.

Interventi locali

La *topic detection* tramite il metodo Reinert del *subcorpus* degli interventi locali ha individuato la presenza di 4 *cluster* principali, così caratterizzati:

- classe 1, riferita ai contenuti di calcolo e grafici specifici su edifici in muratura;
- classe 2, riferita alla richiesta di "dettagli", di migliori specifiche di esecuzione;
- classe 3, attinente al richiamo normativo;
- classe 4, riferita al procedimento in generale e agli interventi a finanziamento in particolare.

Dalla lettura del vocabolario delle forme del *cluster* 1, si rileva come sia spesso oggetto delle richieste di integrazione degli interventi locali il tema della corrispondenza tra il calcolo e gli elaborati grafici, evidenziando un problema di coerenza e congruenza interna tra gli elaborati progettuali depositati.

Il *cluster* 2, di valenza generale, è preponderante come dimensione ed è del tutto analogo nei contenuti a quanto emerso in questo studio per le analisi sugli altri corpora riguardo alle richieste di dettagli esecutivi e di definizione di certi particolari strutturali.

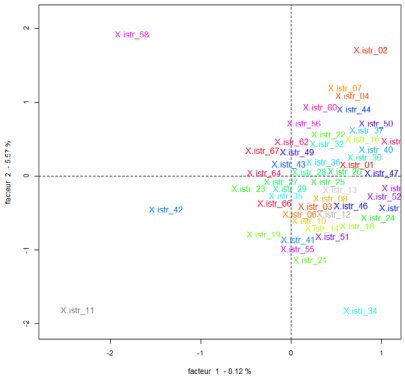
La classe 3 evidenzia i richiami regolamentari: tra le forme più frequenti nel vocabolario della classe si trovano parole come "orientamento", "interpretativo", "comitato_tecnico_scientifico", a dimostrazione che, per questa tipologia di interventi minimali, è forte l'appoggio del tecnico istruttore ai contenuti delle linee guida di emanazione regionale. Per contro, appare evidente come la minor frequenza della forma "ntc_2018" indichi un altrettanto minor utilizzo del richiamo normativo, cogente, di emanazione ministeriale. Le linee guida regionali si esprimono con orientamenti tecnici complementari ed ulteriori, di buona prassi, rispetto alla norma tecnica ministeriale, che ovviamente non possono né sostituire, né superare. Il dato che ne deriva è che sarebbe invece auspicabile un maggior riferimento ed utilizzo del richiamo alla norma tecnica, per garantire maggior robustezza alla richiesta istruttoria, da integrarsi poi con le buone prassi delle linee guida.

Nella classe 4, di carattere generale, si riscontra la presenza di richieste relative ad interventi finanziati, suggerita dalla presenza delle forme "dichiarazione", "carenza", "assenza", che richiamano un elaborato specifico, richiesto per queste particolari tipologie di progetti, le quali usufruiscono di contributi regionali per la prevenzione sismica, evidenziando come spesso i contenuti di tale elaborato specifico disattendano i requisiti minimi richiesti per accedere al finanziamento.

Da un'analisi delle corrispondenze applicata al *subcorpus* degli Interventi locali sulla base della variabile "tecnico istruttore associato" (Fig. 15) sono state individuate le similarità e le distanze tra i singoli istruttori, ossia i tecnici istruttori che più si avvicinano o allontanano dall'uniforme distribuzione delle richieste di integrazioni.

La Fig. 15 mostra chiaramente come, ad essere poco baricentrici, siano buona parte degli istruttori già identificati come tali nell'analisi del corpus delle Istruttorie 2015-2020. Nel caso degli interventi locali, si aggiungono gli istruttori n. 2 e n. 34, opponendosi tra loro sull'asse verticale rappresentativo del fattore 2.

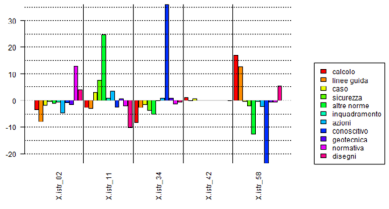
Fig. 15 Analisi delle corrispondenze *subcorpus* Interventi locali in base alla variabile "tecnico istruttore associato".



Una misura del valore χ^2 di associazione (Fig. 16) degli istruttori definiti poco baricentrici alle medesime *keyword* già impiegate nel precedente paragrafo (cfr. Tab. 1) mostra come l'istruttore 2 si distanzi dagli altri per le forme associate agli elaborati grafici ed alla normativa tecnica, mentre l'istruttore 34 si differenzi per l'uso frequente di forme legate alla fase conoscitiva e di indagine sulle strutture esistenti oggetto di intervento.

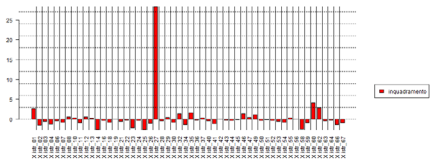
Uno dei temi delle richieste che desta maggior interesse nell'ambito della tipologia di interventi in analisi è quello dell'inquadramento. Quest'ultimo, infatti, che riguarda la distinzione tra le tipologie di intervento sulle costruzioni esistenti, è spesso motivo di richiesta d'integrazione nell'ambito degli interventi inquadrati dai progettisti come interventi locali, e conseguente innesco di una fase di contrapposizione tra l'ufficio ed il tecnico progettista esterno, che evidentemente non ha correttamente inquadrato quanto depositato rispetto alla norma tecnica. Ciò è dovuto in parte anche alla stessa normativa tecnica, che presenta un certo margine di interpretazione, inducendo così il progettista a cercare di classificare l'intervento in una fattispecie più 'leggera' – quella, appunto, dell'intervento locale – che richiede il soddisfacimento di verifiche strutturali con estensione minore rispetto alla totalità della costruzione o il raggiungimento di minori livelli prestazionali in termini di sicurezza *post operam* della costruzione.

Fig. 16 Associazioni tra le *keyword* e gli istruttori poco baricentrici per il *subcorpus* "Interventi locali".



D'altra parte, la successiva Fig. 17, che rappresenta il valore χ^2 di associazione tra la variabile "istruttore tecnico associato" nelle sue modalità (ricomprendenti, questa volta, tutta la popolazione degli istruttori e non solo quelli definiti poco baricentrici) e il gruppo di parole chiave identificate dal termine "inquadramento" (cfr. Tab. 1), mostra come tale tema sia in relazione soprattutto con le istruttorie del tecnico n. 27 (sebbene positivamente associato a numerosi altri istruttori), evidenziando come, nello specifico, l'istruttore potrebbe porre eccessiva attenzione al tema dell'inquadramento, forse anche non sufficientemente supportata da evidenze tecniche e normative.

Fig. 17 Associazioni tra la popolazione dei tecnici istruttori e il gruppo di parole chiave "inquadramento".



Interventi di miglioramento e adeguamento

Il *subcorpus* degli interventi di miglioramento e di adeguamento degli edifici esistenti, così come definiti dal Decreto Ministeriale del 17 gennaio 2018, comprende due tipologie di interventi inquadrate in modo differente dalla normativa tecnica, ma trattati in questa sede in modo unitario in quanto entrambi presuppongono l'esecuzione di un'analisi di sicurezza dell'intera unità strutturale della costruzione, non limitata solo ad alcune parti come nel caso degli interventi locali. Le differenze tra le due tipologie sono quindi riconducibili esclusivamente al diverso livello di sicurezza da ottenere post-intervento. In particolare, l'adeguamento restituisce l'edificio ad un livello di sicurezza statica e sismica pari a quello previsto dalla normativa tecnica per un edificio nuovo, mentre il miglioramento prescrive il raggiungimento della piena sicurezza statica e di un incremento della sicurezza sismica. Dal punto di vista dei contenuti dei progetti, queste due categorie sono pertanto assimilabili, almeno per gli scopi di questa ricerca.

L'analisi del *subcorpus* con il metodo Reinert ha individuato la presenza di 5 classi principali, così sintetizzabili:

- classe 1, riferita alla valutazione della sicurezza degli interventi ed ai livelli di sicurezza prescritti;

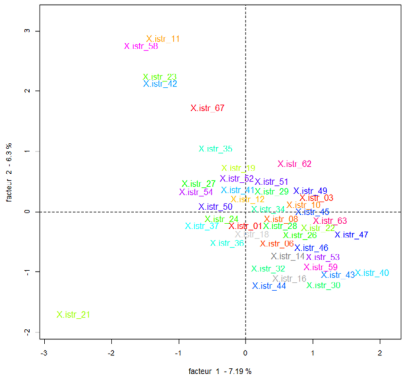
- classe 2, riferita alle indagini conoscitive sulle strutture e sui materiali;
- classe 3, riferita ai contenuti degli elaborati grafici dell'intervento;
- classe 4, riferita alla richiesta di "dettagli", di migliori specifiche di esecuzione;
- classe 5, riferita all'ambito geologico e geotecnico per la caratterizzazione dei terreni di fondazione.

Le classi individuate afferiscono ad aree tematiche distanti tra loro, con l'eccezione delle classi 3 e 4, riguardanti l'una i contenuti degli elaborati grafici dell'intervento, e l'altra la richiesta di migliori specifiche di esecuzione. I dettagli esecutivi dei contenuti progettuali si esplicitano infatti attraverso l'elaborato grafico.

La classe 5, che individua il tema geologico-geotecnico, presenta una forte rilevanza rispetto alla massa totale del *subcorpus* analizzato (circa il 25%), delineando una sottovalutazione di tale argomento negli interventi di miglioramento e di adeguamento.

La Fig. 18 mostra i risultati di un'analisi delle corrispondenze applicata al *subcorpus* degli Interventi di miglioramento e adeguamento in base alla variabile "tecnico istruttore associato". Analogamente al *subcorpus* degli Interventi locali, si osservano anche in questo caso alcuni tecnici istruttori più distanti dagli altri, nello specifico i n. 11, 21 e 58 (Fig. 18).

Fig. 18 Analisi delle corrispondenze *subcorpus* Interventi di miglioramento e adeguamento in base alla variabile "tecnico istruttore associato".



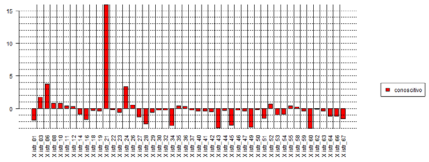
Preme sottolineare come la posizione poco baricentrica non individui un'istruttoria errata nei contenuti o genericamente non corretta, concentrandosi esclusivamente sulla maggiore o minore distanza delle forme lessicali usate dagli istruttori. La presenza di istruttori 'periferici' o poco baricentrici delinea una non uniformità delle istruttorie degli stessi rispetto alla popolazione degli istruttori, ma non necessariamente determina un'errata impostazione dell'istruttoria o una non corretta interpretazione dei contenuti progettuali.

Nei due grafici che seguono (Fig. 19 e 20) è mostrato il valore χ^2 di associazione tra la popolazione degli istruttori (variabile "tecnico istruttore associato" e relative modalità) e le parole chiave relative ai gruppi delle *keyword* rientranti sotto le etichette "conoscitiva" e "calcolo" (cfr. Tab. 1).

Gli interventi di miglioramento e di adeguamento sono di fatto gli interventi sugli edifici esistenti che richiedono un maggior onere di elaborazione della fase progettuale da parte del professionista esterno. Questi interventi, infatti, presuppongono l'esecuzione di un'approfondita fase conoscitiva delle strutture dell'edificio esistente sul quale si interviene, e una dettagliata fase di analisi e di calcolo. Le due fasi conoscitiva e di calcolo sono dunque entrambe centrali per la corretta definizione di queste tipologie di intervento. Un sovrautilizzo di questi temi nelle richieste di integrazione suggerisce che viene posta una particolare attenzione su temi sostanziali per le tipologie di intervento in esame, mentre un sottoutilizzo delinea una sottostima della rilevanza di questi aspetti, determinando la necessità di una riflessione sui contenuti di tali richieste. Si tornerà a breve su questo punto.

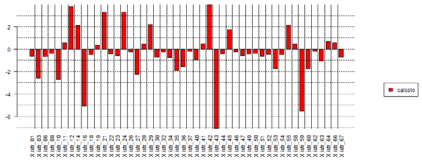
La Fig. 19 mostra come i temi sottesi alla fase conoscitiva abbiano valori di associazione tendenzialmente neutrali con i vari istruttori, con l'eccezione dell'istruttore n. 21, che mostra un particolare utilizzo di forme come "indagine", "saggio", "conoscenza", "fattore di confidenza (fc)", e "livello di conoscenza (lc)" nelle proprie istruttorie.

Fig. 19 Associazioni tra la popolazione dei tecnici istruttori e il gruppo di parole chiave "conoscitiva".



Nel caso, viceversa, delle *keyword* relative alla tematica del calcolo, come "fascicolo", "tabulato", "verifica", appare evidente dalla Fig. 20 un sottoutilizzo da parte di tre istruttori in particolare (n. 16, 43 e 59).

Fig. 20 Associazioni tra la popolazione dei tecnici istruttori e il gruppo di parole chiave "calcolo".



La variabilità nell'utilizzo del lessico appartenente ai due gruppi di parole chiave potrebbe anche essere giustificata dalla particolare circostanza data ad alcuni istruttori di avere assegnato a controllo pratiche che non mostrano particolari carenze su questi temi, tali da non essere oggetto di specifica richiesta. D'altro canto, la probabilità di una simile circostanza diventa nulla se si assume a priori un'equa distribuzione della qualità dei progetti depositati, ipotesi del tutto realistica, essendo la qualità una variabile determinata da un fattore esterno e, pertanto, da ritenersi indipendente rispetto al procedimento di controllo operato dall'ufficio. Inoltre, i progetti devono essere redatti secondo legge da tecnici iscritti nei rispettivi albi professionali, determinando con questo che ciascun professionista, nell'ambito delle proprie competenze, possiede i requisiti professionali specifici per garantire i livelli minimi di sicurezza che la norma richiede, e dunque, l'assenza di progetti di migliore o peggior pregio. Infine, anche ammettendo la presenza di progetti che non richiedano integrazioni nei temi in esame, questi ultimi sarebbero comunque residuali e non determinerebbero effetti significativi nei valori analizzati, in considerazione della numerosità delle istruttorie analizzate.

Interventi di nuova costruzione

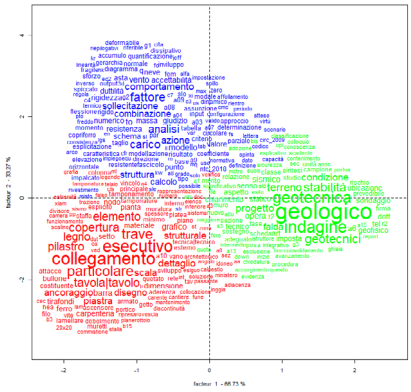
Il *subcorpus* degli Interventi di nuova costruzione riguarda le richieste d'integrazione attinenti ai progetti delle strutture di nuova realizzazione. Per queste strutture, la normativa tecnica richiede prestazioni specifiche, sia in termini di livelli di sicurezza non derogabili, sia nel senso della prescrizione di specifiche verifiche e dell'utilizzo di particolari dettagli costruttivi, differenziati in funzione della tipologia di struttura, del materiale costituente e del comportamento strutturale della costruzione in risposta alle azioni sollecitanti imposte (siano esse di tipo statico o dinamico, come ad esempio quelle di natura sismica).

L'analisi con il metodo Reinert applicata al *subcorpus* in esame ha individuato 3 *cluster* principali ben distinti, come evidenziato dall'analisi delle corrispondenze applicata ai *cluster* rappresentata in Fig. 21.

In particolare, le classi sono così caratterizzate:

- classe 1, riferita alla richiesta di dettagli, di migliori specifiche di esecuzione;
- classe 2, attinente al calcolo in generale;
- classe 3, riferita ai temi geologico-geotecnici e caratterizzazione dei terreni di fondazione.

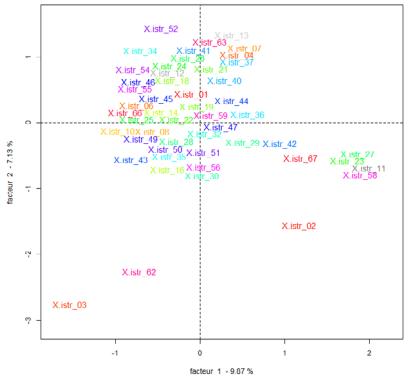
Fig. 21 Distanze tra i *cluster* sul piano cartesiano.



L'analisi delle corrispondenze applicata al *subcorpus* Interventi di nuova costruzione a partire dalla variabile "tecnico istruttore associato" ha individuato, anche in questo caso, alcuni tecnici istruttori più distanti dagli altri (Fig. 22). In particolare, si riscontrano nelle posizioni più periferiche del grafico gli istruttori n. 2, 3, 52 e 62, che sovrautilizzano o sottoutilizzano specifiche forme nelle loro richieste di integrazione ri-

spetto all'uniforme distribuzione del comportamento dei restanti tecnici istruttori.

Fig. 22 Analisi delle corrispondenze *subcorpus* Interventi di nuova costruzione in base alla variabile "tecnico istruttore associato".



Per dare una corretta interpretazione dei fattori predominanti che determinano questo risultato, è stata misurata l'intensità della relazione tra gli istruttori in posizione periferica (n. 2, 3, 52 e 62) e l'insieme delle *keyword* di cui alla Tab. 1 del presente capitolo, determinando il valore χ^2 di associazione tra i 4 istruttori 'periferici' e le *keyword* (Fig. 23).

Nella Fig. 23 è in particolare evidente come gli istruttori n. 2 e 3 si allontanino dall'utilizzo equidistribuito dei temi di normativa, geotecnica, calcolo, dissipazione ed elementi secondari.

Gli argomenti significativi che dovrebbero contraddistinguere gli interventi di nuova costruzione sono, in generale, il calcolo e la dissipazione di energia sismica e le caratteristiche di duttilità della costruzione.

Le Fig. 24 e 25 mostrano come varia il ricorso a questi specifici argomenti tra i vari istruttori, sulla base del valore χ^2 di associazione tra la variabile "tecnico istruttore associato" nelle sue varie modalità e i gruppi di parole chiave relativi al "calcolo" e alla "dissipazione".

Se da un lato il calcolo è ambito classico dei contenuti dei progetti strutturali, dall'altro la duttilità e la dissipazione sono temi relativamente recenti nell'analisi del comportamento delle strutture, essendo entrati efficacemente nel dettato normativo solo con il Decreto Ministeriale 14 gennaio 2008. L'analisi eseguita mostra, per entrambe le tematiche, un parziale sottoutilizzo nelle richieste di integrazione, dovuto ad alcuni specifici istruttori (più numerosi nella materia del "calcolo" rispetto a quella della "dissipazione").

Fig. 23 Associazioni tra le *keyword* e gli istruttori poco baricentrici per il *subcorpus* “Interventi di nuova costruzione”.

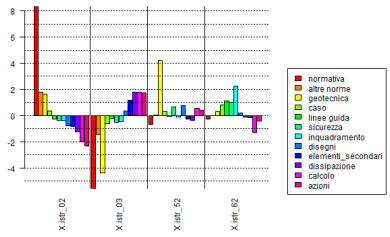


Fig. 24 Associazioni tra la popolazione dei tecnici istruttori e il gruppo di parole chiave “calcolo”.

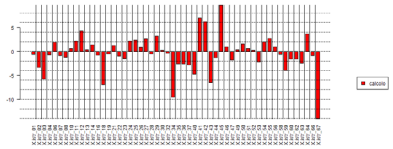
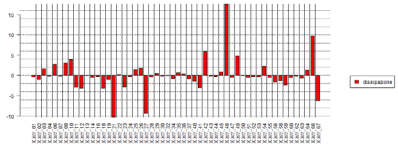


Fig. 25 Associazioni tra la popolazione dei tecnici istruttori e il gruppo di parole chiave “dissipazione”.



Conclusioni

Il capitolo ha presentato l’analisi, tramite tecniche di text mining, dei contenuti delle richieste di integrazione prodotte dai tecnici istruttori del Settore Sismica della Regione Toscana nel corso della fase istruttoria del procedimento di controllo dei progetti strutturali delle costruzioni edilizie che sono depositati presso il medesimo Settore e che per legge sono assoggettati all’attività di controllo.

L’obiettivo principale del lavoro svolto è quello di fornire uno strumento di valutazione della fase istruttoria stessa, favorendo una maggiore uniformità delle istruttorie e il conseguente miglioramento della qualità delle scelte conseguenti alla fase istruttoria, a garanzia della riduzione

della discrezionalità amministrativa, oltre che della pubblica incolumità derivante da una maggiore sicurezza delle costruzioni e del bene pubblico. La riduzione o l'eliminazione dei meccanismi distorsivi del buon andamento del procedimento amministrativo – nel Settore considerato da sempre constatati ma finora mai analizzati – è infatti un aspetto rilevante per l'affermazione dei principi di equilibrio e parità di trattamento, che dovrebbero contraddistinguere l'azione della p.a., oltre che per la percezione esterna dell'attività dell'ufficio. Nel caso studio considerato, la difformità dell'attività istruttoria ha infatti alimentato nel tempo contenziosi e critiche da parte dell'opinione pubblica.

Le evidenze prodotte dall'analisi hanno messo in luce, in particolare, i temi alla base della mancanza di uniformità caratterizzante il procedimento in esame, suggerendo su quali focalizzare l'attenzione dei tecnici, perché più significativi, e su quali invece alleggerirla, in quanto meno concorrenti alla garanzia della sicurezza delle costruzioni.

I risultati ottenuti hanno rimarcato la presenza di difformità nelle richieste istruttorie sia a livello del singolo istruttore, sia con riferimento al dato aggregato su base territoriale dell'ufficio di competenza e con riferimento alla specifica formazione delle competenze dei tecnici del settore. Sebbene l'eliminazione della discrezionalità in senso assoluto sarebbe utopica, anche in considerazione della materia trattata e della normativa tecnica di riferimento, che spesso trova margine di interpretazione, lo studio ha comunque individuato elementi del procedimento che non sono da ritenersi fisiologici e sui quali pertanto si può incidere con opportuni correttivi.

La modalità di analisi su un singolo progetto sottoposto alla popolazione degli istruttori (Progetto test) presentata nella prima parte del capitolo si è dimostrata più efficace nell'individuazione dei temi salienti trattati nelle istruttorie. I temi individuati tramite questa modalità di indagine consentono infatti di far emergere gli argomenti trattati poco o poco coerentemente dal testo di norma tecnica e che per questo determinano maggiori difficoltà applicative anche nello svolgimento delle istruttorie, configurando la necessità, su quei temi specifici, di interpretazioni normative condivise da applicare uniformemente sul territorio, o comunque in generale la necessità di approfondimenti tecnico-scientifici più di dettaglio per cogliere correttamente la richiesta di prestazione dettata da normativa. Questa modalità di analisi è riuscita inoltre ad evidenziare carenze nell'impostazione delle istruttorie, in particolare sull'inquadramento del progetto rispetto alla norma, evidenziando di fatto un possibile pro-

blema di sottovalutazione dell'inquadramento dell'intervento da parte di taluni istruttori; questo può avere potenziali ripercussioni sulla sicurezza degli edifici, oltre che conseguenze sulla corretta applicazione della disciplina sanzionatoria e degli aspetti di rilevanza penale determinati dalla violazione della normativa tecnica per le costruzioni. Nonostante, quindi, l'obiettivo dell'analisi non sia quello di indagare la correttezza o meno dei contenuti della richiesta prodotta dal tecnico, alla luce dei risultati ottenuti si può affermare che la somministrazione periodica di un progetto alla popolazione dei tecnici istruttori potrebbe favorire in certi termini un controllo sulla misura della correttezza istruttoria. Le principali criticità dei risultati ottenuti sono determinate dalla scarsa efficacia del Progetto Test nell'individuare la variabilità e le distanze tra gli istruttori, e, allo stato attuale, dalla limitata estensione del database delle richieste di integrazioni utilizzato per l'individuazione dei temi da approfondire, da considerare fisiologica in rapporto al numero degli istruttori, soprattutto se, come nel nostro caso, connessa ad un tasso di partecipazione al progetto non globale e all'eventuale presenza di istruttori che ritengono il progetto sostanzialmente corretto, non richiedendo integrazioni.

L'analisi condotta sull'insieme di tutte le istruttorie per il periodo 2015-2020 e presentata nella seconda parte del capitolo, poi declinata con maggior dettaglio nei vari *subcorpora* estratti, si è rivelata essere più efficace nell'individuazione di comportamenti non uniformi tra tecnici istruttori o tra presidi territoriali, individuando quei comportamenti che nel corso del capitolo sono stati definiti poco baricentrici o periferici. Questo tipo di analisi, trattando un dataset di progetti in generale tutti diversi l'uno dall'altro, è riuscito tuttavia a cogliere meno efficacemente gli argomenti tecnici più di dettaglio che potrebbero essere oggetto di approfondimento interpretativo. La *topic detection* del corpus delle istruttorie ha inoltre consentito l'individuazione di alcuni *cluster* 'deboli', che seppure non rilevanti in termini quantitativi, forniscono comunque informazioni utili dal punto di vista qualitativo, rivelando un sovrautilizzo a livello territoriale e da parte di taluni istruttori poco giustificabile. Se da un lato si attribuisce scarso merito tecnico a queste parti di contributo istruttorio, dall'altro la non uniformità nella risposta verso l'esterno, e comunque un approccio poco snello del procedimento, evidenziano una problematica da risolvere.

In tutti gli studi, oltre la conferma di quello che ci si attende, si spera sempre di far emergere la sorpresa, terreno fertile per poi far germogliare altre innovazioni o altra ricerca.

«*Mining implies extracting precious nuggets of ore from otherwise worth-*

> *less rock. If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. [...] in the case of text, it can be interesting to take the mining-for- nuggets metaphor seriously»*[11] (Hearst 1999).

Nel nostro caso, la sorpresa maggiore è dipesa dal fatto che la 'nuvola' degli istruttori rilevata dall'analisi delle corrispondenze applicata alle istruttorie 2015-2020 non è molto dispersa, a meno di pochi casi e considerando che le difformità possono essere dovute anche alle diverse tipologie di istruttorie trattate. Quello che si coglie dai risultati è una certa specializzazione degli istruttori poco baricentrici, che tendono a distanziarsi dagli altri in relazione a particolari tematiche, e che richiedono nelle istruttorie approfondimenti talvolta di esteso dettaglio. Questo aspetto, senza nulla togliere alla correttezza delle richieste di integrazione, evidenzia come, nelle istruttorie emesse dai tecnici 'periferici', potrebbero essere spesso sottovalutati temi più generali e di inquadramento, che nella maggior parte dei casi sono gli aspetti dirimenti circa il corretto allineamento del progetto al disposto normativo. Il corretto inquadramento dell'intervento può incidere in modo sostanziale sulla sicurezza della costruzione raggiunta con gli interventi, in quanto, è proprio in base ad esso che la norma definisce l'estensione e l'approfondimento delle verifiche di sicurezza, e il raggiungimento dei livelli di sicurezza prestabiliti *post operam*. In alcune tipologie di intervento, infatti, le verifiche devono essere estese all'intero organismo strutturale, mentre in altre sono limitate al solo intorno degli elementi sui quali si interviene: è evidente che il non corretto inquadramento potrebbe determinare una sottovalutazione della sicurezza, se non addirittura un peggioramento del comportamento di altre parti di struttura non studiate dal progetto e non considerate nelle verifiche. Allo stesso tempo, per alcune tipologie di interventi sono stabiliti livelli minimi di sicurezza non derogabili, sia come incrementi relativi tra prima e dopo l'intervento che in termini assoluti. Anche in questo caso, il non corretto inquadramento potrebbe incidere negativamente sulla sicurezza raggiunta.

La seconda sorpresa riguarda i pochi riferimenti alla normativa tecni-

[11] «Il *mining* implica l'estrazione di pepite di minerale prezioso da roccia altrimenti priva di valore. Se il *data mining* seguisse davvero questa metafora, ciò significherebbe che le persone stanno scoprendo nuovi fatti all'interno delle loro collezioni di dati. Tuttavia in pratica non è proprio così. [...] nel caso del testo, può essere interessante prendere sul serio la metafora dell'estrazione di pepite» (trad. nostra).

ca sulle costruzioni nelle istruttorie, che emergono raramente tra le forme con più alta frequenza. Questo risultato non conforta, considerando che da un lato la norma è il riferimento che definisce i livelli minimi di sicurezza, dall'altro è l'unica base inattaccabile di qualunque istruttoria, anche a garanzia dell'operato del tecnico istruttore. Inoltre, spesso la normativa tecnica viene vista come la definizione di un mero limite entro il quale il tecnico progettista deve muoversi. La norma, in realtà, garantendo la sicurezza e la pubblica incolumità, individua criteri che traducono in termini operativi un contratto sociale. I livelli di sicurezza stabiliti dalle verifiche prescritte secondo norma sono infatti la trasposizione, su una base di calcolo probabilistico, del livello di sicurezza socialmente accettabile riferito alla possibilità di perdita, in termini di vite umane ed in termini di perdita economica. In quest'ottica, il mancato rispetto della norma tecnica, così come il non corretto controllo previsto per legge, determina, oltre ad un potenziale rischio per la sicurezza della popolazione, anche la violazione di un contratto sociale.

Riguardo all'affidabilità degli esiti della ricerca, occorre precisare che i risultati ottenuti e tutta la ricerca derivano dall'applicazione di un approccio di tipo quali-quantitativo all'analisi del testo, che associa all'analisi dei contenuti una parte di analisi e interpretazione di tipo qualitativo, condotta da chi effettua la ricerca e che potrebbe essere interpretata diversamente, o potrebbe essere indirizzata a ricercare evidenze diverse.

In conclusione, un possibile sviluppo e applicazione dell'analisi delle istruttorie presentata nel capitolo potrebbe essere quello di condurre l'analisi già in fase di deposito del progetto dell'intervento, richiedendo ai progettisti, tra gli elaborati, una relazione descrittiva, da produrre in una data forma tipicizzata, in modo tale da estrarne in automatico un'analisi dei contenuti e fornire al tecnico istruttore, oltre ai temi salienti del progetto sui quali porre maggiore attenzione, anche l'indicazione di specifiche criticità al verificarsi di un certo vocabolario, significative per individuare i disallineamenti rispetto alle prescrizioni di norma tecnica.

Ulteriori sviluppi potrebbero individuare filtri ed automatismi da implementare sul portale web di trasmissione dei progetti strutturali, per aumentare l'efficacia del controllo. Al momento in cui si scrive, sulla base dei risultati dell'analisi delle istruttorie 2015-2020 presentata in questa sede ed in particolare delle evidenze prodotte dall'analisi dei c.d. *cluster* 'deboli', in Regione Toscana è allo studio la revisione dei criteri dei controlli automatici che sono condotti nella fase di trasmissione dei progetti, oltre che una generale revisione delle parti testuali di narrativa fissa e predeterminata che compongono la base dei documenti prodotti dall'uffi-

cio nello svolgimento dell'attività istruttoria.

Infine, ma non per ultimo, uno tra gli sviluppi futuri più interessanti potrebbe essere l'implementazione dell'analisi del contenuto anche ad altri procedimenti istruttori del Settore Sismica o di altri settori, auspicabile soprattutto nell'ottica di dare ulteriore forza e centralità alla fase istruttoria del procedimento amministrativo.

La matrice per la sostenibilità: dall'armonizzazione dei sistemi contabili ad Agenda 2030. Il caso della Regione Toscana

Simone De Lellis[1]

Bilancio, Missioni, Programmi, Sostenibilità, Agenda 2030, Regione Toscana.

Introduzione

Programmare significa confrontarsi con un'ipotesi di futuro, ma per riuscire a farlo in modo adeguato è importante soffermarsi sul processo di armonizzazione contabile. Infatti, riuscire a rendere più trasparente e veritiera la rappresentazione contabile della situazione finanziaria, economica e patrimoniale di un ente è, senza dubbio, la prima e irrinunciabile condizione da soddisfare affinché la funzione di programmazione possa svolgersi in modo efficace.

Con questo lavoro si vuole coniugare il tema della sostenibilità economica, sociale e ambientale, proposto dall'Agenda 2030, con il tema del bilancio ed in particolare con il sistema di classificazione della spesa pubblica individuato dal Decreto legislativo 23 giugno 2011, n. 118 (D.lgs. 118/11, Allegato 14), proponendo un'analisi della dimensione economica della sostenibilità attraverso la creazione di una matrice di correlazione tra le Missioni/Programmi della spesa e gli Obiettivi di Agenda 2030. Temi che riteniamo molto importanti soprattutto nel periodo storico che

[1] Ha lavorato come ricercatore all'Istituto Regionale Programmazione economica della Toscana (IRPET) e attualmente è funzionario della programmazione presso Regione Toscana.

stiamo attraversando, dove risulta fondamentale l'ottimale allocazione delle risorse.

La Regione Toscana ha avviato, sin dalla fine del 2018[2], un percorso di formazione della Strategia regionale di sviluppo sostenibile. In questo contesto si colloca l'idea di questo progetto, con cui si vuole offrire uno strumento di supporto per l'analisi delle politiche di bilancio, che risultano fondamentali anche per capire al meglio lo sviluppo delle strategie per la sostenibilità.

Nel prossimo futuro le politiche, così come la concessione dei finanziamenti, potrebbero essere formulate dalle amministrazioni pubbliche sulla base degli Obiettivi di sviluppo sostenibile (*Sustainable Development Goals – SDGs*) dell'Agenda 2030, tramite il loro impiego come strumenti di riferimento della programmazione delle politiche locali, regionali, nazionali ed europee.

A partire dall'ipotesi di una sinergia tra gli standard di classificazione della spesa pubblica in Missioni/Programmi, a cui viene associata la classificazione COFOG, e gli *SDGs* dell'Agenda 2030, nel capitolo sono presentate la costruzione e alcune possibili applicazioni di una matrice di correlazione tra le modalità di distribuzione della spesa pubblica e gli Obiettivi di sviluppo sostenibile. La matrice e le analisi sono state realizzate con il supporto di strumenti e tecniche di text mining, attraverso un approccio di tipo quali-quantitativo (tramite l'uso del software Iramuteq), che sarà descritto dettagliatamente nei paragrafi successivi.

La matrice della sostenibilità può essere impiegata come modello standard declinabile a tutti i livelli di amministrazione pubblica (locale, regionale, nazionale ed europeo), per supportare la programmazione, l'analisi e la verifica delle politiche e degli interventi, sulla base della distribuzione della spesa tra gli *SDGs* di Agenda 2030.

Nei prossimi paragrafi, a seguito di un breve *excursus* sull'Agenda 2030 e della definizione del sistema di classificazione della spesa impiegato nell'analisi, sarà presentato il modello della matrice per la sostenibilità, di cui è stata effettuata una prima applicazione tramite un'analisi della distribuzione delle previsioni di spesa[3] per competenza della Regio-

[2] Delibera della Giunta Regionale della Toscana n. 1079 del 1° ottobre 2018.

[3] I dati di previsione della spesa possono essere reperiti, come *open data*, dalla Banca Dati Amministrazioni Pubbliche (BDAP) della Ragioneria Generale dello Stato (RGS) del Ministero dell'Economia e delle Finanze (MEF). Per le Regioni ed Enti Locali nella sezione "2. Bilancio di previsione spese": http://www.bdap.tesoro.it/sites/openbdap/cittadini/bilancideglienti/bilancientipubbammentvig/bilanciregionientiorganismi/Pagine/SchedaContenutoBilanciArmonizzati.aspx. Per lo Stato Italiano da "Le tavole del Bilancio sem-

ne Toscana tra gli *SDGs*, con riferimento al periodo di programmazione del Programma Regionale di Sviluppo (PRS) 2016-2020. In aggiunta, sono portati, come elementi di *benchmark*, i dati dell'insieme di tutte le regioni italiane, del Comune di Firenze e dello Stato italiano. Il confronto tra questi differenti livelli di governo, oltre a fornire un importante termine di paragone per la Regione Toscana, è volto a dimostrare la valenza dell'idea progettuale e della correlazione individuata tra le categorie di spesa e gli *SDGs*, in considerazione delle diverse competenze in gioco.

La dimensione economica della sostenibilità

Agenda 2030 e lo sviluppo sostenibile

L'Agenda 2030 e gli Obiettivi di sviluppo sostenibile furono approvati a New York il 25 settembre 2015. In quella data i 193 Paesi membri delle Nazioni Unite adottarono all'unanimità la risoluzione 70/1 intitolata *Trasformare il nostro mondo: l'Agenda 2030 per lo sviluppo sostenibile*. Il 1° gennaio 2016 entrarono quindi in vigore a livello internazionale gli Obiettivi di sviluppo sostenibile (*Sustainable Development Goals – SDGs*). L'Agenda 2030 e gli *SDGs* costituiscono il quadro di riferimento per lo sviluppo, dopo la conclusione della fase degli Obiettivi di Sviluppo del Millennio (MDGs), che avevano orientato l'azione internazionale nel periodo 2000-2015.

L'Agenda globale comprende 17 Obiettivi articolati in 169 *target*. Gli Obiettivi (Fig. 1), interconnessi e indivisibili, bilanciano le tre dimensioni dello sviluppo sostenibile: crescita economica, inclusione sociale, tutela dell'ambiente. La realizzazione degli Obiettivi di sviluppo sostenibile, a carattere universale, è rimessa all'impegno di tutti gli Stati, attraverso l'adozione di strategie nazionali.

A livello europeo, lo sviluppo sostenibile è al centro delle politiche dell'Unione fin dalla sua inclusione nei Trattati, a partire dal Trattato di Amsterdam del 1997. Dopo una fase iniziale di lenta reazione dell'Unione nel recepimento dell'Agenda 2030, la Commissione Europea formalizzò il massimo impegno per allinearsi al documento ONU, dichiarandolo «interamente coerente con la visione europea».

plificato dello Stato in formato editabile", Tavola 3.2.2 – Analisi delle spese di previsione per Missioni e Programmi, indicando la Legge di Bilancio, per il 2019: http://www.rgs.mef.gov.it/VERSIONE-I/attivita_istituzionali/formazione_e_gestione_del_bilancio/bilancio_di_previsione/bilancio_semplificato/.

Fig. 1 Gli obiettivi di sviluppo sostenibile di Agenda 2030.

In Italia, la Strategia nazionale per lo sviluppo sostenibile fu adottata nel 2002, e in seguito aggiornata con l'entrata in vigore il 2 febbraio 2016 delle prescrizioni contenute nell'articolo 3 della legge 28 dicembre 2015, n. 221. Il 13 marzo 2017 il Ministero dell'ambiente presentò quindi la bozza della nuova Strategia nazionale per lo sviluppo sostenibile e, pochi giorni dopo, avviò un processo di consultazione con la società civile. La Strategia fu presentata al Consiglio dei Ministri il 2 ottobre 2017 e approvata dall'ex Comitato interministeriale per la programmazione economica (CIPE) il 22 dicembre 2017, al temine di un intenso lavoro coordinato dal Ministero dell'ambiente in stretta collaborazione con la Presidenza del Consiglio dei Ministri, con il Ministero degli affari esteri e della cooperazione internazionale e con il Ministero dell'economia. La Strategia fa riferimento alla struttura delle 5P introdotta dall'Agenda 2030 (*People, Planet, Prosperity, Peace and Partnership*). Centrale nell'attuazione dell'Agenda 2030 è il coinvolgimento delle regioni nel disegno di strategie territoriali di sviluppo sostenibile. Dal punto di vista della partecipazione della società civile e della diffusione degli Obiettivi di sviluppo sostenibile, una realtà significativa è rappresentata dall'Alleanza italiana per lo Sviluppo Sostenibile (ASviS)[4].

Il monitoraggio dei risultati previsto dall'Agenda 2030 si basa su un insieme di 232 indicatori statistici globali (*global indicator framework*), individuati per misurare il raggiungimento dei *target* degli *SDGs* ed elaborati dall'*Interagency and Expert Group on SDG Indicators* (*IAEG-SDGs*), composta da rappresentanti degli Stati membri e, in qualità di osservatori, da esponenti di agenzie regionali ed internazionali. Gli indicatori, indivi-

[4] L'Alleanza Italiana per lo Sviluppo Sostenibile (ASviS) è nata il 3 febbraio del 2016, su iniziativa della Fondazione Unipolis e dell'Università di Roma "Tor Vergata", per far crescere nella società italiana, nei soggetti economici e nelle istituzioni la consapevolezza dell'importanza dell'Agenda 2030 per lo sviluppo sostenibile e per mobilitarli alla realizzazione degli Obiettivi di sviluppo sostenibile (*SDGs - Sustainable Development Goals*). L'Alleanza riunisce attualmente oltre 300 tra le più importanti istituzioni e reti della società civile.

duati in accordo con la *UN Statistical Commission* e adottati dall'Assemblea Generale il 6 luglio 2017 (risoluzione A/RES/71/313), sono integrati dagli indicatori per i livelli nazionali e regionali, sviluppati, invece, dai singoli Stati membri. L'UE propone anche un monitoraggio attraverso un *set* di 100 indicatori stilato da Eurostat. Per l'Italia è l'Istat, in particolare, a svolgere un ruolo attivo di coordinamento nazionale nella produzione degli indicatori per la misurazione dello sviluppo sostenibile e il monitoraggio dei suoi obiettivi.

La Toscana fu la prima Regione in Italia ad introdurre nel proprio Statuto il principio dello sviluppo sostenibile (art. 4, comma 1, lettera n). Con la Legge regionale 7 agosto 2018, n. 48 (L.R. 48/18) "Norme in materia di economia circolare. Modifiche alla L.R. 1/2015" fu disposto il passaggio della programmazione regionale verso l'economia circolare, attraverso l'individuazione di obiettivi e contenuti minimi definiti dal Programma Regionale di Sviluppo (PRS) ed il coordinamento con i piani di settore regionali. La Strategia della Toscana per lo sviluppo sostenibile prese avvio con la partecipazione della Toscana, con Deliberazione n. 1079 del 1° ottobre 2018, ad un bando del Ministero dell'ambiente e della tutela del territorio e del mare (oggi Ministero della transizione ecologica) dedicato al finanziamento di attività di supporto alla realizzazione degli adempimenti previsti dall'art. 34 del D.lgs n. 152/06, mediante il progetto 'Predisposizione del percorso di formazione della Strategia regionale di sviluppo sostenibile'. Con la Decisione n. 16 del 18 febbraio 2019, la Giunta Regionale diede formalmente attuazione al progetto di predisposizione della Strategia, prevedendo il rafforzamento della governance interna attraverso l'istituzione della Cabina di Regia istituzionale e del Tavolo tecnico di coordinamento. Alla base della Decisione vi fu anche una decisa volontà di ascolto dal basso per la definizione della Strategia regionale, promuovendo forme di partecipazione innovative e rivolte a coinvolgere tutte le fasce della popolazione. A tal proposito, ulteriore strumento di attuazione è il Forum regionale per lo Sviluppo Sostenibile, presieduto dall'Assessore all'ambiente e la difesa del suolo e coordinato dalla Direzione Ambiente ed Energia. Obiettivo è quello di favorire lo scambio di informazioni e il *networking* tra gli attori della sostenibilità, a tutti i livelli.

Il sistema armonizzato della contabilità

La necessità di avviare il processo di armonizzazione contabile del settore pubblico discende principalmente dalle esigenze di integrazione poste dall'Unione Europea con la Direttiva 2011/85/UE, che delineò il

quadro dei requisiti per gli schemi di bilancio degli Stati membri, fornendo chiare indicazioni in merito alla necessità di armonizzare i conti pubblici e assicurare la produzione di dati di bilancio e di informazioni contabili di qualità tali da garantire la comparabilità tra gli Stati membri dell'Unione Europea. Di qui la fissazione di regole minime comuni per i quadri di bilancio nazionali finalizzate a renderli più trasparenti, confrontabili e il più possibile completi e veritieri, evidenziando la necessità di coordinamento tra tutti i sotto-settori delle amministrazioni e di uniformare le regole e le procedure contabili.

Il legislatore italiano introdusse la materia dell''Armonizzazione dei bilanci pubblici' nella Carta costituzionale con la riforma del titolo V del 2001, dapprima inserendola nell'ambito della competenza concorrente, e poi, con la Legge costituzionale n. 1 del 2012, trasformandola in competenza esclusiva statale, introducendo il principio del pareggio di bilancio, coerentemente con le regole di governance economica europea in vigore. Il processo di armonizzazione nazionale proseguì con l'approvazione della Legge Delega n. 42 del 2009 sul federalismo fiscale e della Legge di contabilità e finanza pubblica n. 196 del 2009, che intendeva garantire la gestione unitaria e coordinata della finanza pubblica e, per il suo tramite, l'unità economica della Repubblica (articoli 117 e 120 della Costituzione), favorire l'attuazione del federalismo fiscale e consentire l'aderenza dell'ordinamento contabile pubblico italiano alla normativa comunitaria. In particolare, la Legge n. 196 del 2009 formalizzò la classificazione del bilancio dello Stato che è ora articolata su tre livelli di aggregazione – le Missioni, i Programmi e le Azioni – al fine di favorire una migliore comprensione delle scelte allocative in relazione alle principali politiche pubbliche da perseguire attraverso la spesa.

Infine, furono emanati il D.lgs. 91/11 per le amministrazioni pubbliche in generale e il D.lgs. 118/11 (successivamente modificato e integrato dal D.lgs. 126/14) per le amministrazioni pubbliche territoriali, quali regioni, enti locali ed enti del Servizio sanitario nazionale. Per le Università, la scelta fatta dal legislatore fu quella di individuare criteri e norme specifiche, che hanno trovato collocazione nella Legge n. 240/10 e nel successivo D.lgs. 18/12.

Il D.lgs. 118/11 introdusse, tra le altre novità, regole contabili uniformi e un comune piano dei conti integrato per consentire il consolidamento e il monitoraggio in fase di previsione, gestione e rendicontazione, ma anche schemi comuni di bilancio articolati, sul lato della spesa, in Missioni e Programmi, con le relative azioni, in armonia con la classifi-

cazione economica e funzionale individuata dagli appositi Regolamenti e Direttive comunitari in materia, in simmetria con quanto già avveniva per il bilancio dello Stato.

Se le Missioni sono definite in base al riparto di competenze stabilito dagli articoli 117 e 118 della Costituzione, i Programmi rappresentano gli aggregati omogenei di attività volte a perseguire gli obiettivi istituzionali definiti nell'ambito delle Missioni, e la loro denominazione riflette le principali aree di intervento delle Missioni di riferimento, consentendo una rappresentazione di bilancio omogenea per tutti gli enti pubblici. Nello specifico, la ripartizione dei Programmi è raccordata allo standard di classificazione internazionale di secondo livello adottato dall'Unione Europea, denominato *Classification Of the Functions Of Government* (COFOG)[5], secondo le corrispondenze individuate nel *Glossario delle Missioni e dei Programmi*, che costituisce una guida per la classificazione delle spese (Allegato n. 14 del D.lgs. 118/11).

In questo modo, le spese sono classificate secondo criteri omogenei, allo scopo di assicurare maggiore trasparenza delle informazioni riguar-

[5] La *Classification of the Functions of Government* (COFOG) è uno standard di classificazione internazionale della spesa delle Amministrazioni pubbliche per funzione, attraverso cui si aspira ad una valutazione omogenea dell'attività economica svolta dalle pubbliche amministrazioni anche dei Paesi UE. La classificazione COFOG, utilizzata quale *benchmark* di riferimento per i Programmi del bilancio, prevede tre successivi livelli di analisi e permette di classificare tutte le voci di spesa delle amministrazioni pubbliche. Sono previste dieci Divisioni (funzioni di 1° livello), analizzate al loro interno in Gruppi (funzioni di 2° livello) e successivamente in Classi (funzioni di 3° livello), per consentire, tra l'altro, una valutazione omogenea delle attività delle pubbliche amministrazioni svolte dai diversi paesi. Le Divisioni rappresentano i fini primari perseguiti dalle amministrazioni, i Gruppi riguardano le specifiche aree di intervento delle politiche pubbliche e le Classi identificano i singoli obiettivi in cui si articolano le aree di intervento. L'adozione della COFOG per tutte le amministrazioni pubbliche, disposta in Italia dall'art. 2, comma 1, lett. c) della Legge n. 196/2009, mira a garantire il consolidamento della spesa attraverso un riferimento uniforme e comune. La classificazione per Missioni e Programmi, vista invece la possibilità delle amministrazioni di modificare la struttura per Programmi (per rappresentare specifiche e contingenti politiche di spesa), non necessariamente assicura uno schema standard fisso per aggregare le spese delle diverse amministrazioni. Da ciò deriva che la flessibilità insita nella classificazione per Missioni e Programmi costituisce un parziale limite in una prospettiva di consolidamento della spesa mentre, d'alta parte, la rigidità della COFOG (stabilita a livello internazionale e adottata in Europa con il Sistema europeo dei conti nazionali e regionali (Sec 2010), insieme alle altre classificazioni funzionali), costituisce un vantaggio per i confronti non solo internazionali ma anche tra gli enti all'interno dei confini nazionali. Pertanto, la classificazione COFOG rappresenta lo strumento che permette di far colloquiare le diverse rappresentazioni della spesa per Missioni e Programmi. La spesa delle amministrazioni pubbliche secondo la COFOG, al pari delle entrate pubbliche, viene registrata in base al criterio della competenza economica adottato dal sistema dei conti nazionali.

danti la destinazione delle risorse pubbliche, agevolare la 'lettura' secondo la finalità di spesa, consentire pertanto la più ampia comparabilità dei dati di bilancio e permetterne l'aggregazione.[6]

Ai fini del presente capitolo, sono state prese in considerazione, per i bilanci delle amministrazioni regionali e locali, 19 Missioni, con i relativi Programmi e Gruppi COFOG di riferimento, che rappresentano le funzioni principali e gli obiettivi strategici perseguiti dalle amministrazioni e corrispondono a specifiche Missioni nel bilancio dello Stato. Per l'applicazione della matrice al caso statale, le Missioni previste a livello statale dalla Legge 31 dicembre 2009, n. 196 "Legge di contabilità e finanza pubblica" e sue successive modificazioni e integrazioni sono state associate con quelle previste dal D.lgs. 118/11. La Tab. 1 riporta le Missioni che saranno utilizzate nel seguito del lavoro per tutti i livelli di governo.

Tab. 1 Missioni nei bilanci delle amministrazioni regionali e locali.

Missioni nei bilanci delle amministrazioni regionali e locali	
Missione 01	"Servizi istituzionali, generali e di gestione"
Missione 02	"Giustizia"
Missione 03	"Ordine pubblico e sicurezza"
Missione 04	"Istruzione e diritto allo studio"
Missione 05	"Tutela e valorizzazione dei beni e delle attività culturali"
Missione 06	"Politiche giovanili, sport e tempo libero"
Missione 07	"Turismo"
Missione 08	"Assetto del territorio ed edilizia abitativa"
Missione 09	"Sviluppo sostenibile e tutela del territorio e dell'ambiente"

[6] Lo scambio delle informazioni tra i singoli Stati membri dell'UE si basa sul Sistema europeo dei conti nazionali e regionali (Sec 95) dell'Unione Europea, costituito da un insieme di principi, definizioni, classificazioni e regole contabili comuni a cui i diversi Paesi membri devono attenersi ai fini dell'elaborazione dei conti da inviare alla Commissione Europea, dando vita ad un sistema contabile comparabile, che permette di descrivere in maniera sistematica e dettagliata il complesso di un'economia, le sue componenti e le sue relazioni con altre economie. Nel corso del 2014 è stato adottato in Italia, in coerenza con quanto avvenuto da parte di tutti gli Stati membri dell'Unione Europea, il nuovo sistema Sec 2010 dei conti economici nazionali e regionali. Il Sec 2010 (come già avvenuto per il precedente sistema Sec 95, Regolamento CE del Consiglio, n. 2223/1996) è stato adottato in Europa con il Regolamento (UE) del Parlamento europeo e del Consiglio, n. 549/2013, fissando il complesso dei principi e delle metodologie da applicare nella costruzione dei conti economici nazionali, anche attraverso specifiche regole per i diversi settori istituzionali in cui è articolata l'economia degli Stati membri dell'Unione Europea.

Missione 10	"Trasporti e diritto alla mobilità"
Missione 11	"Soccorso civile"
Missione 12	"Diritti sociali, politiche sociali e famiglia"
Missione 13	"Tutela della salute"
Missione 14	"Sviluppo economico e competitività"
Missione 15	"Politiche per il lavoro e la formazione professionale"
Missione 16	"Agricoltura, politiche agroalimentari e pesca"
Missione 17	"Energia e diversificazione delle fonti energetiche"
Missione 18	"Relazioni con le altre autonomie territoriali e locali"
Missione 19	"Relazioni internazionali"

Un nuovo modello per lo sviluppo sostenibile

La matrice per la sostenibilità tra le categorie della spesa e gli Obiettivi di Agenda 2030

La 'matrice per la sostenibilità' è stata realizzata tramite un'analisi documentale ed una *content analysis* di tipo quali-quantitativo, con l'utilizzo del software Iramuteq.

Per costruire la matrice, sono stati innanzitutto selezionati i testi da utilizzare per comporre i vari corpora da sottoporre ad analisi. Sono stati preparati 17 corpora, uno per ciascun *SDGs*[7] dell'Agenda 2030, contenenti la descrizione dei 17 Obiettivi, dei relativi 169 *Target* e dei 273 Indicatori associati[8]. Successivamente, i medesimi testi sono stati utilizzati per l'individuazione di alcune parole o espressioni chiave particolarmente rappresentative dei singoli argomenti in quanto significativamente frequenti nei corpora, utilizzando Iramuteq come *word counter*. A titolo esemplificativo, le parole o espressioni chiave individuate per il *Goal* 8 'Lavoro dignitoso e crescita economica' sono: "sviluppo economico", "attività produttive", "occupati", "disoccupati", "disoccupazione". Nei casi di parole o espressioni con più significati, le stesse sono state sostituite con sinonimi o parole composte. Sono esempi la parola "sviluppo" (utilizzata nelle

[7] I testi degli Obiettivi e *Target* sono stati estratti dalla Risoluzione 70/1 intitolata "Trasformare il nostro mondo: l'Agenda 2030 per lo sviluppo sostenibile" adottata dall'Assemblea Generale delle Nazioni Unite il 25 settembre 2015.

[8] Si veda il Rapporto *SDGs* 2019 dell'Istat.

diverse espressioni "paesi in via di sviluppo", "cooperazione allo sviluppo", "sviluppo economico"), la parola "ordine" (utilizzata nelle espressioni "forze dell'ordine", "ordine pubblico") e la parola "sociale" (utilizzata nelle diverse espressioni "esclusione sociale", "protezione sociale").

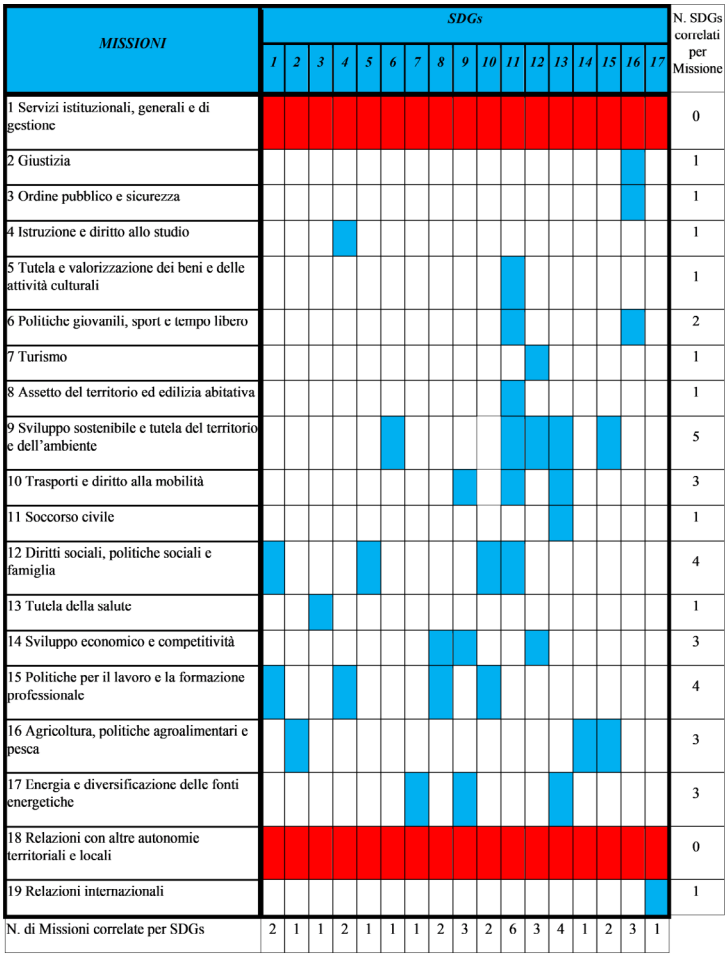
Infine, è stato preparato l'ultimo corpus relativo alle categorie di spesa, contenente la descrizione delle 19 Missioni e dei corrispondenti Programmi e Gruppi COFOG come riportati nell'Allegato 14 del D.lgs 118/11.

Tramite una ricerca dei valori di associazione del χ^2 delle parole o espressioni chiave precedentemente individuate con le singole Missioni, eseguita dal software, è stato quindi possibile individuare le correlazioni tra le categorie di spesa e gli *SDGs* dell'Agenda 2030.

Il risultato finale dell'analisi statistica realizzata da Iramuteq è rappresentato da una matrice di correlazione (Fig. 2), dove ciascuna categoria di spesa (Missione) è associata a ciascun *SDGs*. Nella matrice, ogni cella evidenziata in blu raffigura una forte associazione tra le variabili *SDGs* e le Missioni, e le celle rosse un'associazione inversa, ovvero una sottorappresentazione dei temi delle variabili considerate.

Dalla Fig. 2 emerge come l'Obiettivo 11 'Rendere le città e gli insediamenti umani inclusivi, sicuri, duraturi e sostenibili' risulti associato a ben 6 Missioni, seguito dall'Obiettivo 13 'Adottare misure urgenti per combattere il cambiamento climatico e le sue conseguenze', con 4 Missioni correlate. Entrambi gli Obiettivi sono molto complessi, con *target* ed indicatori differenziati volti ad affrontare tematiche diverse. Analizzando invece il numero di *SDGs* correlati per Missione, si evidenzia come sia la Missione 9 'Sviluppo sostenibile e tutela del territorio e dell'ambiente' ad essere correlata con il maggior numero di Obiettivi (5 in totale). Questo risultato ci sembra coerente, considerato che i temi dello sviluppo sostenibile e dell'ambiente rappresentano due capisaldi degli *SDGs* di Agenda 2030. Al contrario, le Missioni 1 e 18 non sono correlate con gli *SDGs*. Si tratta, infatti, di Missioni trasversali e non ripartibili. Nello specifico, la Missione 1 ricomprende spese per il funzionamento dei servizi generali, dei servizi statistici e informativi, delle attività per lo sviluppo dell'ente in un'ottica di governance e partenariato e per la comunicazione istituzionale. La Missione 18 riguarda invece erogazioni ad altre amministrazioni territoriali e locali di finanziamenti non riconducibili a specifiche Missioni.

Fig. 2 La matrice per la sostenibilità tra le Missioni/Programmi della spesa e gli Obiettivi di Agenda 2030.



MISSIONI	SDGs 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	N. SDGs correlati per Missione
1 Servizi istituzionali, generali e di gestione	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0
2 Giustizia																■		1
3 Ordine pubblico e sicurezza																■		1
4 Istruzione e diritto allo studio				■														1
5 Tutela e valorizzazione dei beni e delle attività culturali											■							1
6 Politiche giovanili, sport e tempo libero											■					■		2
7 Turismo												■						1
8 Assetto del territorio ed edilizia abitativa											■							1
9 Sviluppo sostenibile e tutela del territorio e dell'ambiente						■					■	■	■		■			5
10 Trasporti e diritto alla mobilità									■		■		■					3
11 Soccorso civile													■					1
12 Diritti sociali, politiche sociali e famiglia	■				■					■	■							4
13 Tutela della salute			■															1
14 Sviluppo economico e competitività								■	■			■						3
15 Politiche per il lavoro e la formazione professionale	■			■				■		■								4
16 Agricoltura, politiche agroalimentari e pesca		■												■	■			3
17 Energia e diversificazione delle fonti energetiche							■		■				■					3
18 Relazioni con altre autonomie territoriali e locali	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	0
19 Relazioni internazionali																	■	1
N. di Missioni correlate per SDGs	2	1	1	2	1	1	1	2	3	2	6	3	4	1	2	3	1	

Il caso della Regione Toscana

Allo scopo di sperimentare l'applicazione del modello ad un caso concreto, la 'matrice per la sostenibilità' è stata applicata al caso della Regione Toscana, prendendo a riferimento i dati di previsione della spesa in termini di competenza per il periodo di programmazione relativo al PRS 2016-2020 e, nello specifico, per gli anni 2016, 2017, 2018 e 2019, in modo tale da analizzare e valutare la modalità di distribuzione della spesa

rispetto agli Obiettivi di Agenda 2030 per quasi l'intera durata[9] della X legislatura regionale toscana.

Successivamente, è stato operato un confronto tra la distribuzione della spesa rispetto agli *SDGs* di Regione Toscana e altre categorie di enti pubblici a diversi livelli di governo, ritenuti significativi per i nostri obiettivi di ricerca. Nello specifico, la matrice è stata applicata ai dati di previsione della spesa in termini di competenza per lo stesso arco temporale del 'Comune di Firenze', dello 'Stato italiano', e delle 'regioni italiane' in quanto insieme.

Per spiegare queste scelte, riteniamo fondamentale ricordare, innanzitutto, il quadro normativo che prevede il riparto delle competenze tra i diversi livelli di governo, così come stabilito dagli articoli 117 e 118 della Costituzione, in base alle quali sono definite le Missioni di spesa che rappresentano le funzioni principali e gli obiettivi strategici perseguiti dalle amministrazioni.

In particolare, il Titolo V della Costituzione, alla luce della Legge costituzionale n. 3 del 2001, ha profondamente modificato il regime di riparto delle competenze e funzioni fra Stato e regioni. Alle regioni spetta una competenza legislativa generale, di tipo residuale, per tutte quelle materie non specificamente riservate allo Stato dal secondo comma dell'art. 117, una competenza concorrente su altre specifiche materie di cui al terzo comma dello stesso articolo, e una competenza amministrativa su materie disciplinate dall'art. 118, attribuite in base al principio di sussidiarietà, differenziazione e adeguatezza.

Le diverse sfere di competenza attribuite a ciascun livello di governo del territorio riguardano tutti gli ambiti della sostenibilità (economico, sociale ed ambientale), rendendo particolarmente interessante l'analisi dei dati di spesa appartenenti a livelli territoriali differenti, e consentendo la verifica del livello di correlazione e distribuzione della spesa tra gli *SDGs* rispetto alla distribuzione delle competenze tra le diverse tipologie di enti territoriali, configurando un primo tentativo di trasferimento della matrice, anche al fine di controllare a diversi livelli la validità della correlazione tra le classi di spesa e gli *SDGs*.

A questo punto ci sembra opportuno offrire brevemente il quadro delle previsioni di spesa della Regione Toscana, aggregata per Missioni/Programmi, prendendo come riferimento gli anni 2016-2019. Prima occorre

[9] Per avere una visione completa di legislatura mancano i dati dell'anno 2020, non ancora disponibili al momento in cui sono state eseguite le analisi (settembre 2019).

tuttavia precisare che, per l'applicazione della matrice al caso concreto, i dati della spesa delle Missioni 1 e 18 non sono stati presi in considerazione in quanto correlate con gli *SDGs*.

La spesa regionale nel periodo 2016-2019, al netto della componente passiva di amministrazione, delle Missioni 20, 50, 60 e 99, e delle Missioni trasversali 1 e 18, è pari complessivamente a 35.778,33 milioni di Euro, di cui l'82,12% (29.380,72 milioni di Euro) è destinata alla Missione 13 'Tutela della salute'. Proprio per la grande rilevanza della spesa destinata dalle regioni alle politiche a tutela della salute, come sarà chiarito più approfonditamente nel seguito, abbiamo preferito trattare la Missione 13 separatamente rispetto alle altre. La spesa della Regione Toscana destinata alle Missioni/Programmi, con l'esclusione della Missione 13, per gli anni 2016-2019 è di 6.397,61 milioni di Euro.

La matrice per la sostenibilità della Regione Toscana

L'applicazione del modello della 'matrice per la sostenibilità' alla spesa della Regione Toscana (Fig. 3) evidenzia chiaramente come gli Obiettivi dell'Agenda 2030 cui sono destinate maggiori risorse siano i *Goal* 9 'Imprese, innovazione e infrastrutture' e 11 'Città e comunità sostenibili', rispettivamente con il 58,67% e 58,49% della spesa, seguiti dal *Goal* 13 'Lotta contro il cambiamento climatico' con il 53,99%. Viceversa, gli Obiettivi con la più bassa percentuale di risorse assegnate sono il *Goal* 14 'Vita sott'acqua' e 15 'Vita sulla terra', rispettivamente con lo 0,42% e lo 0,47%, seguiti dai *Goal* 5 'Parità di genere' e 6 'Acqua pulita e servizi igienico-sanitari', cui sono destinati rispettivamente lo 0,96% e l'1% della spesa regionale.

Guardando invece alla distribuzione della spesa per Missione, si evidenzia che le Missioni cui è destinata maggiore spesa sono la 10 'Trasporti e diritto alla mobilità', con il 45,15% e la 14 'Sviluppo economico e competitività', con l'11,32%. Alla Missione 2 'Giustizia' non è destinata alcuna risorsa, mentre alla 3 'Ordine pubblico e sicurezza' lo 0,17% e alla 6 'Politiche giovanili, sport e tempo libero' lo 0,3%.

Come anticipato nel paragrafo precedente, il calcolo delle percentuali di distribuzione della spesa per *SDGs* è stato effettuato prendendo a riferimento il totale della spesa correlato agli *SDGs*, con l'esclusione della Missione 13 'Tutela della salute'. La percentuale di spesa destinata alla tutela della salute, in quanto rappresenta per l'ente Regione una Missione di spesa eccessivamente pesante rispetto alle altre Missioni, è stata calcolata separatamente prendendo come riferimento il totale della spesa

correlato agli *SDGs*.

Fig. 3 Matrice per la sostenibilità della Regione Toscana. Anni 2016-2019.

REGIONE TOSCANA

MISSIONI	*SDGs*																
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
1 Servizi istituzionali, generali e di gestione																	
2 Giustizia																0	
3 Ordine pubblico e sicurezza																0,17	
4 Istruzione e diritto allo studio				5,43													
5 Tutela e valorizzazione dei beni e delle attività culturali											2,51						
6 Politiche giovanili, sport e tempo libero											0,30					0,30	
7 Turismo												1,45					
8 Assetto del territorio ed edilizia abitativa											1,13						
9 Sviluppo sostenibile e tutela del territorio e dell'ambiente						0,96					8,02	8,02	5,75		6,36		
10 Trasporti e diritto alla mobilità									45,15		45,15		45,15				
11 Soccorso civile													0,65				
12 Diritti sociali, politiche sociali e famiglia	4,97				1					4,97	1,38						
13 Tutela della salute (*)			82,12														
14 Sviluppo economico e competitività								11,32	11,08			11,32					
15 Politiche per il lavoro e la formazione professionale	9,38			6,49				9,38		9,38							
16 Agricoltura, politiche agroalimentari e pesca		3,86												0,42	3,86		
17 Energia e diversificazione delle fonti energetiche							2,44		2,44				2,44				
18 Relazioni con altre autonomie territoriali e locali																	
19 Relazioni internazionali																	3,22
Totale spesa per SDGs	14,35	3,86	82,12	11,92	1	0,96	2,44	20,70	58,67	14,35	58,49	20,79	53,99	0,42	10,22	0,47	3,22

(*): la spesa destinata alla tutela della salute rappresenta la maggiore Missione di spesa dell'ente Regione, per questo si è preferito trattare la Missione 13 a

parte. Quindi la percentuale di spesa dedicata alla tutela della salute (82,12%) è stata calcolata sul totale della spesa correlata agli *SDGs*, mentre la percentuale di spesa per tutte le altre Missioni è stata calcolata sottraendo dal totale della spesa quella per la Missione 13.

Confronto tra livelli di governo

La Fig. 4 mostra la distribuzione della spesa per *Goal* ai vari livelli di governo, raffrontando i dati risultanti dalle singole matrici elaborate per la Regione Toscana, l'insieme delle regioni italiane, il Comune di Firenze e lo Stato italiano. Se la matrice della Regione Toscana è stata presentata nel paragrafo precedente, le altre tre sono riportate nell'Appendice 2 al presente volume.

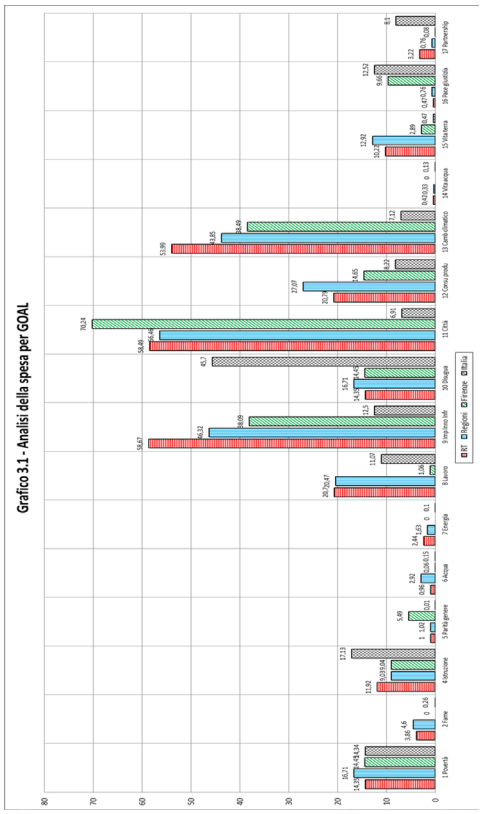
La distribuzione della spesa per l'insieme delle regioni è simile a quella della Regione Toscana. Gli Obiettivi cui è destinata la percentuale maggiore di spesa sono il 9 'Imprese, innovazione e infrastrutture', l'11 'Città e comunità sostenibili' e il 13 'Lotta contro il cambiamento climatico', seppure con proporzioni diverse. Inoltre, rispetto agli altri soggetti l'ente Regione è quello che destina la maggiore percentuale di spesa per il *Goal* 8 'Lavoro dignitoso e crescita economica'. I *Goal* cui è attribuita la minore percentuale di spesa sono invece il 5 'Parità di genere', il 14 'Vita sott'acqua' e il 16 'Pace, giustizia e istituzioni solide'.

Anche il Comune di Firenze destina la maggiore percentuale di spesa ai medesimi *Goal* dell'ente Regione (9, 11 e 13), con un picco di spesa del 70,24% per il *Goal* 11. Il Comune di Firenze si distingue inoltre per la destinazione di una maggiore percentuale di risorse rispetto agli altri soggetti al *Goal* 5 'Parità di genere'. Viceversa, i *Goal* con minore spesa sono il 14 'Vita sott'acqua', il 2 'Sconfiggere la fame' e il 7 'Energia pulita e accessibile'.

Per l'Italia, la modalità di distribuzione delle risorse cambia completamente ed i *Goal* con maggiore spesa risultano essere il 10 'Ridurre le disuguaglianze', il 4 'Istruzione di qualità' e l'1 'Sconfiggere la povertà'. Rispetto agli altri soggetti considerati, è l'Italia a destinare la maggior percentuale delle risorse alle Missioni 16 'Pace, giustizia e istituzioni solide' e 17 'Partnership per gli obiettivi'. I *Goal* con minore spesa sono il 5 'Parità di genere', il 7 'Energia pulita e accessibile' e il 14 'Vita sott'acqua'.

In definitiva, e come sarà meglio chiarito nel paragrafo successivo, la distribuzione della spesa tra gli *SDGs* ai vari livelli di governo risulta fortemente influenzata dal riparto delle competenze.

Fig. 4 Analisi della spesa per *Goal*.

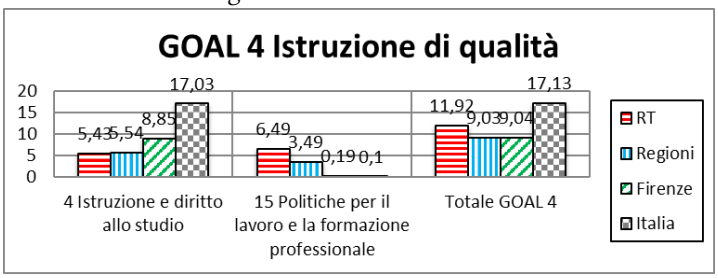


L'analisi degli Obiettivi di sviluppo sostenibile

In questa sezione è presentata l'analisi di alcuni degli Obiettivi per i quali ci è sembrata più significativa la correlazione con le Missioni di spesa e con le diverse competenze ai vari livelli di governo.

Il *Goal* 4 'Istruzione di qualità' (Fig. 5), teso a fornire un'educazione di qualità, equa ed inclusiva, e opportunità di apprendimento per tutti, in base alla matrice per la sostenibilità risulta associato alla Missione 4 'Istruzione e diritto allo studio' e ai relativi Programmi e alla Missione 15 'Politiche per il lavoro e la formazione professionale' e al solo Programma 1502 'Formazione professionale' (non sono quindi correlati i Programmi relativi alle politiche occupazionali e per il mercato del lavoro).

Fig. 5 *Goal* 4 Istruzione di qualità: distribuzione della spesa per Missione e per livello di governo territoriale. Anni 2016-2019.



L'Italia è il soggetto con la maggiore percentuale di spesa dedicata a quest'Obiettivo, pari al 17,13% e derivante quasi esclusivamente dalla Missione 4, di cui il 12,29% per i Programmi riferiti all'istruzione secondaria di primo e secondo grado. La mancanza di risorse per la formazione professionale è attribuibile al fatto che quest'ultima è una materia di competenza regionale.

La Regione Toscana attribuisce all'Obiettivo 4 l'11,92% delle proprie risorse, quasi equamente distribuite tra la Missione 4 (principalmente con i Programmi 404 'Istruzione universitaria' e 402 'Altri ordini di istruzione non universitaria'), e la Missione 15 (con il Programma 1502 'Formazione professionale'). L'insieme delle regioni dedicano a quest'Obiettivo circa il 3% in meno rispetto alla Regione Toscana, per via di una minore percentuale di risorse dedicate al Programma 1502.

Per il Comune di Firenze, invece, la quasi totalità della spesa, pari al 9,04%, destinata all'Obiettivo 4 deriva dai Programmi 406 'Servizi ausiliari all'istruzione', 402 'Altri ordini di istruzione non universitaria' e 401 'Istruzione prescolastica' della Missione 4 'Istruzione e diritto allo studio'.

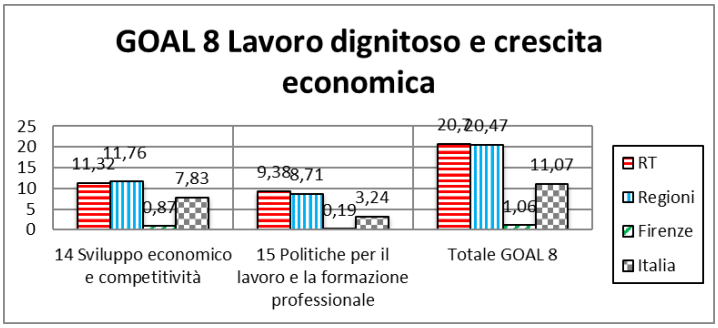
Il *Goal* 8 'Lavoro dignitoso e crescita economica' (Fig. 6) risulta, secondo la matrice per la sostenibilità, maggiormente correlato alla Missione 14 – che vuole incentivare una crescita economica duratura, inclusiva e sostenibile, con politiche a favore dell'industria, delle PMI e dell'artigianato, del commercio, della ricerca e innovazione – e 15, per un'occupazione piena e produttiva ed un lavoro dignitoso per tutti.

Con l'attribuzione di circa il 20% della spesa totale, distribuita equamente tra le politiche di sviluppo economico e competitività da una parte e le politiche del lavoro e formazione professionale dall'altra, è l'ente Regione a destinare la maggior parte delle risorse a quest'Obiettivo. Di queste risorse, più della metà derivano dai Programmi riservati agli enti regionali dedicati alla Politica regionale unitaria.

Segue l'Italia, con l'11,07% della spesa, di cui il 7,83% relativo alle politiche per lo sviluppo economico e competitività, costituite dalla Missione dello Stato 11 'Competitività e sviluppo delle imprese', la Missione 12 'Regolazione dei mercati', la Missione 16 'Commercio internazionale e internazionalizzazione del sistema produttivo' e la Missione 17 'Ricerca e innovazione', ed il 3,24% alle politiche per il lavoro.

Solo l'1,06% delle risorse è destinato a quest'Obiettivo dal Comune di Firenze.

Fig. 6 *Goal* 8 Lavoro dignitoso e crescita economica: distribuzione della spesa per Missione e per livello di governo territoriale. Anni 2016-2019.



Il *Goal* 9 'Imprese, Innovazione e infrastrutture' (Fig. 7) ha l'obiettivo di costruire un'infrastruttura resiliente e promuovere l'innovazione ed una industrializzazione equa, responsabile e sostenibile. Dalla matrice per la sostenibilità è emersa una correlazione con la Missione 10 'Trasporti e diritto alla mobilità', la Missione 14 'Sviluppo economico e competitività', la Missione 17 'Energia e diversificazione delle fonti energetiche' e tutti i rispettivi Programmi associati.

Questo è l'Obiettivo a cui la Regione Toscana destina la maggiore percentuale delle proprie risorse, con il 58,67%. Buona parte di queste risorse derivano dalla Missione 10 con il 45,15%, di cui il 17,87% dal Programma 1001 'Trasporto ferroviario' ed il 20,18% dal Programma 1002 'Trasporto pubblico locale'. Il resto delle risorse proviene dalla Missione 14 (ad esclusione del Programma 1402 'Commercio - reti distributive - tutela dei consumatori' che non risulta correlato con quest'Obiettivo), con l'11,08% e dalla Missione 17 con il 2,44%.

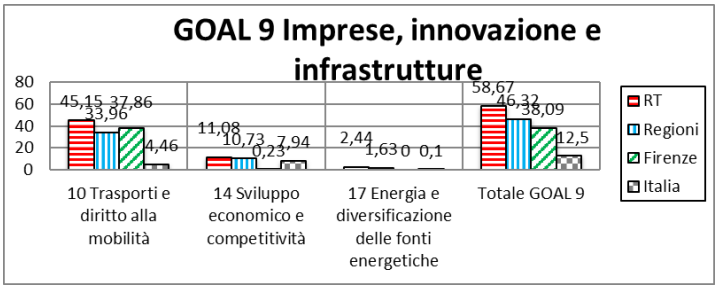
Segue l'insieme delle regioni con il 46,32% delle risorse, la cui differenza di circa il 12% rispetto al dato della Regione Toscana è da attribuire principalmente alla minor distribuzione della spesa nella Missione 10

ed in particolar modo nei Programmi destinati al trasporto ferroviario e al trasporto pubblico locale.

Il Comune di Firenze investe in quest'Obiettivo, il 38,09% delle risorse, da attribuire quasi esclusivamente alle politiche relative al trasporto pubblico locale e gli interventi sulla viabilità e infrastrutture stradali. A riguardo si ricordano, tra l'altro, gli importanti investimenti degli ultimi anni per la realizzazione del sistema tramviario fiorentino.

Infine, l'Italia riserva all'Obiettivo il 12,5% della spesa, di cui il 4,46% delle risorse derivanti dalla Missione dello Stato 13 'Diritto alla mobilità e sviluppo dei sistemi di trasporto' e dal Programma 1411 'Sistemi stradali, autostradali ed intermodali' della Missione 14 'Infrastrutture pubbliche e logistica', ed il 7,94% dalle politiche per lo sviluppo economico e competitività.

Fig. 7 *Goal* 9 Imprese, innovazione e infrastrutture: distribuzione della spesa per Missione e per livello di governo territoriale. Anni 2016-2019.

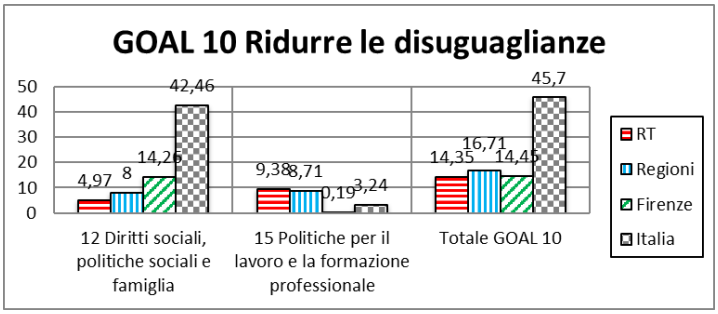


Con riguardo al *Goal* 10 'Ridurre le disuguaglianze' (Fig. 8), l'Italia assegna la maggior parte delle proprie risorse, pari al 45,7%, al raggiungimento di tale Obiettivo. La maggior parte di queste risorse derivano dalle politiche a sostegno dei diritti sociali e delle famiglie, afferenti alla Missione dello Stato 24 'Diritti sociali, politiche sociali e famiglia' (11,11%), alla Missione 25 'Politiche previdenziali' (29,33%), a parte della Missione 27 'Immigrazione, accoglienza e garanzia dei diritti' (0,65%), ad esclusione dei rapporti con le confessioni religiose, e alla Missione 28 'Sviluppo e riequilibrio territoriale' (1,38%). Il restante 3,24% deriva dalle politiche per il lavoro.

Gli altri soggetti attribuiscono una percentuale di spesa al raggiungimento di quest'Obiettivo che varia dal 14% del Comune di Firenze (con una spesa che si colloca quasi totalmente nella Missione 12) e Regione Toscana (con una spesa pari al 9,38% per le politiche per il lavoro

e al 4,97% per le politiche sociali), al 16% dell'insieme delle regioni, distribuito equamente tra la Missione 12 e 15.

Fig. 8 *Goal* 10 Ridurre le disuguaglianze: distribuzione della spesa per Missione e per livello di governo territoriale. Anni 2016-2019.

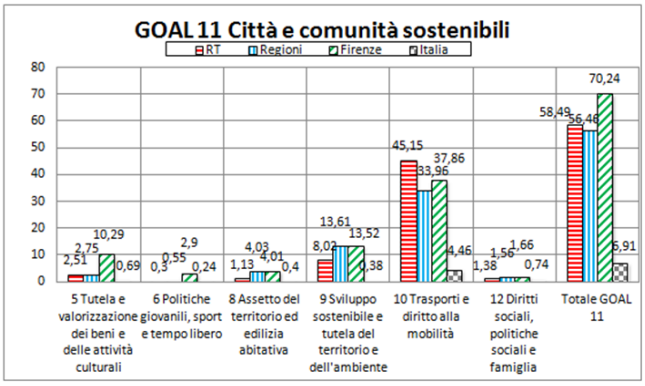


Il *Goal* 11 'Città e comunità sostenibili' (Fig. 9) è teso a rendere le città e gli insediamenti umani inclusivi, sicuri, duraturi e sostenibili. Quest'Obiettivo, per la sua complessità tematica e varietà dei *target* che include, è correlato con ben 6 Missioni: le Missioni 5, 6, 8, 9 e 10 con tutti i rispettivi Programmi, e la Missione 12 con i Programmi 1204 'Interventi per i soggetti a rischio di esclusione sociale' e 1206 'Interventi per il diritto alla casa'.

Il Comune di Firenze distribuisce la maggior parte della propria spesa, con il 70,24%, per il raggiungimento del *Goal*. Segue la Regione Toscana con il 58,49% e l'insieme delle regioni con il 56,46%, infine l'Italia con il 6,91%.

Ci sembra in effetti corretto riscontrare che sia il Comune a destinare la maggiore percentuale di spesa ad un Obiettivo che si colloca vicino al cittadino, avendo ad oggetto proprio le città, seguito dalle Regioni e, a distanza, dall'Italia. Le politiche su cui sono maggiormente distribuite le risorse sono quelle legate ai trasporti, seguite dallo sviluppo economico. Il Comune, inoltre, rispetto agli altri soggetti, investe molto di più per le politiche culturali (10,29%) e giovanili (2,9%).

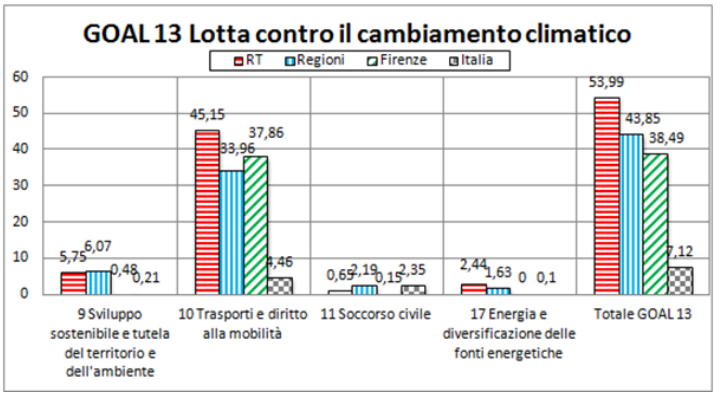
Fig. 9 *Goal* 11 Città e comunità sostenibili: distribuzione della spesa per Missione e per livello di governo territoriale. Anni 2016-2019.



Infine, per quanto riguarda il *Goal* 13 'Lotta contro il cambiamento climatico' (Fig. 10), in questo caso è la Regione Toscana in prima fila tra i soggetti considerati nell'analisi – con il 53,99% delle proprie risorse – nell'adottare misure urgenti per combattere il cambiamento climatico e le sue conseguenze, come i disastri naturali. Segue l'insieme delle regioni con il 43,85% e il Comune di Firenze con il 38,49%. Infine, l'Italia con il 7,12%.

Oltre alle politiche relative ai trasporti e alla mobilità, risultano correlate con quest'Obiettivo le politiche ambientali a difesa del suolo, delle aree protette e forestazione e per la riduzione dell'inquinamento atmosferico, le politiche energetiche ed anche quelle per la protezione civile e per gli interventi a seguito di calamità naturali.

Fig. 10 *Goal* 13 Lotta contro il cambiamento climatico: distribuzione della spesa per Missioni e per livello di governo territoriale. Anni 2016-2019.



Conclusioni e possibili sviluppi della ricerca

Nel corso del capitolo è stata presentata la realizzazione tramite tecniche di text mining e successiva applicazione ad alcuni casi concreti di una matrice di correlazione tra le modalità di distribuzione della spesa pubblica e gli Obiettivi di sviluppo sostenibile definiti dall'Agenda 2030 delle Nazioni Unite, a partire dall'ipotesi di una sinergia tra gli standard di classificazione della spesa pubblica in Missioni/Programmi, a cui viene associata la classificazione COFOG, e gli *SDGs* dell'Agenda 2030, e dalla volontà di rendere trasparente la sostenibilità della spesa pubblica, alla luce dei processi di armonizzazione dei bilanci intervenuti negli anni e delle esigenze di allocazione delle risorse derivanti dal particolare contesto storico.

La 'matrice per la sostenibilità' vuole essere uno strumento per incrementare la base conoscitiva di un ente pubblico che cerca di misurare e valutare il proprio posizionamento rispetto agli *SDGs* e che vuole programmare in modo consapevole il proprio futuro sostenibile.

Dalla realizzazione della matrice e dalla sua applicazione pratica al caso della Regione Toscana e ad altri livelli di governo è emerso come quasi tutte le Missioni di spesa, ad eccezione della Missione 1 'Servizi istituzionali, generali e di gestione' e della Missione 18 'Relazioni con altre autonomie territoriali e locali' – con l'esclusione delle Missioni trasversali o non ripartibili o comunque dei servizi forniti in maniera indivisibile – siano correlate con gli *SDGs*. L'Obiettivo associato al maggior numero di Missioni è l'11 'Rendere le città e gli insediamenti umani inclusivi, sicuri, duraturi e sostenibili', con ben 6 Missioni associate. Al contrario, la Missione associata al maggior numero di Obiettivi è la 9 'Sviluppo sostenibile e tutela del territorio e dell'ambiente', con ben 5 Obiettivi associati.

In particolare, la distribuzione della spesa della Regione Toscana e dell'insieme delle regioni sono molo simili tra loro, con una destinazione di spesa maggiore in entrambi i casi nei confronti dei *Goal* 9 'Imprese, innovazione e infrastrutture', 11 'Città e comunità sostenibili' e 13 'Lotta contro il cambiamento climatico'. Per il Comune di Firenze la spesa risulta più accentuata in direzione del *Goal* 11, e in misura minore verso i *Goal* 13 e 9. Discorso a parte merita l'Italia, con una spesa decisamente accentuata per il *Goal* 10 'Ridurre le disuguaglianze', ma anche nei confronti dei *Goal* 4 'Istruzione di qualità', 16 'Pace, giustizia e istituzioni solide' e 17 'Partnership per gli obiettivi'.

Focalizzando l'attenzione sull'ente Regione Toscana, la distribuzione della spesa – con l'esclusione, per le motivazioni più volte ricordate,

della Missione 13 'Tutela della salute' – è maggiormente concentrata nella realizzazione degli Obiettivi 9 'Imprese, innovazione e infrastrutture' (58,67%), 11 'Città e comunità sostenibili' (58,49%) e 13 'Lotta contro il cambiamento climatico' (53,99%), e riserva percentuali più alte di spesa rispetto a tutti gli altri soggetti considerati per gli Obiettivi 7 'Energia pulita e accessibile' (2,44%) e 8 'Lavoro dignitoso e crescita economica' (20,70%). Le Missioni con la maggiore percentuale di spesa sono la 10 'Trasporti e diritto alla mobilità' (45,15%), la 14 'Sviluppo economico e competitività' (11,32%) e la 15 'Politiche per il lavoro e la formazione professionale' (9,38%). La spesa risulta invece inferiore per gli Obiettivi 5 'Parità di genere' (1%), 6 'Acqua pulita e servizi igienico-sanitari' (0,96%), 16 'Pace, giustizia e istituzioni solide' (0,47%) e 14 'Vita sott'acqua' (0,42%), e nelle Missioni 6 'Politiche giovanili, sport e tempo libero' (0,3%), 3 'Ordine pubblico e scurezza' (0,17%) e 2 'Giustizia' (0%).

Alla fine dell'anno 2021, la matrice per la sostenibilità presentata nel capitolo è stata impiegata dalla Regione Toscana nell'ambito del Bilancio di previsione 2020-2022[10] e dell'analisi dei dati di spesa del Programma Attuativo Regionale del Fondo di Sviluppo e Coesione (PAR FSC) 2007-2013 e 2014-2020, analizzando come la spesa del Programma si sia distribuita nei due periodi di programmazione tra gli Obiettivi di Agenda 2030.

In conclusione, il lavoro presentato in questo studio offre molteplici opportunità per futuri studi e approfondimenti in materia, che grazie all'utilizzo di metodi automatici di analisi del contenuto, potrebbero concentrare l'analisi su altri aspetti, quali:

- un *benchmark* più ampio con i dati di spesa di altri soggetti ritenuti significativi (ad esempio una comparazione tra comuni o tra regioni);
- l'elemento giuridico, indagando in modo approfondito la correlazione tra gli *SDGs* e le competenze dei diversi livelli di governo;
- gli indicatori statistici ed il posizionamento della Regione Toscana rispetto all'attuazione degli *SDGs*;
- un'analisi contabile di legislatura, non solo in termini di competenza, ma anche in termini di tipologia di spesa (spesa corrente e per investimenti), di risorse allocate (impegni e pagamenti), di andamento temporale della spesa e della sua distribuzione tra le varie Direzioni regionali (per una mappatura della struttura regionale rispetto agli *SDGs*);
- i dati della spesa in termini di cassa desunti dai bilanci consuntivi

[10] Allegato h, Nota Integrativa: analisi distribuzione della spesa per l'anno 2020.

e rendiconti finanziari (Conti Pubblici Territoriali), da affiancare all'analisi qui proposta dei dati di previsione della spesa in termini di competenza derivanti dai bilanci di previsione;
- la programmazione della distribuzione della spesa tra obiettivi/interventi per esempio del PRS 2021/2025 in base agli *SGDs* di Agenda 2030 che più si vorranno raggiungere;
- la classificazione COFOG, al posto delle Missioni/Programmi, come punto di partenza per sviluppare la 'matrice per la sostenibilità';
- l'analisi dei dati di spesa (suddivisi in Missioni e Programmi) del Piano Nazionale di Ripresa e Resilienza (PNRR) in base agli *SDGs* di Agenda 2030, collocando la spesa destinata al raggiungimento delle Missioni del PNRR tra gli *SDGs* di Agenda 2030.

L'approfondimento di tutti questi aspetti potrebbe agevolare in futuro la programmazione delle politiche per la sostenibilità rispetto alle politiche di bilancio. Spendere di più o di meno non significa necessariamente migliorare o peggiorare una situazione esistente. Per tale ragione, la lettura dei dati di bilancio potrebbe o dovrebbe integrarsi con la valutazione degli esiti della programmazione e con i *target* e gli indicatori di Agenda 2030.

Appendice 1

Le interviste sono state effettuate in modalità *zoom meeting* nelle date tra il 24 agosto 2021 e il 20 settembre 2021.

Intervista n. 1 – Roberta Valletti, ricercatrice IRES Piemonte.

Intervista n. 2 – Roberto Impicciatore, professore associato Dipartimento di Scienze Statistiche "Paolo Fortunati", Università di Bologna.

Intervista n. 3 – Denis Baldan, operatore sociale Caritas Venezia.

Intervista n. 4 – Carlotta Giordani, Osservatorio regionale sull'immigrazione, Veneto lavoro.

Appendice 2

Fig. 1 Matrice per la sostenibilità dell'insieme delle regioni italiane. Anni 2016-2019.

INSIEME DELLE REGIONI ITALIANE																
MISSIONI	*SDGs*															
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
1 Servizi istituzionali, generali e di gestione																	
2 Giustizia																0,06	
3 Ordine pubblico e sicurezza																0,15	
4 Istruzione e diritto allo studio				5,54													
5 Tutela e valorizzazione dei beni e delle attività culturali											2,75						
6 Politiche giovanili, sport e tempo libero											0,55					0,55	
7 Turismo												1,7					
8 Assetto del territorio ed edilizia abitativa											4,03						
9 Sviluppo sostenibile e tutela del territorio e dell'ambiente						2,92					13,61	13,61	6,07		8,32		
10 Trasporti e diritto alla mobilità									33,96		33,96		33,96				
11 Soccorso civile													2,19				
12 Diritti sociali, politiche sociali e famiglia	8				1,02					8	1,56						
13 Tutela della salute (*)			76,95														
14 Sviluppo economico e competitività								11,8	10,73			11,76					
15 Politiche per il lavoro e la formazione professionale	8,71			3,49				8,71		8,71							
16 Agricoltura, politiche agroalimentari e pesca		4,6												0,33	4,6		
17 Energia e diversificazione delle fonti energetiche							1,63		1,63				1,63				
18 Relazioni con altre autonomie territoriali e locali																	
19 Relazioni internazionali																	0,76
Totale spesa per SDGs	16,7	4,6	76,95	9,03	1,02	2,92	1,63	20,5	46,32	16,71	56,46	27,07	43,85	0,33	12,92	0,76	0,76

(*) vedi nota Fig. 3 (cfr. cap. 10).

Fonte: elaborazione personale su dati della Banca Dati Amministrazioni Pubbliche (BDAP) della Ragioneria Generale dello Stato (RGS).

Fig. 2 Matrice per la sostenibilità del Comune di Firenze. Anni 2016-2019.

COMUNE DI FIRENZE																
MISSIONI	*SDGs*															
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
1 Servizi istituzionali, generali e di gestione																	
2 Giustizia																0,05	
3 Ordine pubblico e sicurezza																6,71	
4 Istruzione e diritto allo studio				8,85													
5 Tutela e valorizzazione dei beni e delle attività culturali											10,29						
6 Politiche giovanili, sport e tempo libero											2,9					2,9	
7 Turismo												0,26					
8 Assetto del territorio ed edilizia abitativa											4,01						
9 Sviluppo sostenibile e tutela del territorio e dell'ambiente						0,06					13,52	13,52	0,48		2,89		
10 Trasporti e diritto alla mobilità									37,86		37,86		37,86				
11 Soccorso civile													0,15				
12 Diritti sociali, politiche sociali e famiglia	14,3				5,49					14,26	1,66						
13 Tutela della salute (*)			0,01														
14 Sviluppo economico e competitività								0,87	0,23			0,87					
15 Politiche per il lavoro e la formazione professionale	0,19			0,19				0,19		0,19							
16 Agricoltura, politiche agroalimentari e pesca		0												0	0		
17 Energia e diversificazione delle fonti energetiche							0		0				0				
18 Relazioni con altre autonomie territoriali e locali																	
19 Relazioni internazionali																	0,08
Totale spesa per SDGs	14,5	0	0,01	9,04	5,49	0,06	0	1,06	38,09	14,45	70,24	14,65	38,49	0	2,89	9,66	0,08

(*) vedi nota Fig. Fig. 3 (cfr. cap. 10).

Fonte: elaborazione personale su dati della Banca Dati Amministrazioni Pubbliche (BDAP) della Ragioneria Generale dello Stato (RGS).

Fig. 3 Matrice per la sostenibilità dell'Italia. Anni 2016-2019.

ITALIA

MISSIONI (1)	*SDGs*																
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
1 Servizi istituzionali, generali e di gestione																	
2 Giustizia (2)																2,6	
3 Ordine pubblico e sicurezza (3)																9,68	
4 Istruzione e diritto allo studio (4)				17													
5 Tutela e valorizzazione dei beni e delle attività culturali (5)											0,69						
6 Politiche giovanili, sport e tempo libero (6)											0,24					0,24	
7 Turismo (7)												0,01					
8 Assetto del territorio ed edilizia abitativa (8)											0,4						
9 Sviluppo sostenibile e tutela del territorio e dell'ambiente (9)						0,15					0,38	0,38	0,21		0,21		
10 Trasporti e diritto alla mobilità (10)									4,46		4,46		4,46				
11 Soccorso civile (11)													2,35				
12 Diritti sociali, politiche sociali e famiglia (12)	11,1				0,01					42,46	0,74						
13 Tutela della salute (13)(*)			15,49														
14 Sviluppo economico e competitività (14)								7,83	7,94			7,83					
15 Politiche per il lavoro e la formazione professionale (15)	3,24			0,1				3,24		3,24							
16 Agricoltura, politiche agroalimentari e pesca (16)		0,26												0,13	0,26		
17 Energia e diversificazione delle fonti energetiche (17)							0,1		0,1				0,1				
18 Relazioni con altre autonomie territoriali e locali																	
19 Relazioni internazionali (18)																	8,1
Totale spesa per SDGs	14,3	0,26	15,49	17,1	0,01	0,15	0,1	11,1	12,5	45,7	6,91	8,22	7,12	0,13	0,47	12,52	8,1

(1) Le Missioni previste dal D.lgs 118/2011 sono state associate con quelle previste a livello statale dalla Legge 31 dicembre 2009, n. 196 (Legge di contabilità e finanza pubblica) e sue successive modificazioni e integrazioni.
(2) Corrisponde alla Missione "6 Giustizia".
(3) Corrisponde alla Missione "5 Difesa e sicurezza del territorio" e la Missione "7 Ordine pubblico e sicurezza".
(4) Corrisponde alla Missione "22 Istruzione scolastica" e la Missione "23 Istruzione universitaria e formazione post-universitaria".
(5) Corrisponde alla Missione "21 Tutela e valorizzazione dei beni e attività culturali e paesaggistici".
(6) Corrisponde alla Missione "30 Giovani e sport".
(7) Corrisponde alla Missione "31 Turismo".
(8) Comprende la Missione "19 Casa e assetto urbanistico" e i Programmi 14.8 e 14.9 della Missione "14 Infrastrutture pubbliche e logistica".
(9) Corrisponde alla Missione "18 Sviluppo sostenibile e tutela del territorio e dell'ambiente". I GOAL 6, 11 e 12 comprendono il Programma "14.5 Sistemi idrici, idraulici ed elettrici" della Missione "14 Infrastrutture pubbliche e logistica".
(10) Comprende la Missione "13 Diritto alla mobilità e sviluppo dei sistemi di trasporto" e il Programma 14.11 della Missione "14 Infrastrutture pubbliche e logistica".
(11) Comprende la Missione "8 Soccorso civile" e il Programma 14.10 della Missione "14 Infrastrutture pubbliche e logistica".
(12) Comprende la Missione "24 Diritti sociali, politiche sociali e famiglia". Il GOAL 10 include la Missione "25 Politiche previdenziali", parte della Missione "27 Immigrazione, accoglienza e garanzia dei diritti" ad esclusione dei rapporti con le confessioni religiose e la Missione "28 Sviluppo e riequilibrio territoriale".
(13) Comprende la Missione "20 Tutela della salute" e il Programma "3.6 Concorso dello Stato al finanziamento della spesa sanitaria" della Missione "3 Relazioni finanziarie con le autonomie territoriali".
(14) Comprende la Missione "11 Competitività e sviluppo delle imprese", la Missione "12 Regolazione dei mercati", la Missione "16 Commercio internazionale e internazionalizzazione del sistema produttivo", la Missione "17 Ricerca e innovazione". Il GOAL 9 include la Missione "15 Comunicazioni".
(15) Corrisponde alla Missione 26 "Politiche per il lavoro".
(16) Corrisponde alla Missione 9 "Agricoltura, politiche agroalimentari e pesca".
(17) Comprende la Missione "10 Energia e diversificazione delle fonti energetiche" e il Programma "14.5 Sistemi idrici, idraulici ed elettrici" della Missione "14 Infrastrutture pubbliche e logistica".
(18) Corrisponde alla Missione "4 L'Italia in Europa e nel Mondo".

(*) vedi nota Fig. 3 (cfr. cap. 10).

Fonte: elaborazione personale su dati della Ragioneria Generale dello Stato (RGS) del Ministero dell'Economia e delle Finanze (MEF).

Riferimenti bibliografici

Aggarwal C. C., Zhai C. (2012) (a cura di), Mining Text Data, Springer.

Alhassan I., Sammon D., Daly M. (2016), Data governance activities: an analysis of the literature, in Journal of Decision Systems, vol. 25(sup1), pp. 64-75.

Arthur P., Bode K. (2014) (a cura di), Advancing Digital Humanities: Research, Methods, Theories, Palgrave Macmillan.

Bartoli Langeli A. (2006), Notai. Scrivere documenti nell'Italia medievale, Viella, Roma, pp. 59-86.

Beaudouin V. (2016), Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis, Glottometrics 33, pp. 56-72.

Bellavitis A., Frank M., Sapienza V. (2017) (a cura di), Garzoni. Apprendistato e formazione tra Venezia e l'Europa in età moderna, Universitas studiorum, Mantova.

Benzécri J.-P. (1982), Histoire et préhistoire de l'analyse des données, Dunod, Paris.

Benzécri J.-P. (1992), Correspondence Analysis Handbook, Marcel Dekker, Inc., New-York, Basel, Hong Kong.

Berger A. L., Della Pietra S. A., Della Pietra V. J. (1996), A maximum entropy approach to natural language processing, in Computational Linguistics, vol. 22(1), pp. 39-71.

Berry M. W. (2004), (a cura di), Survey of Text Mining. Clustering, Classification, and Retrieval, Springer-Verlag, New-York.

Blei D. M., Lafferty J.D. (2006), Dynamic topic models, in Proceedings of the 23rd international conference on machine learning, pp 113-120.

Blei D.M., Lafferty J.D. (2007), A correlated topic model of science, Statistics, vol. 1(1), pp. 17-35.

Blei D.M., Lafferty J.D. (2009), Topic models, in Sahami A., Srivastava M. (a cura di), Text Mining: Theory and Applications, Taylor and Francis, pp. 71-93.

Blei D.M., Ng A., Jordan M. (2003), Latent Dirichlet Allocation, in The Journal of Machine Learning Research, vol. 3, pp. 993-1022.

Bolasco S. (1999), Analisi multidimensionale dei dati, Carocci, Roma.

Bolasco S. (2005), Statistica testuale e text mining: alcuni paradigmi applicativi, in Quaderni di Statistica, vol. 7, pp. 17-53.

Busa R. (1974-1980), Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices ed concordantiae, Frommann Holzboog, Stoccarda.

Capano G., Howlett M., Jarvis D. S., Ramesh M., Goyal N. (2020), Mobilizing policy (in) capacity to fight COVID-19: Understanding variations in state responses, in Policy and Society, vol. 39(3), pp. 285-308.
Celentano A., Cortesi A., Mastandrea P. (2004), Informatica Umanistica: una disciplina di confine, in Mondo Digitale, vol. 4, pp. 44-55.
Celotto A. (2015), In nome del popolo italiano. Quante sentenze vengono pronunciate ogni anno in Italia?, in Huffpost Italia, https://www.huffingtonpost.it/alfonso-celotto/nome-popolo-italiano-sentenze-pronunciate-italia_b_6995584.html.
Cerrillo-Martínez A., Casadesús-de-Mingo A. (2021), Data governance for public transparency, in El profesional de la información, vol. 30(4).
Ceron A., Curini L., Iacus S. M. (2014), Social Media e Sentiment Analysis. L'evoluzione dei fenomeni sociali attraverso la rete, Sxi – Springer per l'innovazione, Milano.
Cortelazzo M. A., Nadalutti P., Tuzzi A. (2013), Improving Labbé's Intertextual Distance: Testing a Revised version on a Large Corpus of Italian Literature, in Journal of Quantitative Linguistics, vol. 20(2), pp. 125-152.
Criado J. I., Sandoval-Almazan R., Gil-Garcia J. R. (2013), Government innovation through social media, in Government Information Quarterly, vol. 30(4), pp. 319-326.
De Mauro A. (2019), Big Data Analytics. Analizzare e interpretare dati con il machine learning, Apogeo.
Delmastro M., Nicita A. (2019), Big Data. Come stanno cambiando il nostro mondo, Il Mulino, Bologna.
Dente B. O. (2011), Le decisioni di policy, Il Mulino, Bologna.
Duranti L. (1998), Diplomatics: New Uses for an Old Science, The Scarecrow Press, Lanham.
Eder M., Rybicki J., Kestemont M. (2016), Stylometry with R: A Package for Computational Text Analysis, in The R Journal, vol. 8(1), pp. 107-121.
Erboso A. (2019), Codice diplomatico veneziano Lanfranchi, elenco sommario, Archivio di Stato di Venezia.
Flores R. D. (2017), Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data, in American Journal of Sociology, vol. 123(2), pp. 333-384.
Foucault M. (1966), Le parole e le cose. Un'archeologia delle scienze umane, trad. it. (2016), Rizzoli, Milano.
Gieryn T. F. (1983), Boundary-work and the demarcation of science from non-science: strains and interests in professional ideologies of scientists, in American Sociological Review, vol. 48(6), pp. 781-795.

Gieryn T. F. (1995), Boundaries of Science, in Jasanoff S., Markle G. E., Petersen J. C., Pinch T. (a cura di), Handbook of Science and Technology Studies, Sage, Newbury Park, pp. 393-443.

Gilardi F., Shipan C. R., Wüest B. (2021), Policy Diffusion: The Issue-Definition Stage, in American Journal of Political Science, vol. 65(1), pp. 21-35.

Gilardi F., Wüest B. (2018), Text-as-Data Methods for Comparative Policy Analysis, Working Paper, https://www.fabriziogilardi.org/resources/papers/Gilardi-Wueest-TextAsData-Policy-Analysis.pdf.

Gilardi F., Wüest B. (2020), Using text-as-data methods in comparative policy analysis, in Handbook of research methods and applications in comparative policy analysis, Edward Elgar Publishing.

Giuliano L., La Rocca G. (2008), L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso, Led edizioni, Milano.

Gold M. (2012), Debates in Digital Humanities, University of Minneapolis Press, Minneapolis and London.

Goodman S. W. (2021), 'Good American citizens': a text-as-data analysis of citizenship manuals for immigrants, 1921–1996, in Journal of Ethnic and Migration Studies, vol. 47(7), pp. 1474-1497.

Greenacre M.J. (2007), Correspondence analysis in practice, Chapman & Hall, London.

Griffiths T., Steyvers M. (2004), Finding scientific topics, in Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 101(1), pp. 5228-5235.

Grün B., Hornik K. (2011), Topicmodels: An R Package for Fitting Topic Model, in Journal of Statistical Software, vol. 40(13), pp. 1-30.

Hearst M. A. (1999), Untangling text data mining, in Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, Association for Computational Linguistics, pp 3-10.

Holmes D. (1998), The Evolution of Stylometry in Humanities Scholarship, in Literary and Linguistic Computing, vol. 13(3), pp. 111-117.

Jardine N., Van Rijsbergen C.J. (1971), The use of hierarchical clustering in information retrieval, in Information Storage and Retrieval, vol. 7, pp. 217-240.

Jenkins H. (2007), Cultura Convergente, Apogeo.

Jones B. D., Baumgartner F. R. (2005), The politics of attention: How government prioritizes problems, University of Chicago Press.

Joula P. (2006), Autorship attribuition, Foundation and Trends, in Information Retrival, vol. 1(3), pp. 233-334.

Juola P. (2008), Killer Applications in Digital Humanities, in Literary and Linguistic Computing, vol. 23(1), pp. 73-83.
Kahneman D., Sibony O., Sunstein C. R. (2021), Noise: A flaw in human judgment, Little, Brown.
Kolbjørnsrud V., Amico R., Thomas R. J. (2016), How artificial intelligence will redefine management, in Harvard business review, vol. 2(1), pp. 3-10.
Krippendorff K. (1983), Analisi del contenuto. Introduzione metodologica, ERI, Torino.
Labbé C., Labbé D. (2001), Inter-Textual Distance and Authorship Attribution Corneille and Moliére, in Journal of Quantitative Linguistics, vol. 8, pp. 213-231.
Labbé D. (2007), Experiments on authorship attribution by intertextual distance in English, in Journal of Quantitative Linguistics, vol. 14, pp. 33-80.
Laney D. (2001), 3D Data Management: controlling data volume, velocity and variety, Meta Group Research Note.
Lanfranchi L. (1942), Lavori e programmi per una pubblicazione delle carte veneziane anteriori al 1200, in Archivio veneto, vol. 30(5), pp. 246-252.
Lanfranchi L. (1944-1986), Codice diplomatico veneziano, dattiloscritto inedito, Archivio di Stato di Venezia.
Lanfranchi L. (1984), Per un codice diplomatico veneziano del secolo XIII, in Viridarium floridum. Studi di storia veneta offerti dagli allievi a Paolo Sambin, Antenore, Padova, pp. 355-363.
Lasswell H. D. (1927), Propaganda Technique in the World War, Alfred A. Knopf, New York.
Lasswell H. D. (1949), The Language of Politics: Studies in Quantitative Semantics, George Stewart, New York, trad. it. (1979), Il linguaggio della politica Studi di semantica quantitativa, ERI, Torino.
Lebart L., Salem A. (1988), Analyse statistique des données textuelles: questions ouvertes et lexicometrie, Dunod, Paris.
Lebart L., Morineau A., Warwick K.M. (1984), Multivariate descriptive statistical analysis. Correspondence analysis and related techniques for large matrices, Wiley, New York.
Leonard T. C. (2008), Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness, in Constitutional Political Economy, vol. 19(4), pp. 356-360.
Losito G. (1993), L'analisi del contenuto nella ricerca sociale, Franco Angeli, Milano.

Luhn H. (1959), Auto-encoding of documents for information retrieval systems, in M. Boaz (a cura di), Modern Trends in Documentation, Pergamon Press, London, pp. 45-58.

Maciejewski M. (2017), To do more, better, faster and more cheaply: Using big data in public administration, in International Review of Administrative Sciences, vol. 83(1_suppl), pp. 120-135.

Melloni E., Righettini M.S. (2019), Durante e dopo l'emergenza: il ruolo della valutazione delle performance per rafforzare la capacità amministrativa delle pubbliche amministrazioni, in Rassegna Italiana di Valutazione, vol. 23(75), pp. 42-59.

Michel J.B., Shen Y.K., Aiden A.P., Veres A., Gray M.K., Pickett J.P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A., Aiden E.L (2011), Quantitative Analysis of Culture Using Millions of Digitized Books, in Science, vol. 331(6014), pp. 176-182.

Monroe B. L., Schrodt P. A. (2008), Introduction to the special issue: The statistical analysis of political text, in Political Analysis, vol. 16(4), pp. 351-355.

Murtagh F. (2005), Correspondence Analysis and Data Coding with Java and R, Chapman & Hall/CRC, London.

Nanetti A., Benvenuti D. (2021), Engineering historical Memory and the interactive Exploration of Archival Documents: the online Application for Pope Gregory X's Privilege for the monastic community of Mount Sinay (1274) as a prototype, in Umanistica digitale, vol. 10, pp. 325-357.

Orlandi T., Mordenti R. (2003), Lo status accademico dell'informatica umanistica, in Archeologia e calcolatori, vol. 14, pp. 7-32.

Ostrom E. (2005), Understanding institutional diversity, Princeton University press, New Jersey.

Penzo Doria G. (2020), Uno strano DPCM 24 marzo 2020: la diplomatica contemporanea contro i documenti falsi, in FiloDiritto, https://www.filodiritto.com/uno-strano-dpcm-24-marzo-2020-la-diplomatica-contemporanea-contro-i-documenti-falsi.

Porter M. (1980), An algorithm for suffix stripping, in Program, vol. 14(3), pp. 130-137.

Prescott A. (2012), Consumers, creators or commentators?: Problems of audience and mission in the digital humanities, in Arts and Humanities in Higher Education, vol. 11(1-2), pp. 61-75.

R development core team (2016), R: a language and environment for statistical computing [software], Vienna, Austria: R foundation for statistical computing, http://www.r-project.org.

Ratinaud P. (2014), IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires (Version 0.7 alpha 2), http://www.iramuteq.org.

Ratinaud P., Marchand P. (2012), Application de la méthode ALCESTE à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "CableGate" avec IRaMuTeQ, in Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles, Liège, Belgique, pp. 835-844.

Ratinaud P., Marchand P. (2015), Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014), in Mots. Les Langages Du Politique, vol. 108, pp.57-77.

Regione Toscana (2020), PO.R.TO.S. Tratto da POrtale Regione Toscana Sismica, http://www327.regione.toscana.it/web/portos.

Reinert M. (1983), Une methode de classification descendante hierarchique: application a l'analyse lexicale par context, in Les Cahiers de l'Analyse des Données, vol. 8(2), 187-198.

Reinert M. (1990), ALCESTE: Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval, in Bulletin de Méthodologie Sociologique, vol. 26, pp. 24-54.

Reinert M. (1993), Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars, in Language et Société, n. 66, pp. 5-39.

Reinert M. (1998), Mondes lexicaux et Topoi dans l'approche Alceste, in Mellet E., Vuillaume M. (a cura di), Mots chiffrés et déchiffrés, Honoré Champion, Paris, pp. 289-303.

Righettini M. S. (2021), Framing Sustainability. Evidence from Participatory Forums to Taylor the Regional 2030 Agenda to Local Contexts, in Sustainability, vol. 13(8), 4435.

Righettini M. S., Sbalchiero S. (2017a), The regulatory state and its variants in communications sector: Paradigms, strategies and ideas of universality and inclusion in Europe, in Stato e mercato, vol. 37(2), pp. 247-282.

Righettini M. S., Sbalchiero S. (2017b), Institutional entrepreneurship and change in consumer protection policy in the telecommunications sector: innovations in the text-based analysis approach, in Policy and Society, vol. 36(4), pp. 611-631.

Righettini M. S., Sbalchiero, S. (2021), Detecting and Comparing Institutional Change: Evidence from Consumer/User Protection's Salience in

Telecommunications in Nine European Countries, in Rivista Italiana di Politiche Pubbliche, vol. 16(2), pp. 191-217.
Roberts M. E., Stewart B. M., Tingley D., Lucas C., Leder-Luis J., Gadarian S., Albertson B., Rand D. (2014), Structural topic models for open-ended survey responses, in American Journal of Political Science, vol. 58, pp. 1064-1082.
Roberts S., Snee H., Hine C., Morey Y., Watson H. (2016) (a cura di), Digital Methods for Social Science. An Interdisciplinary Guide to Research Innovation, Palgrave Macmillan.
Rudman J. (1998), The state of authorship attribution studies: some problems and solutions, in Comput. Humanit., vol. 31, pp. 351-365.
Sanger J., Feldman R. (2007), The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, Cambridge.
Sbalchiero S. (2018), Finding topics: a statistical model and a quali-quantitative method, in Tuzzi A., Tracing the Life-Course of Ideas in the Humanities and Social Sciences, Springer, pp. 189-210.
Sbalchiero S. (2021), Dal metodo all'esperienza. Fare ricerca con la sociologia comprendente, Padova University Press, Padova.
Sbalchiero S., Eder M. (2020), Topic modeling, long texts and best number of topics: some problems and solutions, in Quality & Quantity, vol. 54, pp. 1095-1108.
Sbalchiero S., Tuzzi A. (2016), Scientists' spirituality in scientists' words. Assessing and enriching the results of a qualitative analysis of in-depth interviews by means of quantitative approaches, in Quality & Quantity, vol. 50(3), pp. 1333-1348.
Siddiki S., Frantz C. (2022), The institutional grammar in policy process research, in Policy Studies Journal, vol. 50(4).
Smyrnaios N., Ratinaud P. (2013), Comment articuler analyse des réseaux et des discours sur Twitter, in tic&société, vol. 7(2), pp. 120-147.
Smyrnaios N., Ratinaud P. (2017), The Charlie Hebdo Attacks on Twitter: A Comparative Analysis of a Political Controversy in English and French, in Social Media + Society, vol. 3(1), pp. 1-13
Slapin J. B., Proksch S. O. (2008), A scaling model for estimating time-series party positions from texts, in American Journal of Political Science, vol. 52(3), pp. 705-722.
Sorokin P. A. (1956), Fads and Foibles in Modern Sociology and Related Siences, Henry Regnery, Chicago.
Sunstein C. R. (2002), Risk and reason: Safety, law, and the environment,

Cambridge University Press, Cambridge.
Thomas W. I., Znaniecki F. (1918- 1920), Il contadino polacco in Europa e in America, trad. it. (1968), Comunità, Milano.
Tuzzi A. (2003), L'analisi del contenuto: Introduzione ai metodi e alle tecniche di ricerca, Carocci editore, Roma.
Trevisani M., Tuzzi A. (2020), Distance measures for exploring pairs of novels in a large corpus of Italian literature, in Book of Short Papers SIS2020, pp. 229-234.
Tuzzi A. (2010), What to put in the bag? Comparing and contrasting procedures for text clustering, in Italian Journal of Applied Statistics/Statistica Applicata, vol. 22(1), pp. 77-94.
Tuzzi A., Cortelazzo M.A. (2018), What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, in Digital Scholarship in the Humanities, vol. 33(3), pp. 685-702.
Tuzzi A. (2018), Tracing the Life-Course of Ideas in the Humanities and Social Sciences, Springer.
Zhao Y., Xub X., Wangc M. (2019), Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews, in International Journal of Hospitality Management, vol. 76, pp. 111-121.

Sitografia

https://2018.datadriveninnovation.org/it/, 18 maggio 2018, Data Driven Innovation.

http://asve.arianna4.cloud/.

https://asvis.it/.

http://briguglio.asgi.it/immigrazione-e-asilo/2005/maggio/anticip-dossier-caritas.html#:~:text=Un'impennata%20si%20ha%20tra,registrazione%20dei%20permessi%20di%20 soggiorno.

http://doi.org/10.5281/zenodo.3567769.

https://eadh.org/projects/read.

http://ec.europa.eu/eurostat/publications/statistical-books.

https://ec.europa.eu/eurostat/statistics-explained/index.php/Government_expenditure_by_function_%E2%80%93_COFOG.

https://forbes.it/2021/04/28/ilde-forgione-la-social-manager-che-ha-rilanciato-gli-uffizi-grazie-a-tiktok/.

https://garzoni.hypotheses.org/.

https://lucene.apache.org/core/2_9_4/api/core/org/apache/lucene/search/Similarity.

http://ml-api.daf.teamdigitale.it/ (link non più attivo).

https://noi-italia.istat.it/pagina.php?L=0&categoria=4&dove=ITALIA.

https://it.wikipedia.org/wiki/Software_libero#Le_%22quattro_libert%C3%A0%22, Software libero, da Wikipedia, l'enciclopedia libera.

http://online.leggiditalia.it/.

http://pnd.beniculturali.it/il-piano/.

http://san.beniculturali.it/web/san/progetti-di-digitalizzazione?p_p_id=prjdgtportlet_WAR_prjsanportlet_INSTANCE_nA2e&p_p_lifecycle=1&p_p_state=normal&p_p_mode=view&p_p_col_id=box_contenuto&p_p_col_count=1&_prjdgtportlet_WAR_prjsanportlet_INSTANCE_nA2e__spage=%2Fportlet_action%2Fsan%2Fprjdgt%3Fstep%3Delenco_completo&_prjdgtportlet_WAR_prjsanportlet_INSTANCE_nA2e_step=elenco_completo.

https://temi.camera.it/leg18/temi/iniziative-per-prevenire-e-contrastare-la-diffusione-del-nuovo-coronavirus.html#iniziative-per-prevenire-e-contrastare-la-diffusione-del-nuovo-coronavirus-.

https://unstats.un.org/sdgs/indicators/indicators-list/.

http://www301.regione.toscana.it/bancadati/atti/indexAttiD.xml, Regione Toscana, Atti dei Dirigenti.

http://www327.regione.toscana.it/web/portos.
https://www.archiviodistato.firenze.it/pergasfi/.
http://www.arpat.toscana.it/.
http://www.bdap.tesoro.it/sites/openbdap/cittadini.
http://www.civesveneciarum.net.
https://www.facebook.com/corrieredellasera/videos/981605565574486.
https://www.facebook.com/uffizigalleries/.
https://www.helpconsumatori.it/secondo-piano/consumers-forum-conciliazione-paritetica-una-sconosciuta/180021/.
http://www.iccd.beniculturali.it/it/progetti/4597/arco-architettura-della-conoscenza-ontologie-per-la-descrizione-del-patrimonio-culturale.
http://www.iramuteq.org.
https://www.istat.it/it/benessere-e-sostenibilit%C3%A0/obiettivi-di-sviluppo-sostenibile/gli-indicatori-istat.
https://www.istat.it/it/files/2019/07/Statistica-report-Bilancio-demografico-2018.pdf.
https://www.leggiditaliaprofessionale.it.
http://www.regione.toscana.it/banchedati.
http://www.rgs.mef.gov.it/VERSIONE-I/e_government/amministrazioni_pubbliche/arconet/.
http://www.rgs.mef.gov.it/VERSIONE-I/home.html.
http://www.r-project.org.
http://www.statistica.beniculturali.it/.
https://www.timemachine.eu/.
https://www.uffizi.it/news/uffizi-facebook-2020.
https://www.uffizi.it/pagine/social_media_policy_uffizigalleries.

Non solo parole. Come si estrae valore pubblico dalle banche dati testuali pubbliche e come questo valore può migliorare processi decisionali e politiche e contribuire all'innovazione del management nelle pubbliche amministrazioni? Il volume offre numerosi esempi in questa direzione e su come l'approccio TestiComeDati (*Tex as data*) possa essere alla portata di tutti. Il volume fornisce spunti di metodo e un ventaglio di possibili applicazioni per le pubbliche amministrazioni che dispongono di una grande mole di dati testuali (documenti di programmazione, leggi, progetti, decreti, consultazioni, delibere, sentenze) ma ne fanno scarso uso per migliorare la qualità delle politiche e delle decisioni. I capitoli illustrano elaborazioni automatiche e semi-automatiche di banche dati di varie dimensioni a supporto della capacità di policy, delle capacità amministrative e della qualità dei servizi pubblici.

Molti degli autori dei capitoli sono dipendenti pubblici che hanno applicato il metodo dei Testi come dati (TasD) all'interno del proprio contesto lavorativo guidati da un team di docenti dell'Universitá di Padova nell'ambito del Master in Policy Innovation and Sustainability Impact Assessment (PISIA).

ISBN 978-88-6938-318-2

€ 30,00