

Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche



RELAZIONE FINALE

**TOPIC MODELING APPLICATO ALLA COMUNICAZIONE
DIGITALE DI MUSEI SCIENTIFICI ITALIANI:
IMPLEMENTAZIONE, LIMITI E MIGLIORAMENTI DI
BERTOPIC PER LA LINGUA ITALIANA**

Relatore Prof. Andrea Sciandra

Dipartimento di Scienze Statistiche

Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata

Correlatore Prof. Stefano Crabu

Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata

Laureando Andrea Simonetto

Matricola N 2121340

Anno Accademico 2025/2026

Indice

1	Introduzione	7
2	Approcci classici al <i>Topic Modeling</i>	11
2.1	<i>Latent Dirichlet Allocation</i>	11
2.2	Structural Topic Model	13
3	BERTopic: architettura e pipeline	17
3.1	Motivazione: oltre il Bag-of-Words	17
3.2	La pipeline standard	19
3.2.1	Embedding dei documenti con Sentence-BERT	19
3.2.2	Riduzione dimensionale con UMAP	20
3.2.3	Clustering con HDBSCAN	21
3.2.4	Vectorization con CountVectorizer	23
3.2.5	Pesatura con c-TF-IDF	23
3.2.6	Rappresentazione finale dei topic	24
3.3	Punti di forza e limiti	24
4	Estensioni e varianti di BERTopic proposte	29
4.1	Modelli di embedding alternativi	29
4.2	Riduzione dimensionale alternativa	29
4.3	Metodi di clustering alternativi	30
4.4	Metodo di rappresentazione alternativo	31
4.5	Integrazione di Gibbs Sampling in BERTopic	32
5	Dati: i musei scientifici italiani su social media	35

5.1	Contesto e motivazione del corpus	35
5.2	<i>Preprocessing</i>	36
5.2.1	<i>Preprocessing</i> per STM e LDA	37
5.2.2	<i>Preprocessing</i> per BERTopic	39
5.3	Metadati disponibili per STM	39
6	Metriche di confronto e Disegno sperimentale	41
6.1	Obiettivo del confronto	41
6.2	Coerenza semantica (UMass)	42
6.3	Diversità dei topic	44
6.4	Distanza di Hellinger	45
6.5	Esclusività e punteggio FREX	47
6.6	Valutazione qualitativa	48
6.7	Configurazioni dei modelli	50
6.7.1	LDA	50
6.7.2	STM	50
6.7.3	BERTopic	51
7	Risultati	53
7.1	Panoramica generale degli argomenti	53
7.2	Confronto tra LDA, STM e BERTopic	54
7.2.1	LDA e STM	54
7.2.2	Modelli probabilistici e BERTopic	56
7.3	Varianti di BERTopic	60
7.3.1	Modello di embedding monolingue e rappresentazione con RAKE	60
7.3.2	Analisi comparativa delle architetture spaziali: Ridu- zione Dimensionale e Clustering	65
7.3.3	Rappresentazione con RAKE	68
7.3.4	Introduzione del Gibbs Sampling	71
8	Conclusioni	75
8.1	Il rumore lessicale e la gestione dei token sconosciuti	75
8.2	Coerenza semantica e interpretabilità dei topic	77
8.3	Le varianti e l'assenza di un miglioramento netto	78

INDICE

8.4	Direzioni future	79
A	Top words dei topic	81
A.1	Nota introduttiva	81
A.2	LDA	81
A.3	STM	83
A.4	BERTopic	84
B	Valutazione qualitativa dei topic	87
B.1	Nota introduttiva	87
B.2	LDA	87
B.3	STM	88
B.4	BERTopic	89
B.5	Grafi di sovrapposizione dei topic	91

Capitolo 1

Introduzione

La crescita esponenziale della quantità di testo disponibile in formato digitale (articoli, documenti istituzionali, contenuti generati dagli utenti sui social network, archivi storici digitalizzati) ha reso l'elaborazione manuale di grandi corpora un'operazione proibitiva in termini di tempo e di costo. In questo contesto, le tecniche di *Natural Language Processing* (NLP) svolgono un ruolo centrale nel processo di estrazione di conoscenza dai dati testuali non strutturati. Tra queste, il *topic modeling* si è affermato come uno degli strumenti più diffusi per l'analisi automatica di ampie collezioni documentali. Il *topic modeling* è una tecnica statistica finalizzata a rivelare la struttura semantica latente di un corpus, sulla base di due assunzioni fondamentali: ogni documento è rappresentato come una miscela di argomenti (*topic*), e ogni *topic* è a sua volta una distribuzione di probabilità sull'insieme del vocabolario. L'obiettivo del modello è dunque quello di inferire, a partire dai soli dati osservati, le strutture tematiche che li hanno generati, fornendo una rappresentazione interpretabile dei contenuti a livello di documento. L'apprendimento di rappresentazioni del testo costituisce uno dei problemi fondamentali del NLP, ed è stato affrontato lungo traiettorie diverse. Da un lato, il topic modeling si concentra su strutture tematiche di alto livello, utili per organizzare e interpretare grandi volumi di testo. Dall'altro, i modelli linguistici come word2vec, BERT e GPT sfruttano *embedding* contestuali a li-

vello di parola, particolarmente efficaci per operazioni a granularità fine come l'analisi del sentiment o il *named entity recognition*. I due approcci risultano in larga misura complementari: il topic modeling fornisce una sintesi tematica del corpus, mentre i modelli linguistici arricchiscono di contesto semantico le singole occorrenze. L'avvento dei *Large Language Models* (LLM) ha aperto nuove possibilità anche per il topic modeling stesso. Approcci di tipo neurale, basati su architetture *transformer*, sfruttano le rappresentazioni contestualizzate prodotte da modelli pre-addestrati per generare topic più coerenti e semanticamente interpretabili rispetto alle tecniche classiche. Storicamente, i modelli di topic modeling possono essere ricondotti a tre grandi famiglie: algebrici (come la *Latent Semantic Indexing*), probabilistici (di cui la *Latent Dirichlet Allocation* è l'esempio più noto, insieme alle sue estensioni come lo *Structural Topic Model*) e neurali (di cui BERTopic costituisce uno dei rappresentanti più recenti). A un secondo livello, la distinzione fondamentale riguarda il tipo di rappresentazione testuale impiegata: *bag of words* (BoW), che ignora ordine e contesto delle parole, oppure *embedding* contestuali, che ne preservano le relazioni semantiche. Le applicazioni del topic modeling spaziano oggi su un'ampia varietà di domini: *information retrieval*, classificazione documentale, analisi dei social media, bioinformatica, studio dell'evoluzione tematica in archivi storici e tanti altri. La sua versatilità lo rende uno strumento di riferimento ogni volta che sia necessario estrarre informazioni strutturate da raccolte testuali eterogenee e di grandi dimensioni.

L'applicazione di queste tecniche alla lingua italiana introduce, tuttavia, difficoltà specifiche. La maggior parte dei modelli linguistici pre-addestrati e degli strumenti di pre-elaborazione è stata sviluppata per l'inglese, mentre l'italiano resta una lingua a bassa disponibilità di risorse nel panorama NLP, sia per la relativa scarsità di dataset dedicati, sia per il supporto limitato alle specificità morfologiche e sintattiche della lingua. Questo si traduce, in pratica, in una minore qualità delle rappresentazioni vettoriali disponibili e in un comportamento meno prevedibile dei modelli neurali rispetto a quanto avviene per l'inglese.

Il presente lavoro si articola in due fasi principali. Nella prima si conduce un confronto sistematico tra tre modelli di riferimento (LDA, STM e BERTopic)

applicati a corpora in lingua italiana, valutandone le prestazioni mediante metriche di coerenza e diversità dei topic. Nella seconda si propongono alcune strategie di miglioramento di BERTopic, intervenendo sulle componenti di *embedding*, riduzione dimensionale e *clustering*, con l'obiettivo di ottenere rappresentazioni tematiche più accurate e semanticamente significative per il testo italiano.

Questa tesi è organizzata come segue. Il Capitolo 2 presenta i modelli classici di topic modeling (LDA e STM) discutendone formulazione generativa, metodi di inferenza e limiti applicativi. Il Capitolo 3 è dedicato a BERTopic e alla sua architettura modulare. Il Capitolo 4 è dedicato alla spiegazione delle modifiche effettuate alla pipeline standard di BERTopic. Il Capitolo 5 descrive il *dataset* e la metodologia sperimentale adottata. Il Capitolo 6 presenta le modalità di valutazione adottate e le scelte di configurazione dei modelli posti a confronto. Il Capitolo 7 presenta i risultati del confronto e le strategie di miglioramento proposte. Il Capitolo 8 discute le conclusioni e le possibili direzioni future di ricerca.

Capitolo 2

Approcci classici al *Topic Modeling*

2.1 *Latent Dirichlet Allocation*

La *Latent Dirichlet Allocation* (LDA), introdotta da Blei, Ng e Jordan nel 2003 [3], rappresenta uno dei modelli probabilistici generativi più influenti per l'analisi non supervisionata di collezioni di dati discreti, in particolare di corpora testuali. La sua formulazione nasce per superare i limiti strutturali dei modelli precedenti, quali il *Probabilistic Latent Semantic Indexing* (pLSI) di Hofmann [18], il cui numero di parametri cresce linearmente con la dimensione del corpus rendendolo vulnerabile al sovradattamento e incapace di assegnare probabilità coerenti a documenti non osservati in fase di addestramento. La LDA si configura invece come un modello gerarchico bayesiano a tre livelli, in cui ciascun documento è rappresentato come una mistura latente di argomenti (*topic*) e ciascun topic come una distribuzione di probabilità sul vocabolario.

Dal punto di vista generativo, fissato a priori il numero K di topic, il modello assume la seguente procedura per la generazione di un corpus di M documenti: per ogni topic $k \in \{1, \dots, K\}$ si campiona una distribuzione sulle parole $\varphi_k \sim \text{Dir}(\beta)$; per ogni documento $d \in \{1, \dots, M\}$ si campiona una distribuzione sui topic $\theta_d \sim \text{Dir}(\alpha)$; infine, per ciascuna delle N_d parole del documento si estrae un'assegnazione latente $z_{i,d} \sim \text{Mult}(\theta_d)$ e, condizionata-

mente, la parola osservata $w_{i,d} \sim \text{Mult}(\varphi_{z_{i,d}})$. Le ipotesi sottostanti sono la rappresentazione BoW (irrilevanza dell'ordine delle parole) e la scambiabilità condizionata delle osservazioni [3]. La probabilità congiunta del modello si fattorizza come:

$$p(w, z, \theta, \varphi \mid \alpha, \beta) = \prod_{k=1}^K p(\varphi_k \mid \beta) \prod_{d=1}^M p(\theta_d \mid \alpha) \prod_{i=1}^{N_d} p(z_{i,d} \mid \theta_d) p(w_{i,d} \mid \varphi_{z_{i,d}}) \quad (2.1)$$

dove K è il numero di topic, M il numero di documenti del corpus, N_d la lunghezza in parole del documento d -esimo, φ_k il vettore delle probabilità delle parole nel topic k -esimo, θ_d il vettore delle proporzioni dei topic nel documento d , $z_{i,d}$ l'assegnazione latente del topic alla i -esima parola del documento d , $w_{i,d}$ la parola osservata nella corrispondente posizione, mentre α e β sono gli iperparametri delle due distribuzioni di Dirichlet a priori. Quest'ultima, di parametro $\alpha = (\alpha_1, \dots, \alpha_K)$, è definita sul simpleso come:

$$\text{Dir}(p \mid \alpha) = \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K p_j^{\alpha_j - 1} \quad (2.2)$$

dove $p = (p_1, \dots, p_K)$ è un vettore di probabilità appartenente al simpleso $(K - 1)$ -dimensionale, cioè tale che $p_j \geq 0$ per ogni j e $\sum_{j=1}^K p_j = 1$, $\alpha_j > 0$ sono gli iperparametri di concentrazione e $\Gamma(\cdot)$ denota la funzione Gamma di Eulero, estensione continua del fattoriale ai reali positivi. La scelta della Dirichlet come distribuzione a priori non è arbitraria: essa è la coniugata naturale della Multinomiale, proprietà che garantisce che la distribuzione a posteriori conservi la stessa forma funzionale a priori e che riduce l'aggiornamento bayesiano a una somma di pseudo-conteggi sugli iperparametri. Inoltre, impostando $\alpha < 1$ e $\beta < 1$, la massa di probabilità viene spinta verso i vertici del simpleso, inducendo sparsità: ciascun documento risulta concentrato su pochi topic dominanti e ciascun topic su un sottoinsieme ristretto e semanticamente coerente di termini.

Il principale ostacolo applicativo della LDA risiede nell'intrattabilità del calcolo esatto della distribuzione a posteriori $p(z, \theta, \varphi \mid w, \alpha, \beta)$, poiché l'evidenza marginale $p(w \mid \alpha, \beta)$ richiede una sommatoria sull'insieme combinatorio di tutte le possibili assegnazioni latenti, di cardinalità K^N , con $N = \sum_{d=1}^M N_d$ numero totale di token del corpus. Si ricorre pertanto a metodi di inferenza

2.2 Structural Topic Model

approssimata, riconducibili a due famiglie principali: l’inferenza variazionale (*Variational Bayes*), che approssima la posteriori reale con una famiglia parametrica più semplice minimizzando la divergenza di Kullback–Leibler (equivalentemente, massimizzando l’*Evidence Lower Bound* (ELBO)) già proposta nel lavoro originale [3]; e i metodi Markov Chain Monte Carlo, in particolare il *Collapsed Gibbs Sampling* di Griffiths e Steyvers [15], che marginalizza analiticamente θ e φ sfruttando la coniugazione e campiona iterativamente le sole assegnazioni $z_{i,d}$ fino al raggiungimento della distribuzione stazionaria della catena di Markov.

Nonostante il suo vastissimo utilizzo negli scorsi anni, la LDA presenta diverse problematiche che spesso si rivelano difficili da trascurare. Oltre alla già citata assunzione del BoW, che ignora le informazioni a livello di contesto, è necessario anche specificare a priori il numero di topic K , la cui scelta errata produce topic incoerenti se troppo basso o ridondanti se troppo elevato. Altra conseguenza della struttura gerarchica descritta in precedenza è l’assunzione di indipendenza indotta dalla distribuzione di Dirichlet, che è strutturalmente incapace di catturare le correlazioni tra gli argomenti (un’ipotesi difficilmente difendibile, dal momento che in un corpus reale i topic tendono naturalmente a sovrapporsi sul piano dei contenuti). A tutto ciò si aggiunge il costo computazionale elevato su corpora di grandi dimensioni e la forte dipendenza dei risultati da un’accurata fase di pre-elaborazione linguistica.

2.2 Structural Topic Model

Lo *Structural Topic Model* (STM), introdotto da Roberts, Stewart e Tingley [25], rappresenta un’estensione del framework della LDA e del CTM progettata per superare uno dei principali limiti nell’analisi empirica delle scienze sociali: l’incapacità di incorporare metadati a livello di documento all’interno del processo generativo. Nella pratica della ricerca, i documenti sono quasi invariabilmente accompagnati da informazioni supplementari: l’identità dell’autore, l’orientamento politico, la data di pubblicazione, l’appartenenza a un gruppo sperimentale, ecc. L’esclusione di queste caratteristiche dal model-

lo può introdurre errori di misurazione sistematici qualora si vogliano stimare gli effetti di tali covariate sui contenuti testuali. L'STM stima congiuntamente due dimensioni della produzione discorsiva: la *prevalenza tematica* (quanto spesso un argomento è trattato in un documento) e il *contenuto tematico* (con quale vocabolario specifico viene discusso).

Dal punto di vista generativo, fissato il numero K di topic, l'STM modella le proporzioni tematiche del documento d attraverso una distribuzione Logistico-Normale condizionata sul vettore di covariate X_d :

$$\vec{\theta}_d \mid X_d \gamma, \Sigma \sim \text{LogisticNormal}(\mu = X_d \gamma, \Sigma) \quad (2.3)$$

dove γ è una matrice di coefficienti di regressione di dimensioni $p \times (K - 1)$ e Σ è la matrice di covarianza $(K - 1) \times (K - 1)$ che cattura le correlazioni residue tra argomenti. Si osservi che il caso particolare in cui X_d si riduce alla sola intercetta (ovvero in assenza di covariate di prevalenza) corrisponde alla specificazione del *Correlated Topic Model* (CTM) di Blei e Lafferty [4, 5], predecessore diretto dell'STM, che generalizza la LDA sostituendo il prior di Dirichlet sulle proporzioni tematiche con una distribuzione Logistico-Normale, così da catturare le correlazioni residue tra argomenti attraverso la matrice Σ (assunzione più realistica rispetto alla quasi-indipendenza imposta dal prior di Dirichlet). Il contenuto tematico è invece modellato tramite un modello log-lineare ispirato all'architettura SAGE (*Sparse Additive Generative Model of Text*) [13], in cui la distribuzione sulle parole per l'argomento k nel documento d è:

$$\beta_{d,k,v} \propto \exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right) \quad (2.4)$$

dove y_d denota il livello assunto dalla covariata di contenuto per il documento d (ad esempio l'autore), m_v è la frequenza logaritmica di base della parola v , $\kappa_{k,v}^{(t)}$ è la deviazione specifica dell'argomento k , $\kappa_{y_d,v}^{(c)}$ cattura le preferenze lessicali globali del gruppo y_d , e $\kappa_{y_d,k,v}^{(i)}$ è il termine di interazione che misura come il gruppo y_d modifichi il proprio vocabolario specificamente in relazione all'argomento k . Questo ultimo termine costituisce lo strumento attraverso cui l'STM rende misurabile il fenomeno del *framing*: la tendenza di livelli differenti della covariata di contenuto a discutere uno stesso argomento ricorrendo a scelte lessicali sistematicamente diverse.

Poiché la distribuzione a posteriori esatta è intrattabile (per la non-coniugazione tra il prior Logistico-Normale e la verosimiglianza multinomiale) l'STM ricorre a un algoritmo di *Expectation-Maximization Variazionale Semi-Collassato* [25]. Nell'E-step (passo di *Expectation*: partendo dai parametri attuali del modello, si stima a quali topic appartengono i documenti), le variabili di assegnazione tematica a livello di token vengono integrate analiticamente, mentre la non-coniugazione del prior è gestita tramite un'approssimazione di Laplace. Nell'M-step (passo di Massimizzazione: date le assegnazioni tematiche appena stimate, si aggiornano i parametri del modello per renderle il più plausibili possibile), i parametri globali γ , κ e Σ vengono aggiornati con l'introduzione di prior regolarizzanti (di tipo L1 o L2) per prevenire il sovradattamento indotto dall'elevata dimensionalità parametrica. Una delle innovazioni metodologicamente più rilevanti dell'STM riguarda la stima degli effetti delle covariate sulla prevalenza tematica. L'approccio tradizionale (estrarre post-hoc le proporzioni $\hat{\theta}_d$ e utilizzarle come variabile dipendente in regressioni OLS) trascura l'incertezza di misurazione propria della stima tematica, producendo stime distorte [25]. L'STM ovvia a questo problema attraverso il *Method of Composition*: campionando ripetutamente dalla distribuzione variazionale a posteriori di $\vec{\theta}$ e γ , l'incertezza del modello generativo viene propagata all'interno della stima degli effetti medi, producendo intervalli di confidenza che incorporano sia la variabilità campionaria sia l'errore di misurazione. Infine, per l'etichettatura dei topic Roberts et al. [25] introducono il punteggio FREX (*Frequency-Exclusivity*), definito come media armonica ponderata tra il rango di frequenza intra-topic e il rango di esclusività lessicale di una parola rispetto agli altri argomenti. Rispetto ai ranking basati sulla sola probabilità marginale, il FREX evita che parole ad alta frequenza di corpus prive di specificità tematica dominino la rappresentazione dei topic, restituendo invece le *anchor words* semanticamente discriminanti.

Capitolo 3

BERTopic: architettura e pipeline

L'evoluzione del *topic modeling* ha attraversato una fase di transizione radicale con l'avvento del *deep learning* e, in particolare, con l'introduzione dei modelli linguistici basati sull'architettura *Transformer*. In questo panorama, BERTopic [16] si è imposto come framework di riferimento allo stato dell'arte. Sviluppato da Maarten Grootendorst, BERTopic non si configura come un singolo modello probabilistico generativo (come i modelli presentati in precedenza) bensì come un'architettura modulare e componibile che orchestra una pipeline sequenziale di algoritmi. L'innovazione centrale di questo framework risiede nel disaccoppiamento esplicito tra la fase di rappresentazione semantica del testo, la fase di raggruppamento spaziale e la fase di estrazione lessicale dei topic.

3.1 Motivazione: oltre il Bag-of-Words

I metodi classici di topic modeling, tra cui LDA, CTM e STM, fondano la propria impalcatura inferenziale sull'ipotesi statistica del BoW. In questo framework, un documento viene rappresentato come un vettore delle frequenze dei termini in uno spazio di dimensionalità pari alla cardinalità del vocabolario, trattando le parole come variabili casuali condizionatamente indipendenti. Questa assunzione di intercambiabilità comporta la completa distruzione

dell'ordine sequenziale dei token e, conseguentemente, della struttura sintattica e del contesto semantico. Dal punto di vista della statistica multivariata, l'approccio BoW introduce tre criticità strutturali. La prima è la sparsità estrema della matrice termine-documento: poiché un documento contiene solo una frazione del vocabolario totale, la matrice risultante è dominata da zeri, il che inficia l'efficienza degli stimatori statistici e rende i modelli vulnerabili al rumore. La seconda è l'ortogonalità semantica: nel modello vettoriale BoW ogni parola occupa una dimensione ortogonale indipendente, cosicché due sinonimi (es. “automobile” e “vettura”) vengono mappati su assi distinti e dunque la loro similarità risulta prossima a zero. La terza è l'incapacità di gestire la polisemia: un termine che assume significati opposti a seconda del contesto (es. “banco”) viene collassato in un'unica dimensione vettoriale, corrompendo la purezza semantica dei topic estratti.

Rappresentazioni contestuali

Il superamento del bag-of-words si è concretizzato con lo sviluppo delle rappresentazioni contestuali rese possibili dall'architettura *transformer* e dal meccanismo dell'attenzione. A differenza dei word embedding statici di prima generazione (Word2Vec, GloVe), che associano a ogni lemma un unico vettore denso in $\mathbb{R}^d$, i modelli basati su *Transformer* (BERT, RoBERTa, XLNet) calcolano rappresentazioni vettoriali dinamiche che variano in funzione del contesto. Il meccanismo abilitante è la *Scaled Dot-Product Attention*[30]: dati i vettori *Query* (Q), *Key* (K) e *Value* (V) derivati da ciascun token, l'attenzione contestuale è definita da:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (3.1)$$

dove d_k è la dimensione del vettore delle chiavi, impiegato come fattore di normalizzazione per stabilizzare il gradiente. L'adozione di queste rappresentazioni contestualizzate sovverte l'approccio classico: BERTopic non analizza la frequenza di singole stringhe di caratteri, ma la distribuzione topologica di vettori densi nello spazio semantico. Un documento che discute di dinamiche finanziarie sarà proiettato in una regione vicina a un documento sulle politi-

che monetarie, anche in assenza di intersezione lessicale diretta, producendo topic con una coerenza interna significativamente superiore.

3.2 La pipeline standard

L'innovazione metodologica di BERTopic consiste nel scomporre il problema del topic modeling in sotto-problemi distinti, affrontando ciascuno con l'algoritmo più idoneo. La pipeline standard si articola in cinque fasi sequenziali: embedding dei documenti, riduzione dimensionale, clustering, vectorization e pesatura c-TF-IDF.

3.2.1 Embedding dei documenti con Sentence-BERT

Il primo step della pipeline prevede la trasformazione del corpus testuale in una matrice di vettori densi tramite Sentence-BERT (SBERT), introdotto da Reimers e Gurevych [24]. SBERT adotta un'architettura a rete siamese: due sottoreti identiche con pesi condivisi processano ciascuna frase indipendentemente attraverso un encoder Transformer, producendo una rappresentazione token-level di dimensioni $N \times d$, dove N è la lunghezza della sequenza e d la dimensione nascosta del modello (tipicamente 768 per i modelli BERT-base).

Poiché è necessario un unico vettore documento di dimensione fissa, la rappresentazione token-level viene compressa tramite un'operazione di *pooling*. La strategia di default in BERTopic è il MEAN-pooling, che calcola la media aritmetica dei vettori di output contestualizzati:

$$\mathbf{u} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}_i \quad (3.2)$$

dove $\mathbf{h}_i \in \mathbb{R}^d$ è il vettore di output del token i -esimo. Gli embedding di due documenti $\mathbf{u}$ e $\mathbf{v}$ vengono poi confrontati tramite similarità del coseno, senza necessità di processarli congiuntamente (*late-interaction*).

I pesi del modello vengono ottimizzati tramite *fine-tuning* su dataset di similarità testuale. Una delle funzioni obiettivo più utilizzate è la *Triplet Loss*, che processa simultaneamente una frase di ancoraggio s_a , una frase positiva

s_p e una negativa s_n , minimizzando:

$$L_{TL} = \max(\|\mathbf{s}_a - \mathbf{s}_p\| - \|\mathbf{s}_a - \mathbf{s}_n\| + \varepsilon, 0) \quad (3.3)$$

dove $\varepsilon > 0$ è un margine che garantisce una separazione geometrica minima tra coppie semanticamente concordi e discordi. La qualità degli embedding prodotti da SBERT condiziona direttamente tutte le fasi successive della pipeline, poiché le relazioni semantiche del corpus vengono interamente codificate in questa rappresentazione vettoriale.

3.2.2 Riduzione dimensionale con UMAP

Gli embedding ad alta dimensionalità non possono essere impiegati direttamente in algoritmi di clustering basati sulla densità a causa della "maledizione" della dimensionalità: all'aumentare della dimensionalità dello spazio, il volume cresce esponenzialmente e i punti diventano sempre più sparsi, cosicché le distanze tra osservazioni tendono a uniformarsi e perdono capacità discriminante. BERTopic integra pertanto UMAP (*Uniform Manifold Approximation and Projection*), formulato da McInnes, Healy e Melville [20], un algoritmo non-lineare di riduzione della dimensionalità.

A differenza di PCA (*Principal Component Analysis*)[19], che proietta i dati su assi di massima varianza tramite trasformazioni lineari e risulta pertanto incapace di preservare strutture curvilinee o non convesse, UMAP opera una riduzione non-lineare che approssima la varietà geometrica sottostante ai dati. Questa proprietà gli consente di preservare simultaneamente la micro-struttura (la prossimità tra documenti semanticamente simili all'interno di un cluster) e la macro-struttura globale (le distanze relative tra topic distinti). L'algoritmo costruisce dapprima, nello spazio ad alta dimensionalità, un *fuzzy simplicial set*: un grafo di vicinanza pesato in cui la forza del legame tra due documenti x_i e x_j è definita da:

$$w_{i,j} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right) \quad (3.4)$$

dove $d(\cdot, \cdot)$ è una funzione di distanza definita sullo spazio degli embedding di partenza, ρ_i è la distanza dal vicino più prossimo di x_i (che garantisce

3.2.3 Clustering con HDBSCAN

la connettività anche dei punti più isolati) e σ_i è un fattore di riscaldamento adattivo calibrato in modo che la somma dei pesi dei vicini di x_i sia pari a $\log_2(k)$, con k il numero di vicini considerato. Poiché i pesi risultanti sono asimmetrici ($w_{i,j} \neq w_{j,i}$), UMAP li simmetrizza tramite un operatore di unione sfocata:

$$P_{ij} = w_{i,j} + w_{j,i} - w_{i,j} \cdot w_{j,i} \quad (3.5)$$

La proiezione in bassa dimensionalità avviene minimizzando la *Fuzzy Cross-Entropy* tra le probabilità originali P_{ij} e le corrispondenti probabilità Q_{ij} nel sottospazio ridotto:

$$C = \sum_{i \neq j} \left[P_{ij} \log \left(\frac{P_{ij}}{Q_{ij}} \right) + (1 - P_{ij}) \log \left(\frac{1 - P_{ij}}{1 - Q_{ij}} \right) \right] \quad (3.6)$$

Il primo termine introduce forze di attrazione tra punti vicini nello spazio originale; il secondo introduce forze di repulsione tra punti lontani. È proprio quest'ultimo termine che garantisce la conservazione della struttura globale del corpus, impedendo la fusione di cluster semanticamente distinti.

3.2.3 Clustering con HDBSCAN

Ottenuta una rappresentazione spaziale a bassa dimensionalità, BERTopic aggrega i documenti in cluster tramite HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*), proposto da Campello, Moulavi e Sander [7]. La scelta di HDBSCAN rispetto ad algoritmi partizionali come K-Means è motivata da tre vantaggi strutturali: non richiede la specificazione a priori del numero di cluster; è in grado di identificare cluster di forma arbitraria; classifica esplicitamente i punti di rumore (etichetta -1), evitando che documenti non riconducibili a nessun topic vengano forzatamente assegnati a un argomento.

Il punto di partenza è la *Mutual Reachability Distance* tra due punti a e b :

$$d_{\text{mreach}}(a, b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a, b)\} \quad (3.7)$$

dove $\text{core}_k(x)$ è la distanza tra x e il suo k -esimo vicino più prossimo, detta *core distance*. La formulazione con il massimo persegue un duplice scopo: se entrambi i punti si trovano in regioni dense (core distance piccola), la distanza

effettiva $d(a, b)$ domina il calcolo, preservando la geometria locale; se invece uno dei due è periferico (core distance elevata), la metrica si dilata proporzionalmente, scoraggiando la formazione di connessioni spurie tra regioni a densità differente. In questo modo i punti isolati vengono tenuti a distanza dai cluster principali finché la soglia di densità non scende sufficientemente.

Su di essa HDBSCAN costruisce il *Minimum Spanning Tree* (MST) del grafo completo in cui ogni nodo rappresenta un documento e ogni arco ha peso pari alla *mutual reachability* distance tra i due punti connessi [7]. Dato un grafo pesato con n nodi, l'MST è il sottografo aciclico che collega tutti i nodi con esattamente $n - 1$ archi minimizzando la somma totale dei pesi: esso costituisce lo scheletro della distribuzione dei dati, conservando soltanto le connessioni strutturalmente necessarie ed eliminando quelle ridondanti. La proprietà chiave ai fini del clustering è che gli archi a peso elevato dell'MST corrispondono ai punti di separazione naturale tra gruppi: rimuovendoli in ordine decrescente di peso la struttura si frammenta progressivamente nei suoi componenti più coesi.

Il MST viene quindi condensato in una gerarchia di cluster: abbassando progressivamente una soglia di densità $\lambda = 1/d$, ogni scissione dell'albero genera due sotto-cluster figli. Questo processo prosegue fino alla dissoluzione nei singoli punti, producendo un dendrogramma completo da cui HDBSCAN estrae i cluster più stabili attraverso la persistenza λ , ovvero la differenza tra il valore di λ in cui il cluster nasce e quello in cui scompare.

L'estrazione del partizionamento finale da questa gerarchia avviene secondo il principio dell'*Excess of Mass* (EoM): tra tutti i possibili insiemi di cluster annidati, viene selezionato quello che massimizza la *persistenza complessiva* dei cluster scelti rispetto alle soglie di densità. L'idea è che un cluster autenticamente rappresentativo della struttura dei dati debba mantenersi stabile lungo un ampio intervallo di livelli di densità, mentre frammentazioni transitorie indicano sotto-strutture meno significative.

Formalmente, a ogni cluster C della gerarchia è associato un livello $\lambda_{\text{birth}}(C)$, in corrispondenza del quale esso si è separato dal proprio cluster genitore, e a ogni punto $p \in C$ un livello λ_p , in corrispondenza del quale p ha lasciato

3.2.4 Vectorization con CountVectorizer

C (entrando in un sotto-cluster o cadendo nel rumore). La *stabilità* di C è allora definita come:

$$S(C) = \sum_{p \in C} (\lambda_p - \lambda_{\text{birth}}(C)) \quad (3.8)$$

e quantifica la "massa" accumulata dal cluster nell'intervallo di densità in cui è rimasto coeso. L'algoritmo percorre quindi la gerarchia condensata dal basso verso l'alto: per ogni nodo confronta la stabilità del cluster padre con la somma delle stabilità dei figli, conservando il padre se $S(\text{padre}) \geq \sum_i S(\text{figlio}_i)$ (ovvero quando la coesione complessiva è meglio descritta da un singolo cluster) e selezionando altrimenti i figli. Il risultato è un partizionamento piatto dei dati che combina la robustezza di un'analisi multi-scala con l'interpretabilità di un'assegnazione discreta, e che ammette esplicitamente l'esistenza di punti non assegnati ad alcun cluster (rumore).

3.2.4 Vectorization con CountVectorizer

Le procedure descritte fino a HDBSCAN forniscono una categorizzazione topologica dei documenti, ma non una descrizione lessicale comprensibile dei topic. BERTopic affronta questo problema tramite una macro-aggregazione post-hoc: tutti i documenti assegnati a uno stesso cluster vengono concatenati in un unico documento rappresentativo della classe. A questo punto, il sistema si interfaccia con metriche frequentistiche classiche costruendo la matrice termine-classe delle frequenze: per ciascun cluster, i documenti aggregati vengono tokenizzati e i conteggi dei termini vengono accumulati in un vettore di frequenze. Crucialmente, questa operazione avviene a valle del clustering, il che significa che le scelte sul vocabolario (come un'eventuale rimozione di parole vuote) non influenzano in alcun modo la proiezione geometrica effettuata nelle fasi precedenti.

3.2.5 Pesatura con c-TF-IDF

Per identificare i termini più rappresentativi di ciascun topic, Grootendorst ha sviluppato il c-TF-IDF (*Class-based Term Frequency - Inverse Document Frequency*) [16]. Mentre il TF-IDF classico opera a livello di singolo documento, la versione class-based traspone l'analisi al livello di cluster. Il peso

$W_{t,c}$ di un termine t nel topic c è definito come:

$$W_{t,c} = \text{tf}_{t,c} \times \log\left(1 + \frac{A}{\text{tf}_t}\right) \quad (3.9)$$

dove $\text{tf}_{t,c}$ è la frequenza normalizzata (norma L1) del termine nel macro-documento del cluster c , tf_t è la frequenza cumulata del termine in tutti gli altri cluster, e A è la media del numero di parole per classe, che funge da fattore di regolarizzazione. La normalizzazione L1 di $\text{tf}_{t,c}$ è fondamentale per correggere l'asimmetria indotta da cluster di dimensioni molto diverse, convertendo le frequenze assolute in proporzioni distribuzionali comparabili tra topic.

3.2.6 Rappresentazione finale dei topic

Il c-TF-IDF produce per ciascun topic un ranking decrescente di termini per peso $W_{t,c}$. BERTopic seleziona ordinariamente i primi termini come etichette rappresentative del cluster. La natura modulare della pipeline permette di raffinare ulteriormente questa rappresentazione senza riprocessare le fasi precedenti. Il modulo *Maximal Marginal Relevance* (MMR) penalizza iterativamente i termini ridondanti, selezionando parole che massimizzano congiuntamente la rilevanza per il topic e la diversità semantica reciproca, misurata tramite similarità del coseno. In alternativa, modelli generativi (BART, T5, o LLM tramite API) possono essere interrogati sui documenti più rappresentativi del cluster per produrre etichette testuali discorsive, elevando l'interpretabilità del risultato per l'analista.

In sintesi, la pipeline modulare di BERTopic combina la profondità semantica delle rappresentazioni contestuali di Sentence-BERT, la topologia simmetrizzata di UMAP, la densitometria adattiva di HDBSCAN e la discriminazione lessicale del c-TF-IDF, definendo un framework scalabile e teoricamente fondato per l'estrazione di topic su corpora testuali di grandi dimensioni.

3.3 Punti di forza e limiti

La pipeline descritta nelle sezioni precedenti conferisce a BERTopic vantaggi strutturali netti rispetto ai modelli classici, frutto di una combinazione di

tecnologie che operano sequenzialmente e che massimizzano la ritenzione delle informazioni semantiche lungo l'intera catena di elaborazione.

Il vantaggio più rilevante risiede nell'utilizzo di rappresentazioni contestualizzate prodotte da modelli Transformer. A differenza dei modelli tradizionali basati sul BoW (come LDA o STM) che quantificano la co-occorrenza dei termini ignorando l'ordine delle parole e il contesto semantico, i Transformer analizzano il testo bidirezionalmente tramite il meccanismo di *self-attention*, consentendo di disambiguare il significato di ciascun vocabolo in funzione delle parole circostanti. Questa profondità semantica permette a BERTopic di raggruppare documenti che discutono del medesimo concetto anche in assenza di sovrapposizione lessicale diretta, superando la rigidità di una esatta corrispondenza lessicale. La superiorità di questo approccio risulta particolarmente marcata nel trattamento di testi brevi e non strutturati (come post sui social media o recensioni) dove modelli come LDA collassano metodologicamente per insufficienza di co-occorrenze affidabili.

Un secondo punto di forza è la modularità architetturale del framework [16]. BERTopic disaggrega il processo di estrazione dei topic in fasi indipendenti e interscambiabili, consentendo di sostituire liberamente ciascun componente in funzione del dominio applicativo, della lingua del corpus o delle risorse computazionali disponibili. Questa flessibilità permette al framework di evolvere in sincronia con la scoperta di nuove tecniche metodologiche e con i rapidi progressi dell'intelligenza artificiale. A ciò si aggiunge la capacità di modellare i topic in senso dinamico, tracciando la deriva concettuale dei temi nel tempo, e di incorporare un meccanismo di riduzione gerarchica che consente di fondere argomenti semanticamente affini, offrendo un controllo granulare sulla struttura tematica estratta.

Nonostante questi vantaggi, BERTopic presenta limiti strutturali di rilievo che ne condizionano l'applicabilità, specialmente in contesti di ricerca rigorosa.

Il principale limite sul piano teorico è la natura di *hard clustering*: ogni documento viene assegnato a un unico topic in modo esclusivo, secondo una logica di partizionamento geometrico che risulta intrinsecamente riduzioni-

sta se applicata a testi complessi (quali articoli accademici, contratti legali o resoconti clinici) i quali per loro natura trattano simultaneamente più argomenti in proporzioni variabili. Nei modelli generativi, al contrario, ogni documento è rappresentato come una miscela probabilistica di topic latenti [3], una prospettiva che riflette con maggiore fedeltà la realtà discorsiva dei testi articolati. Sebbene BERTopic abbia introdotto funzioni post-hoc come *approximate_distribution()* per generare distribuzioni approssimate per documento, si tratta di euristiche retroattive che non colmano l'assenza di un vero modello probabilistico nativo.

Un secondo limite, strettamente connesso alla natura di HDBSCAN [7], riguarda la gestione degli outlier. L'algoritmo classifica come rumore tutti i documenti che non rientrano in aree di densità geometrica sufficiente, assegnando loro l'etichetta convenzionale -1 ed escludendoli dall'analisi primaria. Nei corpus reali questa quota può risultare consistente, determinando una perdita informativa non trascurabile. Le strategie di riduzione degli outlier messe a disposizione dal framework permettono di recuperare parzialmente questi documenti, ma ciascuna introduce propri compromessi in termini di accuratezza semantica e costo computazionale.

Un terzo limite riguarda la tendenza del modello a generare un numero molto elevato di topic. Lasciato alle configurazioni predefinite, BERTopic produce frequentemente molti micro-cluster ridondanti e iper-frammentati (fenomeno noto come *over-splitting*) rendendo i risultati difficilmente interpretabili e richiedendo un intervento manuale per consolidare tematiche artificialmente separate. Nonostante BERTopic metta a disposizione la funzione *reduce_topics()*, che consente di raggruppare automaticamente i topic in base alla similarità coseno tra le rispettive rappresentazioni vettoriali, questo strumento non sempre è in grado di eliminare il problema.

Sul piano della riproducibilità, BERTopic sconta una vulnerabilità significativa legata all'instabilità stocastica introdotta da UMAP nella fase di riduzione dimensionale [20]. Poiché l'ottimizzazione del layout manifold dipende da una inizializzazione casuale, esecuzioni distinte dello stesso modello sullo stesso corpus possono produrre mappature divergenti, con cluster che si scompongono, si aggregano o scompaiono tra una computazione e l'altra. Il ricorso

a seed fissi attenua il problema ma non lo risolve alla radice, mascherando l'incertezza intrinseca che caratterizza i confini semantici reali dei cluster.

Sul piano computazionale, la complessità quadratica del meccanismo di *self-attention* dei Transformer può rendere l'analisi su corpus estesi estremamente onerosa, mentre HDBSCAN richiede un dispendio di memoria considerevole per il calcolo delle distanze a coppie e la costruzione degli alberi gerarchici. A differenza di LDA e STM, che operano su computazioni matriciali ottimizzate e risultano significativamente più efficienti su CPU, BERTopic richiede infrastrutture specializzate che ne limitano l'accessibilità in contesti con risorse limitate.

Infine, emerge un paradosso architetturale nella fase di rappresentazione dei topic. Nonostante l'investimento computazionale iniziale sugli embedding, la generazione delle etichette descrittive dei cluster ricade sul c-TF-IDF [16], una variante del *bag-of-words* che non eredita la capacità di disambiguazione contestuale prodotta nella fase di embedding. Le descrizioni tematiche risultanti sono pertanto insensibili al contesto e vulnerabili all'emergere di termini semanticamente vuoti o interferenze da metadati sistematici (come nomi di autori in corpus scientifici) che opacizzano l'interpretabilità dei risultati e richiedono pre-elaborazioni esterne non supportate nativamente dall'architettura del framework.

Questi limiti motivano le sperimentazioni presentate nel capitolo successivo, in cui si esplorano modifiche alla pipeline di BERTopic con l'obiettivo di provare a migliorare le prestazioni sul corpus italiano in esame.

Capitolo 4

Estensioni e varianti di BERTopic proposte

4.1 Modelli di embedding alternativi

BERTopic utilizza di default il modello *all-MiniLM-L6-v2* per corpora in lingua inglese e *paraphrase-multilingual-MiniLM-L12-v2* per scenari multilingue. Quest’ultimo è in grado di allineare semanticamente testi in oltre 50 lingue all’interno di uno spazio vettoriale condiviso a 384 dimensioni, risultando particolarmente efficace su corpora eterogenei dal punto di vista linguistico. Data la natura prevalentemente italiana dei dati in esame, si è scelto di sperimentare inoltre il modello *dbmdz/bert-base-italian-uncased* [10], disponibile sulla piattaforma Hugging Face. Si tratta di un modello BERT addestrato su un corpus monolingue italiano, in grado di produrre rappresentazioni vettoriali più aderenti alle specificità morfologiche e lessicali della lingua rispetto ai modelli multilingue generalisti.

4.2 Riduzione dimensionale alternativa

UMAP, metodo predefinito di BERTopic, presenta limiti strutturali su corpus di dimensioni ridotte: le sue assunzioni topologiche decadono sotto le 100–500

istanze, producendo cluster fittizi e risultati non riproducibili a causa della natura stocastica dell'ottimizzazione via SGD (*Stochastic Gradient Descent*).

Come alternativa deterministica è stata considerata la PCA, che proietta i dati lungo le direzioni di massima varianza mediante decomposizione spettrale della matrice di covarianza:

$$C = \frac{1}{n}X^T X, \quad C = V\Lambda V^T \quad (4.1)$$

con X matrice di dati centrati. La PCA garantisce piena riproducibilità e complessità computazionale contenuta, ma assume relazioni lineari tra le variabili, rischiando di appiattare strutture geometriche non lineari tipiche degli embedding neurali.

La *Kernel PCA* (KPCA) supera questo limite applicando il kernel trick per eseguire la decomposizione spettrale nello spazio delle caratteristiche indotto dalla funzione kernel:

$$K\alpha = \lambda\alpha \quad (4.2)$$

L'impiego di un kernel a Funzione di Base Radiale (RBF), che misura la similarità tra documenti tramite una funzione gaussiana della loro distanza nello spazio degli embedding, consente di preservare la curvatura geometrica di questi ultimi senza costruire esplicitamente lo spazio ad alta dimensionalità, combinando la robustezza non lineare di UMAP con il determinismo della PCA. Il principale svantaggio è la complessità $O(n^3)$ legata alla decomposizione spettrale della matrice di Gram, che ne limita l'uso su corpora molto grandi.

4.3 Metodi di clustering alternativi

Il principale limite di HDBSCAN in contesti operativi reali è la gestione del rumore: l'algoritmo può classificare come outlier anche più della metà dei documenti, escludendoli dall'analisi tematica. A ciò si aggiunge l'elevata sensibilità ai parametri di inizializzazione (quali il numero minimo di documenti

4.4 Metodo di rappresentazione alternativo

per ogni gruppo, *min_cluster_size*, o la densità minima richiesta affinché un punto venga considerato un punto core, *min_samples*) la cui calibrazione errata produce variazioni drastiche nella granularità dei cluster.

Come alternativa è stato considerato *K-Means*, che minimizza la somma delle distanze euclidee al quadrato dai centroidi:

$$J = \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 \quad (4.3)$$

dove k è il numero di cluster fissato a priori, S_i è l'insieme dei documenti assegnati al cluster i -esimo, $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$ è il centroide del cluster i -esimo e $\|\cdot\|$ denota la norma euclidea.

K-Means garantisce l'assegnazione di ogni documento a un cluster (nessun outlier), un controllo diretto sulla granularità tramite il parametro k e maggiore stabilità. La necessità di specificare k a priori rappresenta tuttavia una criticità non trascurabile, poiché una scelta errata può produrre partizioni semanticamente incoerenti. Per individuare il valore ottimale si è ricorso all'indice di Silhouette, che per ciascun punto x_i è definito come:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4.4)$$

dove $a(i)$ è la distanza media di x_i dagli altri punti del proprio cluster e $b(i)$ è la distanza media minima verso i punti del cluster più vicino. Il valore di k scelto corrisponde a quello che massimizza l'indice di Silhouette medio sull'intero corpus.

4.4 Metodo di rappresentazione alternativo

Il metodo predefinito *c-TF-IDF* presenta due limiti principali: la dipendenza da un corpus di riferimento ampio e stabile per il calcolo dell'IDF globale, e l'estrazione di soli unigrammi, che fatica a descrivere concetti complessi senza ricorrere a n-grammi con costi computazionali crescenti.

In alternativa è stato proposto l'algoritmo *RAKE* (Rapid Automatic Keyword Extraction), che opera sulla struttura interna del testo in modo indipen-

dente dalla distribuzione globale del corpus. RAKE assegna a ciascun termine un punteggio basato sul rapporto tra grado di co-occorrenza e frequenza:

$$\text{Score}(w) = \frac{\text{deg}(w)}{\text{freq}(w)} \quad (4.5)$$

I principali vantaggi di RAKE rispetto a c-TF-IDF sono: l'indipendenza statistica dal corpus, l'estrazione nativa di espressioni multi-parola senza espansione del vocabolario, e la leggerezza computazionale che lo rende adatto a sistemi con aggiornamenti frequenti del modello tematico. RAKE incarna inoltre un approccio alternativo alla rilevazione delle keyword più rilevanti: anziché affidarsi alla frequenza relativa dei termini (come fa c-TF-IDF), misura la centralità di ciascuna parola tramite il suo grado in un grafo di co-occorrenza, ossia il numero di termini distinti con cui essa co-occorre all'interno delle frasi del testo. Parole che fungono da snodi concettuali nella rete lessicale risultano così privilegiate rispetto a quelle semplicemente frequenti.

A differenza dei metodi precedentemente descritti, RAKE non dispone di una implementazione nativa all'interno del framework BERTopic e ha richiesto pertanto lo sviluppo di un modulo di integrazione personalizzato.

4.5 Integrazione di Gibbs Sampling in BERTopic

Un limite strutturale di BERTopic è l'assenza di un meccanismo inferenziale probabilistico esplicito: le assegnazioni documento-topic prodotte dal clustering sono deterministiche e prive di una stima della distribuzione di probabilità sottostante, rendendo il modello fragile in presenza di dati sparsi, rumorosi o di dimensioni ridotte. Per affrontare queste criticità, ci si è ispirati all'approccio *Gibbs-BERTopic* proposto da Zhu e Liu [32], che integra il Gibbs sampling all'interno della pipeline di BERTopic per raffinare iterativamente le assegnazioni tematiche iniziali.

L'implementazione si articola in tre fasi. Nella prima, BERTopic viene eseguito nella sua configurazione standard per ottenere un'assegnazione iniziale dei documenti ai topic e la corrispondente matrice documento-termine. I contatori fondamentali del campionatore vengono inizializzati a partire da questa

assegnazione: n_k è il numero di documenti assegnati al topic k , n_{kw} il numero di occorrenze della parola w nel topic k , e n_{k+} il totale di parole assegnate al topic k . I documenti classificati come outlier (etichetta -1) vengono redistribuiti uniformemente tra i topic validi.

Nella seconda fase viene eseguito il Gibbs sampling secondo un modello mistura Dirichlet-Multinomiale (DMM), in cui a ciascun documento viene assegnato un unico topic. Ad ogni iterazione, per ciascun documento d il topic corrente viene rimosso dalle statistiche (*exclude-self*), dopodiché viene campionato un nuovo topic dalla distribuzione a posteriori:

$$P(Z_d = k \mid \mathbf{Z}_{-d}, \mathbf{W}, \alpha, \beta) \propto (n_k + \alpha) \cdot \prod_{w \in d} \left(\frac{n_{kw} + \beta}{n_{k+} + V\beta} \right)^{c_{dw}} \quad (4.6)$$

dove α e β sono i parametri di concentrazione delle distribuzioni di Dirichlet, V la dimensione del vocabolario e c_{dw} il conteggio della parola w nel documento d . Per garantire stabilità numerica, il calcolo della verosimiglianza viene eseguito in log-spazio tramite:

$$\log \phi_{kw} = \log(n_{kw} + \beta) - \log(n_{k+} + V\beta) \quad (4.7)$$

Completate le iterazioni, vengono stimate la distribuzione topic-parola ϕ e la distribuzione documento-topic θ :

$$\phi_{kw} = \frac{n_{kw} + \beta}{n_{k+} + V\beta}, \quad \theta_{dk} = \frac{\mathbb{K}[Z_d = k] + \alpha}{1 + K\alpha} \quad (4.8)$$

Nella terza fase, la matrice θ (che ora codifica le probabilità di appartenenza tematica di ciascun documento) viene utilizzata come nuovo spazio di rappresentazione: su di essa viene applicata una seconda istanza di UMAP con metrica di Hellinger, appropriata per distribuzioni di probabilità, seguita da un secondo passaggio di HDBSCAN per il riclustering finale. Le nuove assegnazioni vengono infine reintegrate nel modello BERTopic, aggiornando le rappresentazioni c-TF-IDF dei topic.

I parametri adottati nell'implementazione sono: numero di iterazioni $\text{iter} = 5$, $\alpha = 0.1$, $\beta = 0.01$, in linea con le indicazioni del paper di riferimento [\[32\]](#).

Capitolo 5

Dati: i musei scientifici italiani su social media

5.1 Contesto e motivazione del corpus

Negli ultimi due decenni l'avvento dei social network ha trasformato il modo in cui istituzioni e organizzazioni comunicano con il proprio pubblico, e i musei non hanno fatto eccezione: le piattaforme digitali si sono affermate come un canale ordinario per ampliare il bacino di utenza e rendere più continuativa la relazione con i visitatori. Facebook, ad esempio, resta una delle piattaforme su cui i musei italiani mantengono una forte presenza, grazie a un formato che consente una comunicazione efficace attraverso post brevi, eventi, dirette e contenuti multimediali di diversa natura.

Studiare cosa comunicano i musei scientifici italiani (e non soltanto quanto siano performanti in termini di engagement) richiede di confrontarsi con un corpus testuale di grandi dimensioni, prodotto in modo non strutturato e privo di una linea editoriale uniforme. È in questo contesto che il *topic modeling* trova una giustificazione metodologica forte: si tratta di una famiglia di tecniche non supervisionate capaci di far emergere queste aree tematiche latenti che compongono la produzione comunicativa di un'organizzazione. Applicato a un corpus eterogeneo come quello dei musei scientifici, il topic

modeling permette di descrivere in modo riproducibile l'architettura tematica della comunicazione, di confrontare l'identità editoriale dei singoli musei e di osservarne l'evoluzione nel tempo.

I dati utilizzati provengono da *CrowdTangle*, la piattaforma di Meta dedicata, fino alla sua dismissione nell'agosto 2024, all'analisi dei contenuti pubblici su Facebook e Instagram per finalità di ricerca accademica e giornalistica [9]. L'archivio offerto da CrowdTangle copre tutti i post pubblicati da pagine pubbliche, gruppi e profili verificati che abbiano superato la soglia di 110.000 like o che siano stati esplicitamente aggiunti a una lista. [14]. Nello studio in questione sono stati recuperati 5355 record relativi a comunicazioni digitali di nove musei scientifici italiani. I testi sono prevalentemente in italiano, con la presenza sporadica di termini inglesi (terminologia tecnica, nomi di eventi) e di occasionali post integralmente in altre lingue. Per ciascun post sono disponibili, oltre al testo, una serie di metadati di interazione (like, reazioni, condivisioni, commenti, visualizzazioni video) e di contesto (pagina, data, tipologia di contenuto, *overperforming score*) che, come si vedrà nei paragrafi successivi, sono stati impiegati come covariate nel modello STM.

5.2 Preprocessing

Il corpus iniziale è stato costruito unendo i post Facebook delle nove pagine relative a musei scientifici italiani più importanti: Museo Galileo (Firenze), Museo Nazionale della Scienza e della Tecnologia "Leonardo da Vinci" (Milano), Fondazione Scienza e Tecnica (Firenze), Museo "A come Ambiente" — MAcA (Torino), Immaginario Scientifico (Trieste), Polo culturale MIMIT — Museo storico della comunicazione (Roma), INAF — Osservatorio Astronomico di Brera, Museo La Specola dell'INAF — Osservatorio Astronomico di Padova e Museo Marconi di Bologna. Per ciascun post sono state ritenute le variabili relative al nome del Museo, alla data di pubblicazione, alla tipologia di contenuto, all'insieme delle interazioni e ai tre campi testuali disponibili (*Message*, *Image Text*, *Description*).

5.2.1 *Preprocessing* per STM e LDA

Costruzione della variabile testuale Poiché ciascun post può presentare contenuto testuale in più campi (il *messaggio* digitato dall'utente, il testo *embedded* in un'immagine o la *description* di un link condiviso) è stata costruita una variabile unica *Text* selezionando, per ciascun post, il più lungo dei tre campi disponibili. Questa scelta privilegia la maggior densità informativa testuale e riduce i casi in cui il messaggio può risultare vuoto o estremamente sintetico (es. soli emoji o hashtag) a fronte di una descrizione del link assai più articolata.

Rimozione del museo Marconi A seguito di un'ispezione preliminare del corpus, la pagina relativa al Museo Marconi è stata esclusa dall'analisi a causa del numero esiguo di post (solamente 2) e della scarsa rappresentatività rispetto al perimetro tematico considerato. Il corpus effettivamente impiegato nell'analisi comprende quindi post di otto pagine museali.

Aggregazione delle due pagine INAF Le due pagine afferenti all'Istituto Nazionale di Astrofisica (l'Osservatorio Astronomico di Brera e il Museo La Specola dell'Osservatorio di Padova) sono state aggregate in un'unica categoria *INAF (Brera/Padova)*. La motivazione principale è di natura quantitativa: considerate separatamente, le due pagine presentavano un numero di post relativamente contenuto, tale da rendere meno stabile la stima degli effetti museo-specifici nel modello STM. A questa considerazione si aggiunge un argomento di omogeneità tematica: entrambe le pagine afferiscono all'INAF e condividono una linea editoriale incentrata su contenuti astronomici e di divulgazione scientifica, rendendo l'aggregazione coerente sul piano dei contenuti oltre che vantaggiosa sul piano statistico. Le rimanenti pagine, tematicamente distinte tra loro e con una numerosità di post adeguata, sono state conservate come categorie autonome.

5.2.1 *Preprocessing* per STM e LDA

Per i due modelli *bag-of-words* (LDA e STM) la pulizia testuale e la lemmatizzazione sono fasi cruciali, poiché entrambi i metodi operano su matrici documento-termine (DTM) e la qualità del vocabolario condiziona diretta-

mente l'interpretabilità dei topic. La pipeline implementata è articolata nei passaggi seguenti.

Normalizzazione e pulizia del testo Ciascun documento è stato sottoposto a una procedura standard di pulizia, con rimozione di URL ed emoji e normalizzazione del testo per uniformare varianti tipografiche e spaziature.

Filtro linguistico Sui testi così ripuliti è stato applicato l'algoritmo di *language detection* di Google Compact Language Detector 2 (cld2), trattenendo i documenti classificati come italiani e quelli per i quali la lingua non è stata determinata (tipicamente testi molto brevi, di norma comunque in italiano). I post in altre lingue sono stati esclusi.

Lemmatizzazione e *tagging* Come tipico nel *topic modeling* in lingua italiana, ai documenti è stata applicata la procedura di lemmatizzazione. I documenti sono stati annotati con UDPipe [29], utilizzando il modello pre-addestrato per l'italiano.¹ Sono stati trattenuti i soli token con *part of speech* sostantivo, aggettivo, verbo o nome proprio, in linea con la prassi consolidata nella letteratura di *topic modeling* di mantenere le sole classi a contenuto semantico pieno [6]. Oltre alle parole vuote rimosse tramite la lista standard, sono stati eliminati quattro ulteriori lemmi ad altissima frequenza e bassa specificità tematica: i verbi *essere*, *avere* e *fare*, scarsamente informativi sul piano dei contenuti, e il sostantivo *museo*, che, ricorrendo in pressoché ogni post come auto-riferimento istituzionale, non contribuisce alla discriminazione tra topic.

Pruning del vocabolario e dei documenti Sulla DTM è stata condotta un'analisi preliminare volta a individuare due soglie di selezione: una sulla frequenza minima dei termini nel corpus e una sulla lunghezza minima dei documenti. Sulla base dei diagrammi di *vocabulary retention* e *document retention* sono state scelte le soglie di 15 occorrenze globali minime per termine

¹Nello specifico, è stato impiegato il modello *italian-isdt-ud-2.5-191206.udpipe*, addestrato sull'Italian Stanford Dependency Treebank (ISDT) nell'ambito del progetto Universal Dependencies, versione 2.5 (rilasciato il 6 dicembre 2019).

5.2.2 Preprocessing per BERTopic

e 25 lemmi minimi per documento. La prima rimuove i termini troppo rari per stimare in modo affidabile la loro probabilità di appartenenza a un topic, riducendo al contempo la dimensionalità della matrice; la seconda esclude i post troppo corti, per i quali la distribuzione documento-topic θ risulterebbe instabile. Al termine di questa selezione la DTM risulta composta da 3361 documenti e 2756 lemmi.

5.2.2 Preprocessing per BERTopic

Diversamente da LDA e STM, BERTopic non opera su una *bag-of-words* ma su rappresentazioni vettoriali dense generate da un *sentence encoder* basato su Transformer. Tali rappresentazioni codificano il significato del testo a partire dal contesto linguistico, e per produrle in modo efficace richiedono il testo nella sua forma il più possibile naturale: la punteggiatura, le parole funzionali, le forme flesse e perfino l'ordine dei token contribuiscono al contesto e quindi alla qualità dell'embedding. Applicare un encoder pre-addestrato su un corpus già preprocessato, equivarrebbe a sottrargli proprio le informazioni di base su cui il modello è stato addestrato, con un degrado prevedibile della qualità delle rappresentazioni.

Per questa ragione a BERTopic è stato fornito il testo grezzo, privo di lemmatizzazione e di rimozione di parole vuote.

Per garantire la confrontabilità dei risultati con LDA e STM, il modello è stato però addestrato sullo stesso sottoinsieme di documenti selezionati al termine del preprocessing descritto sopra (sopravvissuti cioè al filtro sulla lunghezza minima): le tre tecniche operano quindi sullo stesso corpus, differendo unicamente nella rappresentazione testuale fornita in input.

5.3 Metadati disponibili per STM

Uno dei vantaggi principali del modello STM rispetto ad altre tecniche di *topic modeling* è la possibilità di incorporare metadati a livello di documento come covariate nella stima della prevalenza dei topic, modellando esplicitamente come la distribuzione θ (proporzione di ciascun topic all'interno di un

documento) vari in funzione di variabili osservabili [26]. Per il presente corpus sono state utilizzate quattro covariate, tutte ricavate dai campi forniti da CrowdTangle.

Anno di pubblicazione Estratto dalla data di creazione del post, è incluso come variabile numerica modellata tramite spline cubica con tre gradi di libertà. La scelta della spline, anziché di un termine lineare o di un fattore categoriale, consente di catturare in modo flessibile eventuali andamenti non monotoni della prevalenza tematica nel tempo, senza imporre una forma funzionale a priori.

Museo Variabile categorica a sette livelli, costruita ricodificando il nome originale della pagina secondo lo schema descritto in 5.2. Permette di stimare in che misura ciascun museo si caratterizzi per una distribuzione tematica specifica rispetto al corpus complessivo.

Tipologia di contenuto Variabile categorica che originariamente distingue tra link, foto, video nativi, video YouTube, dirette e altre forme. Per evitare livelli scarsamente popolati e ridurre la dimensionalità del fattore, le quattro modalità relative a contenuti video sono state aggregate in un'unica categoria *Video*. La codifica finale distingue dunque tra link, foto, status e video.

Totale interazioni Variabile numerica corrispondente alla somma di tutte le interazioni ricevute dal post (like, reazioni, commenti, condivisioni). È inclusa nel modello tramite spline cubica con quattro gradi di libertà, così da modellare in modo flessibile l'eventuale relazione non lineare tra la performance del post e la composizione tematica (un'ipotesi rilevante per indagare se alcuni topic siano sistematicamente associati a un *engagement* più elevato).

Capitolo 6

Metriche di confronto e Disegno sperimentale

La valutazione dei modelli di topic modeling si è storicamente basata su misure puramente predittive derivate dalla teoria dell'informazione, in particolare la *perplexity* e la *held-out log-likelihood*, che quantificano la capacità del modello di prevedere le parole di documenti non visti. Indagini empiriche hanno però mostrato che tali misure sono spesso negativamente correlate con l'interpretabilità umana [21]: un modello ottimale nella compressione probabilistica può produrre topic incoerenti per un analista. Questa discrepanza ha spostato il focus verso metriche *human-centric*, che approssimano algoritmicamente il giudizio qualitativo umano valutando la coesione interna dei topic, la loro varietà lessicale, la separazione geometrica fra distribuzioni e la specificità dei singoli termini.

6.1 Obiettivo del confronto

L'analisi sperimentale condotta in questa tesi persegue due obiettivi principali. In primis vi è il confronto tra BERTopic, nella sua configurazione standard, e i due approcci classici di topic modeling probabilistico illustrati nei capitoli precedenti. Si intende così verificare se l'introduzione di rappre-

sentazioni semantiche contestuali e di un paradigma di clustering basato su densità conduca effettivamente a una migliore caratterizzazione tematica del corpus rispetto ai modelli generativi tradizionali fondati sul *bag-of-words*. In secondo luogo, l'analisi mira a valutare se le varianti di BERTopic proposte nel Capitolo 4 producano un miglioramento misurabile rispetto alla configurazione *baseline*.

I due confronti richiedono strategie di valutazione differenti. Il confronto tra BERTopic e i modelli classici non può essere condotto in modo soddisfacente attraverso le sole metriche quantitative, poiché le tre pipeline operano su rappresentazioni testuali profondamente diverse: LDA e STM lavorano su una document-term matrix costruita a valle di un'aggressiva fase di lemmatizzazione e *pruning* lessicale, mentre BERTopic riceve in input i documenti nella loro forma testuale originaria, preservando struttura sintattica e morfologia. Di conseguenza, metriche come la coerenza semantica o l'esclusività, sensibili sia al vocabolario sia alla distribuzione dei termini, non risultano direttamente confrontabili tra i due paradigmi. Per questa ragione si è scelto di affidare il confronto alla valutazione qualitativa umana, fondata sull'ispezione manuale dei topic estratti, della loro interpretabilità e della coerenza interna delle parole chiave associate.

Il confronto tra la configurazione *baseline* di BERTopic e le sue varianti, al contrario, avviene su un terreno omogeneo: tutte le pipeline condividono lo stesso preprocessing e operano sul medesimo spazio di rappresentazione. Diventa quindi possibile ricorrere alle 4 metriche quantitative (Coerenza, Diversità, Esclusività e distanza di Hellinger) esposte in seguito, per identificare in modo oggettivo le configurazioni più performanti e quantificare l'impatto di ciascuna modifica architetturale.

6.2 Coerenza semantica (UMass)

La coerenza UMass, introdotta da Mimno et al.[21], parte dal postulato che un topic sia semanticamente interpretabile se le parole che lo caratterizzano con la maggiore massa probabilistica tendono a co-occorrere frequentemente negli stessi documenti del corpus di addestramento. Se un modello raggruppa

termini che raramente compaiono insieme nel testo reale, è probabile che il topic sia un artefatto statistico privo di significato. A differenza delle metriche estrinseche basate sulla *Pointwise Mutual Information* (UCI/PMI [23], NPMI) o sui vettori di contesto (C_V [27]), che richiedono riferimenti esterni come grandi corpora di confronto, la coerenza UMass è *intrinseca*: viene calcolata direttamente sulle statistiche del corpus usato per l'addestramento, senza latenze legate a fonti esterne.

Dato un topic k , si seleziona l'insieme delle sue M parole principali $W^{(k)} = \{w_1, \dots, w_M\}$, ordinate per probabilità marginale decrescente (tipicamente le prime 10 o 20). Per ogni coppia di parole (w_i, w_j) si calcola la probabilità condizionata logaritmica:

$$\text{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)}, \quad (6.1)$$

dove $D(w_i, w_j)$ è la *co-document frequency*, cioè il numero di documenti in cui w_i e w_j co-occorrono; $D(w_j)$ è la *document frequency* marginale di w_j , cioè il numero di documenti in cui w_j appare indipendentemente dall'altra parola; ed ϵ è un fattore di *smoothing* (di norma $\epsilon = 1$) che previene il logaritmo di zero quando le due parole non co-occorrono mai. Il punteggio aggregato del topic è la somma su tutte le coppie del nucleo delle prime M parole selezionate, con w_m di rango inferiore rispetto a w_l :

$$C_{\text{UMass}}(k) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(w_m, w_l) + 1}{D(w_l)}. \quad (6.2)$$

Un aspetto strutturale rilevante è l'*asimmetria* della metrica: a differenza della NPMI, che si basa su una co-occorrenza simmetrica, l'equazione (6.1) stima una probabilità condizionata unidirezionale, cioè la probabilità di osservare la parola di rango inferiore w_m dato che la parola di rango superiore w_l è presente. Questa scelta modella la natura gerarchica del linguaggio: le parole ad alto rango fungono da termini “ancora” (concetti ampi e frequenti che definiscono il topic), mentre quelle a basso rango sono termini specialistici o rari. Il calcolo penalizza così in modo asimmetrico il caso in cui una parola molto specifica venga associata a una parola-ancora frequente con cui

non ha alcun legame documentale, restituendo un valore logaritmico fortemente negativo [21]. Poiché il numeratore non può eccedere il denominatore ($D(w_i, w_j) \leq D(w_j)$), il punteggio è sempre negativo; un valore meno negativo (tendente a zero) indica un topic le cui parole principali sono fortemente correlate.

Il principale limite empirico è la sensibilità alle *junk words*: in presenza di stop-word o termini grammaticali ubiqui non rimossi in fase di pre-processing, numeratore e denominatore assumono valori simili e producono un punteggio elevato ma privo di sostanza informativa. Per questa ragione la correlazione della coerenza UMass con il giudizio umano è meno solida di quella delle metriche estrinseche, e nel paper di riferimento [21] viene sconsigliato di utilizzarla come unica funzione obiettivo in fase di addestramento, poiché la sua ottimizzazione aggressiva tende a far collassare il modello su insiemi di parole frequentissime, con perdita dei topic rari o specializzati. Da qui la necessità di affiancarle un indice di valutazione complementare, basato generalmente sulla diversità lessicale.

6.3 Diversità dei topic

L'integrazione dei modelli probabilistici con il deep learning (autoencoder variazionali, word embeddings) ha permesso di aumentare i punteggi di coerenza, ma ha introdotto la vulnerabilità del *topic collapse* [31]: un modello può convergere su K topic iper-coerenti che però differiscono solo per permutazioni delle stesse poche parole dominanti, risultando inutile nella pratica. Per penalizzare questa ridondanza, Dieng et al. [12] hanno introdotto la metrica della *Topic Diversity* (TD) contestualmente all'Embedded Topic Model (ETM), un'architettura che modella termini e topic nello stesso spazio di embedding.

La diversità è definita come la proporzione di parole uniche tra le prime M parole più probabili di tutti i topic del modello. Detto K il numero di topic, $W_{\text{top}}^{(k)}$ l'insieme delle M parole a più alta probabilità di emissione sotto il topic

k ($P(w | k)$), e $U_W = \bigcup_{k=1}^K W_{\text{top}}^{(k)}$ la loro unione,

$$\text{TD} = \frac{|U_W|}{M K}, \quad (6.3)$$

dove $|U_W|$ è la cardinalità dell'insieme unione, cioè il conteggio dei token lessicali univoci estratti dall'intero modello senza doppi conteggi. A differenza della coerenza, che è una misura *intra-topic*, la diversità è una misura *inter-topic* che giudica il modello nel suo insieme. Il risultato è confinato in $\text{TD} \in [0, 1]$: un valore prossimo a 0 diagnostica ridondanza e collasso (lo stesso nucleo di termini riciclato su più topic), mentre un valore prossimo a 1 indica topic lessicalmente distinti che esplorano nicchie semantiche del vocabolario.

Coerenza e diversità sono complementari e spesso in contrapposizione. Per evitare di premiare modelli sbilanciati su un solo asse, [12] definiscono la meta-metrica composita *Topic Quality* come prodotto delle due:

$$\text{Topic Quality} = \text{Topic Diversity} \times \text{Topic Coherence}, \quad (6.4)$$

dove la forma moltiplicativa impone un vincolo di ottimizzazione congiunta: un crollo in una delle due componenti annulla il punteggio globale. Un modello che satura i topic con termini frequentissimi e ubiqui aumenta la coerenza ma abbassa la diversità, precipitando la *Topic Quality* verso lo zero; solo un modello che estrae termini contemporaneamente correlati all'interno del tema e distintivi rispetto agli altri raggiunge un punteggio elevato. Gli esperimenti di [12] mostrano inoltre che i modelli basati su embedding (come ETM) limitano meglio l'infiltrazione delle stop-word, mantenendo punteggi di diversità elevati anche su vocabolari di grandi dimensioni con distribuzioni a coda lunga.

6.4 Distanza di Hellinger

Per quantificare quanto due distribuzioni tematiche siano distanti nello spazio latente si ricorre alla distanza di Hellinger, introdotta da Ernst Hellinger [17] nel contesto della teoria della misura. Appartiene alla famiglia delle f -divergenze (a cui afferisce per esempio la divergenza di Kullback–Leibler),

ma a differenza di quest'ultima è una vera metrica: è simmetrica e soddisfa la disuguaglianza triangolare.

Per due distribuzioni continue P e Q con densità f e g , il quadrato della distanza di Hellinger è definito come

$$H^2(P, Q) = \frac{1}{2} \int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx = 1 - \int \sqrt{f(x)g(x)} dx, \quad (6.5)$$

dove l'integrale $\int \sqrt{fg} dx$ è il coefficiente di affinità di Bhattacharyya e l'uguaglianza segue dal fatto che ciascuna densità integra a 1. Nel topic modeling, tuttavia, le distribuzioni sono discrete: si tratta tipicamente delle righe della matrice parola-topic β (su un vocabolario di cardinalità V) o dei pesi documento-topic θ (su K assi tematici). Per due distribuzioni discrete $P = (p_1, \dots, p_n)$ e $Q = (q_1, \dots, q_n)$,

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2}. \quad (6.6)$$

La distanza è dunque proporzionale alla norma euclidea ($L2$) calcolata fra i vettori delle radici quadrate delle probabilità, scalata di $1/\sqrt{2}$: ciò la confina in $[0, 1]$, con $H(P, Q) = 0$ se e solo se $P \equiv Q$ (sovrapposizione totale) e $H(P, Q) = 1$ per supporti disgiunti (ortogonalità completa).

La ragione per cui questa metrica è particolarmente adatta al topic modeling risiede nella natura degli oggetti che il modello produce. Tanto le righe della matrice β (la distribuzione di probabilità sul vocabolario associata a un topic) quanto le righe della matrice θ (la mistura di topic di un documento) sono distribuzioni di probabilità definite sul simpleso, cioè vettori non negativi a somma unitaria. Confrontarle con la distanza euclidea o con la similarità del coseno significa trattarle come generici punti di $\mathbb{R}^n$, ignorando che rappresentano masse di probabilità; la distanza di Hellinger, invece, è costruita per misurare la dissimilarità fra distribuzioni e restituisce un valore con una semantica probabilistica precisa. Due proprietà la rendono preferibile alle alternative più comuni. In primo luogo, essendo limitata in $[0, 1]$ e simmetrica, produce una matrice di distanze valida e interpretabile su scala fissa: ciò consente di darla in input ad algoritmi che richiedono una metrica

(clustering agglomerativo, k -medoids), di mediare o determinare una soglia delle distanze in modo coerente e di confrontare valori ottenuti da modelli o esecuzioni diversi, operazioni precluse alla divergenza di Kullback–Leibler, asimmetrica e illimitata. In secondo luogo, la trasformazione a radice quadrata amplifica il peso delle differenze sulle componenti a bassa probabilità: due topic che condividono le parole ad alta frequenza ma divergono nei termini di nicchia vengono riconosciuti come distinti, mentre la similarità del coseno sulle probabilità grezze sarebbe dominata dai pochi termini ad alta massa, mascherando proprio la porzione di vocabolario che ne caratterizza la specificità.

Nell’ambito del topic modeling la distanza di Hellinger viene impiegata principalmente per due scopi. *(i) Misura di similarità e ridondanza fra topic:* calcolata a coppie fra le distribuzioni β , quantifica la separazione del modello a livello dell’intera distribuzione sul vocabolario, e non soltanto delle prime parole come fa la Topic Diversity, permettendo di individuare coppie di topic quasi duplicati. *(ii) Allineamento e diagnostica di stabilità:* poiché l’inferenza di LDA è stocastica (Gibbs sampling o inferenza variazionale), esecuzioni diverse producono topic non identici; accoppiando i topic di due esecuzioni (ad es. con l’algoritmo ungherese) sulla base della distanza di Hellinger se ne valuta la stabilità, con valori medi prossimi a zero che indicano convergenza e valori elevati che segnalano instabilità.

In questa tesi è stata usata unicamente come misura di ridondanza tra i topic.

6.5 Esclusività e punteggio FREX

Un sistema che ordina i termini per sola frequenza premia inevitabilmente le parole comuni a tutti i topic: ispezionando le top-list della matrice termine-topic β si ottengono così elenchi di vocaboli banali, ripetuti trasversalmente in ogni cluster. Per misurare quanto un termine sia specifico di un topic, Bischof et al. nel 2012 [2] hanno formalizzato il concetto di esclusività. Detta $\beta_{f,k}$ la probabilità condizionata $P(w_f \mid z_k)$ della parola f nel topic k , e S l’insieme dei topic di confronto (che può coincidere con tutti i topic del modello o con un sottoinsieme di topic “fratelli” in architetture gerarchiche),

l'esclusività è la frequenza normalizzata per colonna:

$$\phi_{f,k} = \frac{\beta_{f,k}}{\sum_{j \in S} \beta_{f,j}}. \quad (6.7)$$

Un valore elevato indica che la massa probabilistica della parola proviene quasi esclusivamente dal topic k ; per i termini diffusi su molti topic il denominatore cresce e $\phi_{f,k}$ tende al valore base $1/|S|$. L'esclusività da sola, tuttavia, presenta una distorsione: premia patologicamente le parole “iper-esclusive ma ipo-frequenti” (termini rari, refusi o codici che compaiono poche volte ma sempre nello stesso topic), che ottengono un punteggio prossimo a 1 per semplice assenza di competitori.

Per bilanciare frequenza ed esclusività, Bischof e Airolti [2] hanno definito il punteggio *FREX* (Frequency–Exclusivity) come media armonica dei ranghi percentile delle due dimensioni (formulazione poi adottata negli STM da Roberts et al. [25, 26]). I ranghi sono calcolati tramite la funzione di ripartizione empirica (ECDF), che neutralizza l'effetto degli outlier e dei forti gradienti di volume:

$$\text{FE}_{f,k} = \left(\frac{\omega}{\text{ECDF}_{\phi_{\cdot,k}}(\phi_{f,k})} + \frac{1-\omega}{\text{ECDF}_{\mu_{\cdot,k}}(\mu_{f,k})} \right)^{-1}, \quad (6.8)$$

dove $\text{ECDF}_{\phi_{\cdot,k}}(\phi_{f,k})$ è il rango percentile della parola sull'asse di esclusività, $\text{ECDF}_{\mu_{\cdot,k}}(\mu_{f,k})$ è il rango percentile sull'asse della (log-)frequenza $\mu_{f,k}$, e ω è il peso che regola il bilanciamento fra i due assi (in implementazioni consolidate come il pacchetto *stm*, il valore predefinito è $\omega = 0,7$, scelta che privilegia l'esclusività sulla frequenza). La scelta della media armonica è cruciale: a differenza della media aritmetica, essa penalizza fortemente l'asse debole, per cui un termine ottiene un FREX elevato solo se possiede un rango elevato sia per frequenza che per esclusività.

6.6 Valutazione qualitativa

Come discusso nella sezione precedente, il confronto tra BERTopic e i modelli classici è stato condotto attraverso un protocollo di valutazione qualitativa umana, in grado di restituire un giudizio interpretabile e indipendente dalle

specifiche scelte di preprocessing che caratterizzano ciascuna pipeline. Per ogni modello sono state ispezionate le prime dieci parole più rappresentative di ciascun topic e, ove utile, i documenti maggiormente associati, secondo una griglia di valutazione strutturata in due blocchi.

Il primo blocco riguarda la valutazione del singolo topic, condotta in modo indipendente per ciascuna partizione. Sono state misurate tre dimensioni: la *coerenza semantica*, ovvero il grado in cui le dieci parole presentano verso un tema riconoscibile e ben strutturato, espressa su una scala Likert da 0 (parole prive di filo conduttore) a 5 (topic pulito e ben delineato); l'*interpretabilità*, intesa come la possibilità di assegnare al topic un'etichetta tematica chiara e univoca, anch'essa valutata da 0 (etichetta non formulabile) a 5 (etichetta perfettamente rappresentativa di tutte le parole); e il numero di *parole intruse*, ossia quante delle dieci parole risultano fuori tema o non pertinenti (conteggio da 0 a 10). A questi indicatori si aggiunge la possibilità di proporre un'etichetta tematica sintetica in forma libera, utile come traccia interpretativa per il confronto successivo.

Il secondo blocco, compilato dopo aver esaminato l'intera partizione, introduce una prospettiva comparativa. La *distintività* (o esclusività) misura su scala 0–5 il grado in cui ciascun topic occupa uno spazio semantico autonomo rispetto agli altri, con il valore minimo corrispondente a topic sostanzialmente ridondanti e il massimo a topic chiaramente separati. Per i topic giudicati sovrapposti vengono inoltre indicati gli identificativi dei topic con cui si rileva la ridondanza, ottenendo così una mappa qualitativa delle aree tematiche condivise. Infine, viene stimato il numero di *topic distinti effettivi*, ovvero quanti nuclei tematici realmente autonomi emergono al netto delle sovrapposizioni.

Questo protocollo consente di confrontare in modo omogeneo modelli che producono partizioni eterogenee per numerosità e granularità, restituendo un giudizio complessivo sulla qualità interna dei topic e sulla loro effettiva informatività complessiva.

6.7 Configurazioni dei modelli

6.7.1 LDA

La scelta del numero di topic K rappresenta uno dei nodi critici nella configurazione di un modello LDA, poiché incide direttamente sulla granularità della partizione tematica e, di riflesso, sulla sua interpretabilità. Per individuare il valore ottimale è stata impiegata la funzione *FindTopicsNumber()* del pacchetto R *ldatuning* [22], che consente di valutare simultaneamente la qualità del modello al variare di K secondo quattro metriche complementari proposte in letteratura: *Griffiths2004*, *CaoJuan2009*, *Arun2010* e *Deveaud2014*.

La metrica di Griffiths e Steyvers [15] stima la log-verosimiglianza marginale dei dati attraverso una procedura di campionamento, e va massimizzata: valori elevati indicano un modello che si adatta meglio ai dati osservati. Cao et al. [8] propongono invece una misura basata sulla similarità coseno media tra le distribuzioni topic-parola, da minimizzare: bassa similarità implica topic più distinti e meno ridondanti. Arun et al. [1] introducono una metrica che confronta la decomposizione spettrale delle distribuzioni topic-documento e topic-parola tramite divergenza di Kullback-Leibler simmetrica, anch'essa da minimizzare in corrispondenza del numero ottimale di topic. Infine, la misura di Deveaud et al. [11], da massimizzare, calcola la divergenza di Jensen-Shannon media tra coppie di topic, premiando le configurazioni in cui i topic risultano semanticamente ben separati.

L'esplorazione condotta ha individuato $K = 20$ come valore che meglio bilancia le quattro metriche, offrendo un buon compromesso tra coerenza interna dei topic e separazione tra gli stessi. Fissato K , il modello LDA è stato stimato tramite *Gibbs sampling* collassato, metodo di inferenza preferito alla *Variational Bayes* per la sua maggiore stabilità su corpora di dimensioni medie e per la migliore approssimazione della distribuzione a posteriori.

6.7.2 STM

Il modello STM è stato configurato in modo da sfruttare appieno la sua caratteristica distintiva, ovvero l'integrazione di covariate a livello di docu-

mento nella stima della prevalenza dei topic. Come presentato in precedenza, le variabili impiegate per modellare la prevalenza sono l'anno di pubblicazione del documento, modellato attraverso una *spline* con 3 gradi di libertà; il Museo di appartenenza; il numero totale di interazioni di ogni singolo post, anch'essa modellata via *spline*; e il tipo di media, che distingue tra le diverse tipologie di post (immagine, video, link).

La scelta del numero ottimale di topic è stata effettuata attraverso il *CD score* (*Consistency and Differentiation score*), introdotto da Sciandra et al. [28] nell'ambito di un'analisi dei discorsi di fine anno dei presidenti della Repubblica italiana. La metrica integra in un unico indice le due dimensioni fondamentali della qualità di un modello topico: la coerenza semantica interna ai topic e la loro esclusività lessicale. Operativamente, per ciascun valore di K esplorato vengono stimati un modello STM e le relative metriche di coerenza ed esclusività tramite la funzione *searchK()* del pacchetto R *stm* [26]. Le metriche vengono successivamente normalizzate tramite funzione *min-max* sull'intervallo dei valori di K esplorati e successivamente combinate tramite norma $L2$, così da privilegiare le configurazioni ad alta coerenza ed alta esclusività. L'esplorazione condotta ha individuato $K = 17$ come valore ottimale.

6.7.3 BERTopic

La configurazione base di BERTopic riproduce la pipeline standard descritta nel Capitolo 3, utilizzando i valori di default per tutti gli iperparametri dei singoli moduli. L'unica modifica riguarda il modello di *embedding*, sostituito con *paraphrase-multilingual-MiniLM-L12-v2* per garantire una codifica semantica adeguata alla lingua italiana del corpus.

Per favorire la parsimonia nel numero di topic estratti e rendere agevole la successiva valutazione qualitativa umana, è stata inoltre applicata la funzione *reduce_topics()*, già citata nel Capitolo 3, che aggrega iterativamente i topic più simili sulla base della similarità coseno tra le rispettive rappresentazioni c-TF-IDF. Questa scelta consente di ottenere una partizione tematica sufficientemente compatta da poter essere ispezionata e confrontata manualmente

con i risultati di LDA e STM. In questo caso il numero di topic estratto è stato di 27.

Capitolo 7

Risultati

7.1 Panoramica generale degli argomenti

Pur con le differenze legate al diverso numero di topic e alla diversa rappresentazione del testo, i tre modelli convergono nel far emergere un nucleo di temi ricorrenti e ben definiti, riconducibili alla comunicazione tipica dei musei scientifici. Il filone più trasversale e nitido è quello astronomico, presente in tutti i modelli e spesso articolato in più topic distinti: il sistema solare e l’osservazione del cielo (*terra, sole, pianeta, stella, luce*), le missioni spaziali (*spazio, luna, missione, satellite*) e l’attività osservativa degli osservatori (*planetario, osservatorio, Brera, Schiaparelli, Marte*), riconducibile alle pagine INAF. Altrettanto chiari e stabili sono i topic legati all’identità dei singoli musei: quello incentrato su Leonardo da Vinci e sul Museo della Scienza e della Tecnologia di Milano (*leonardo, vinci, galleria*), quello relativo al Museo Galileo di Firenze (*galileo, firenze, fondazione*), il polo museale di Trieste (*trieste, ricercatore, immaginario*) e il Museo “A come Ambiente” di Torino (*maca, ambiente, laboratorio*). A questi si affianca un gruppo di topic di carattere educativo e organizzativo, anch’esso comune a STM e LDA: da un lato le attività e i laboratori rivolti a scuole, famiglie e bambini (*laboratorio, prenotazione, scuola, studente, bambino*), dall’altro la dimensione logistica della fruizione museale, fatta di visite, biglietti, eventi e contenuti digitali

(*biglietto, online, programma, sabato, virtuale*). Compaiono infine temi legati alle collezioni e agli oggetti iconici dei musei, come le mostre e gli strumenti storici (*mostra, opera, collezione, restauro*) o il sottomarino Toti e i padiglioni dei trasporti (*sottomarino, toti, padiglione, ferroviario*).

La maggiore granularità di BERTopic consente inoltre di isolare con nettezza una serie di temi più specifici ma altrettanto coerenti, che i modelli classici tendevano ad aggregare o a non far emergere. Sul versante dei contenuti scientifici si distinguono topic dedicati alla chimica (*chimica, chemioluminescenza, elettroni*), alle biotecnologie e alla genetica (*biotecnologie, genoma, molecolare*), alla vita marina (*pesci, delfini, oceano, squali*), alle neuroscienze (*neuroscienze, cervello*) e all'intelligenza artificiale (*algoritmi, intelligenza, artificiale*). Accanto a questi emergono temi di mitologia e costellazioni (*mitologia, greca, costellazioni*), e diversi topic di natura comunicativa e istituzionale: l'accessibilità e l'inclusione (*accessibile, inclusività, ciechi*), la filatelia e la *mail art* (*francobollo, filatelia, mailart*) e la donazione del cinque per mille (*fiscale, irpef, contribuente*). Nel loro insieme, questi risultati delineano un'architettura tematica stabile attorno a pochi nuclei forti (astronomia, identità dei musei, attività educative e didattiche) e una periferia di argomenti più specifici, la cui emersione dipende in misura sensibile dalla granularità del modello. L'elenco completo delle parole più rappresentative di ciascun topic è riportato nell'Appendice [A](#).

7.2 Confronto tra LDA, STM e BERTopic

Il confronto tra i modelli è stato articolato in due fasi. In primo luogo si sono valutati i due approcci probabilistici classici, LDA e STM, che condividono lo stesso impianto generativo e sono quindi direttamente confrontabili attraverso le metriche quantitative. Successivamente, i risultati sono stati messi a confronto con quelli ottenuti da BERTopic.

7.2.1 LDA e STM

Il confronto tra STM e LDA è stato inizialmente condotto analizzando congiuntamente *exclusivity* e *semantic coherence* per ciascun topic. La Figura [7.1](#)

7.2.1 LDA e STM

mostra, per ciascun topic dei due modelli, il posizionamento nello spazio definito da exclusivity (asse delle ascisse) e semantic coherence (asse delle ordinate). I topic più performanti si collocano nell'angolo in alto a destra del grafico, dove entrambe le metriche assumono valori elevati.

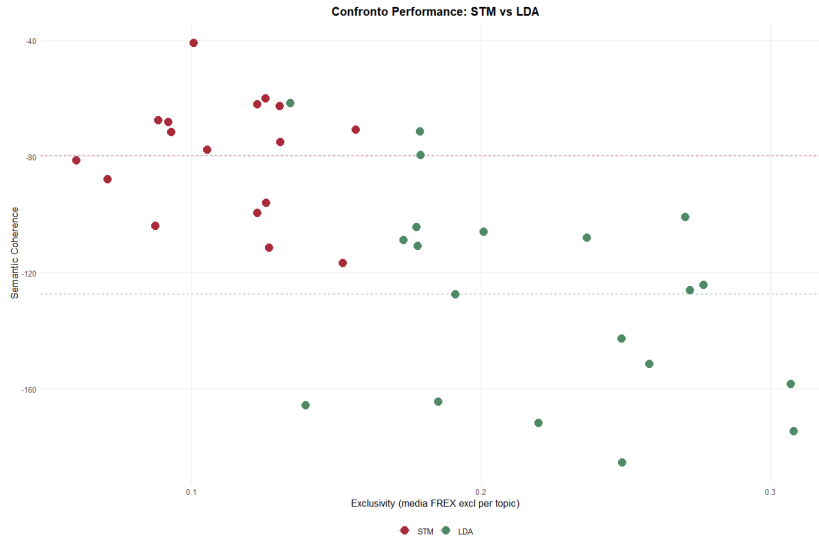


Figura 7.1: Confronto delle performance tra STM (rosso) e LDA (verde) in termini di exclusivity (ascissa) e semantic coherence (ordinata). Le linee tratteggiate indicano il valore medio di Coherence per entrambi i modelli.

Visivamente l'STM sembra individuare topic a coerenza maggiore, ma che presentano un'esclusività più bassa. Le differenze in termini di coerenza semantica ed esclusività sono state verificate mediante test di Wilcoxon per campioni indipendenti. I risultati confermano che STM presenta una coerenza media significativamente superiore rispetto a LDA ($p < 0,001$), indicando che i topic estratti da STM sono internamente più coesi. Analogamente, LDA mostra un'esclusività significativamente più alta rispetto a STM ($p < 0,001$), suggerendo che le parole caratteristiche dei suoi topic tendono ad essere più distintive tra un topic e l'altro.

A complemento di coerenza ed esclusività, il confronto è stato esteso alle altre due metriche aggregate a livello di modello: la *Topic Diversity* e la *distanza di Hellinger*. La Tabella 7.1 riporta le metriche automatiche aggre-

gate per entrambi i modelli, espresse come media $\pm$ deviazione standard, ove applicabile.

Metrica	LDA	STM
Coerenza semantica media	$-127,17 \pm 35,97$	$-79,13 \pm 20,32$
Exclusivity media (FREX)	$0,219 \pm 0,053$	$0,113 \pm 0,027$
Distanza di Hellinger media	$0,145 \pm 0,029$	$0,198 \pm 0,054$
Topic Diversity	0,845	0,724

Tabella 7.1: Metriche automatiche aggregate a livello di modello per LDA e STM. I valori sono espressi come media $\pm$ deviazione standard.

La maggiore Topic Diversity di LDA è coerente con la sua più alta esclusività FREX e indica che le parole più rappresentative dei suoi topic si sovrappongono meno tra loro. La distanza di Hellinger media premia invece STM: quest'ultima suggerisce che i topic di STM, pur condividendo alcune parole prominenti, differiscono in modo più marcato dalla distribuzione delle parole nel corpus. Letti congiuntamente, i quattro indicatori delineano due profili distinti: LDA produce topic con vocabolari più separati in superficie, mentre STM produce topic con maggior coesione interna e una struttura distributiva globalmente più articolata.

7.2.2 Modelli probabilistici e BERTopic

Come esposto in precedenza, a causa del diverso preprocessing dei dati da parte dei modelli probabilistici e di BERTopic, un confronto diretto attraverso le metriche quantitative standard non risulta fruibile. Pertanto si è optato per una valutazione umana qualitativa dei topic estratti dai vari modelli. Le valutazioni sono state condotte dal Prof. Stefano Crabu, professore associato presso il Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA) dell'Università degli Studi di Padova, che ha esaminato in modo indipendente i risultati prodotti dai tre modelli secondo il protocollo descritto nella sezione precedente.

7.2.2 Modelli probabilistici e BERTopic

I tre modelli, in base alle procedure di selezione descritte nella sezione precedente, producono un numero di topic comparabile ma non identico: 20 per LDA, 17 per STM e 27 per BERTopic. Per neutralizzare l'effetto della diversa cardinalità, gli indicatori sintetici comparativi sono espressi in percentuale rispetto al numero di topic di ciascun modello.

Distribuzione dei punteggi sui singoli topic Le Tabelle 7.2, 7.3 e 7.4 riportano la distribuzione di frequenza dei punteggi Likert 0–5 assegnati rispettivamente a Coerenza semantica, Interpretabilità e Distintività, espressi come percentuale sul totale dei topic di ciascun modello. I punteggi grezzi assegnati a ogni singolo topic, insieme ai grafi di sovrapposizione tematica costruiti a partire dalle segnalazioni del valutatore, sono riportati nell'Appendice B. Sotto al conteggio percentuale per ciascun livello sono indicati anche media¹ e mediana.

Punteggio	LDA ($K = 20$)	STM ($K = 17$)	BERTopic ($K = 27$)
0	0,0%	0,0%	7,4%
1	0,0%	11,8%	11,1%
2	10,0%	5,9%	7,4%
3	40,0%	35,3%	7,4%
4	15,0%	11,8%	7,4%
5	35,0%	35,3%	59,3%
Media	3,75	3,53	3,74
Mediana	3,5	3,0	5,0

Tabella 7.2: Distribuzione percentuale dei punteggi di Coerenza semantica.

Sul piano descrittivo si possono evidenziare alcuni andamenti. La distribuzione della Coerenza in BERTopic è fortemente asimmetrica: una larga maggioranza di topic ottiene il punteggio massimo (59,3% con valore 5), mentre una coda non trascurabile di topic risulta del tutto incoerente (oltre il 18% con

¹Sebbene la media sia un indice improprio per valori di una scala likert, è comunque stata inserita per offrire un'indicazione sommaria della tendenza centrale

Punteggio	LDA ($K = 20$)	STM ($K = 17$)	BERTopic ($K = 27$)
0	30,0%	17,6%	40,7%
1	0,0%	0,0%	14,8%
2	25,0%	11,8%	18,5%
3	10,0%	29,4%	7,4%
4	15,0%	17,6%	11,1%
5	20,0%	23,5%	7,4%
Media	2,40	3,00	1,56
Mediana	2,0	3,0	1,0

Tabella 7.3: Distribuzione percentuale dei punteggi di *Interpretabilità*.

punteggio ≤ 1); LDA e STM, al contrario, mostrano distribuzioni più concentrate intorno a valori intermedi. La situazione si rovescia sull'Interpretabilità, dove BERTopic concentra il 55% dei topic a ≤ 1 contro il 30% di LDA e il 17,6% di STM, segnalando una difficoltà specifica nell'assegnare etichette tematiche stabili alle parole estratte. STM mostra il profilo più equilibrato sulle prime due dimensioni; LDA presenta la distribuzione di Distintività più favorevole, con il 40% di topic giudicati pienamente autonomi.

Indicatori riassuntivi del primo e secondo blocco La Tabella 7.5 riassume gli indicatori derivati dal primo blocco del protocollo (Parole intrusive e percentuale di topic ad alta qualità, definito come la percentuale di topic con Coerenza ≥ 4 e Interpretabilità ≥ 4). La Tabella 7.6 riporta gli indicatori comparativi del blocco 2.

I due blocchi raccontano storie tra loro coerenti ma distinte. LDA mostra la più alta concentrazione di topic singolarmente di buona qualità (oltre un terzo soddisfa simultaneamente il criterio di coerenza e interpretabilità), STM ha il miglior dato sulla pulizia lessicale (mediana di parole intrusive pari a 0, contro 1 degli altri due modelli), mentre BERTopic eccelle sull'articolazione complessiva, con il maggior numero di topic semanticamente distinti tra loro: quasi tre quarti dei suoi topic occupano uno spazio tematico proprio, contro circa la metà di quelli di LDA e STM. La maggiore granularità di BERTopic,

7.2.2 Modelli probabilistici e BERTopic

Punteggio	LDA ($K = 20$)	STM ($K = 17$)	BERTopic ($K = 27$)
0	0,0%	0,0%	0,0%
1	0,0%	0,0%	0,0%
2	0,0%	11,8%	7,4%
3	50,0%	52,9%	40,7%
4	10,0%	35,3%	44,4%
5	40,0%	0,0%	7,4%
Media	3,90	3,24	3,52
Mediana	3,5	3,0	4,0

Tabella 7.4: Distribuzione percentuale dei punteggi di *Distintività*.

Indicatore	LDA	STM	BERTopic
Parole intrusive: media	1,60	0,82	2,07
Parole intrusive: mediana	1,0	0,0	1,0
% topic con coerenza ≥ 4 e interpr. ≥ 4	35,0%	29,4%	18,5%

Tabella 7.5: Indicatori sintetici di qualità dei singoli topic (blocco 1).

in altre parole, non si traduce in ridondanza, ma in una migliore separazione tra i nuclei tematici emergenti.

Tabella riassuntiva La Tabella 7.7 e la Figura 7.2 sintetizzano i sette indicatori utilizzati. Il quadro che emerge è chiaramente non univoco: nessun modello domina su tutte le dimensioni, e ogni paradigma mostra un proprio profilo di forze e debolezze.

Letti congiuntamente, i sette indicatori restituiscono tre profili distinti. LDA appare il modello con la migliore qualità interna media nel primo blocco relativo alla qualità dei topic, premiato dalla regolarità con cui produce topic singolarmente leggibili ed etichettabili. STM offre la combinazione più equilibrata tra le dimensioni, distinguendosi soprattutto per la maggior interpretabilità mediana e per la più bassa presenza di parole intrusive. BERTopic, infine, mostra il quadro più polarizzato: una minoranza significativa di topic

Indicatore	LDA	STM	BERTopic
Distintività: mediana	3,5	3,0	4,0
Sovrapposizioni medie per topic	2,85	2,65	1,63
Topic distinti effettivi (%)	55,0%	47,1%	74,1%

Tabella 7.6: Indicatori sintetici di articolazione complessiva del modello (blocco 2).

Indicatore	LDA ($K = 20$)	STM ($K = 17$)	BERTopic ($K = 27$)
Coerenza: mediana	3,5	3,0	5,0
Interpretabilità: mediana	2,0	3,0	1,0
Parole intrusive: media	1,60	0,82	2,07
% topic di alta qualità	35,0%	29,4%	18,5%
Distintività: mediana	3,5	3,0	4,0
Sovrapposizioni medie	2,85	2,65	1,63
Topic distinti effettivi (%)	55,0%	47,1%	74,1%

Tabella 7.7: Quadro riassuntivo degli indicatori qualitativi.

risulta poco interpretabile, ma la quota di topic semanticamente autonomi è nettamente superiore agli altri due modelli, suggerendo che il paradigma neurale tende a privilegiare la separazione tematica del corpus a discapito della pulizia lessicale interna di ciascun topic.

7.3 Varianti di BERTopic

7.3.1 Modello di embedding monolingue e rappresentazione con RAKE

Configurazione Come discusso nella Sezione 4.1, la baseline di BERTopic impiega il modello multilingue *paraphrase-multilingual-MiniLM-L12-v2*, in grado di operare su oltre 50 lingue all'interno di uno spazio vettoriale condiviso a 384 dimensioni. Per verificare se un modello addestrato specificamente

7.3.1 Modello di embedding monolingue e rappresentazione con RAKE

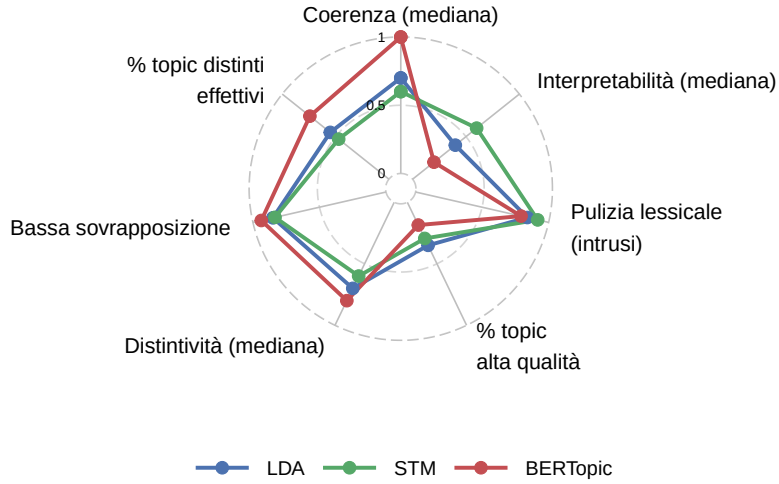


Figura 7.2: Profilo comparativo dei tre modelli sui sette indicatori della Tabella 7.7, normalizzati nell'intervallo $[0, 1]$, orientati in modo che valori più alti corrispondano sempre a una qualità migliore. Gli indicatori *Pulizia lessicale (intrusi)* e *Bassa sovrapposizione* sono ottenuti invertendo le rispettive scale grezze.

sulla lingua italiana possa restituire rappresentazioni più aderenti alle specificità morfologiche e lessicali del corpus, la baseline è stata modificata sostituendo il modello di embedding con *dbmdz/bert-base-italian-uncased* [10], un BERT monolingue italiano disponibile sulla piattaforma Hugging Face. Tutti gli altri moduli della pipeline restano invariati rispetto alla configurazione di base descritta nel Capitolo 3.

Risultati quantitativi La Tabella 7.8 riporta i valori delle quattro metriche (con le rispettive deviazioni standard, a esclusione della Diversity, che è una metrica globale) a confronto con la configurazione di base. La coerenza semantica mostra un lieve miglioramento, con una media meno negativa e una

dispersione leggermente inferiore, mentre l'esclusività e la distanza di Hellinger restano sostanzialmente invariate, con differenze ampiamente contenute entro una deviazione standard. La diversità, al contrario, registra un calo più marcato (da 0.94 a 0.81). Nel complesso, sul piano puramente quantitativo, le rappresentazioni prodotte dal modello monolingue non si differenziano in modo sostanziale da quelle del modello multilingue di riferimento.

Metrica	Baseline	<i>bert-base-italian-uncased</i>
Coerenza	-5.99 ± 3.35	-5.67 ± 2.89
Esclusività FREX (media)	0.73 ± 0.21	0.73 ± 0.22
Distanza di Hellinger	0.40 ± 0.12	0.41 ± 0.12
Diversità	0.94	0.81

Tabella 7.8: Confronto metriche: baseline multilingue *vs.* embedding monolingue italiano.

Ispezione visiva delle parole chiave. L'analisi delle parole chiave estratte dai due modelli evidenzia tuttavia differenze non trascurabili. Il modello monolingue italiano tende a far emergere all'interno dei topic parole vuote (*stopword*) e token privi di significato lessicale. Il fenomeno, pur non riguardando la totalità dei topic, è comunque ricorrente. La Tabella 7.9 riporta le 10 parole più rappresentative di due topic che esemplificano il problema: nel Topic 16 compaiono numerose stopwords (*non, una, te, tuo, altresì*), mentre nel Topic 20 emergono stringhe alfanumeriche e identificativi non interpretabili (ad esempio *plajepxdwsvglfmj9d35tbrwqa, e0pxjrezl4*).

Rappresentazione con RAKE Per affrontare il problema della scarsa qualità delle parole rappresentative trovate, mantenendo l'embedding monolingue italiano, si è provato a sostituire il modulo di rappresentazione dei topic con un'estrazione basata su RAKE.

Risultati con RAKE La Tabella 7.10 confronta le metriche delle tre configurazioni. Poiché la modifica interviene unicamente sulla fase di rappresentazione e non sul clustering, la partizione dei documenti in topic resta invariata

7.3.1 Modello di embedding monolingue e rappresentazione con RAKE

Rango	Topic 16	Topic 20
1	contributo	plajepxdwsvgulfmj9d35tbrwqa
2	beni	museorisorgimentomilano
3	non	postalcad
4	una	postcrossing
5	dichiarazione	francobollo
6	te	e0pxjrezl4
7	mandato	mailart
8	altresì	nicoli
9	irpef	ruggeromaggi
10	tuo	lascito

Tabella 7.9: Top 10 parole per due topic restituite dal modello monolingue italiano.

rispetto al modello monolingue: il confronto isola pertanto il solo effetto della diversa estrazione delle parole chiave (fase di Rappresentazione). L'utilizzo di RAKE invece del c-TF-IDF sembra apportare un leggero miglioramento nel modello monolingua. Tuttavia i risultati sono pressoché identici tra tutte e 3 le configurazioni.

Metrica	Baseline	Monolingue IT	Monolingue IT + RAKE
Coerenza	-5.99 ± 3.35	-5.67 ± 2.89	-5.00 ± 3.09
Esclusività FREX	0.73 ± 0.21	0.73 ± 0.22	0.73 ± 0.21
Distanza di Hellinger	0.40 ± 0.12	0.41 ± 0.12	0.44 ± 0.12
Diversità	0.94	0.81	0.88

Tabella 7.10: Confronto metriche tra baseline multilingue, embedding monolingue italiano ed embedding monolingue italiano con rappresentazione RAKE.

Sul piano qualitativo, l'effetto di RAKE è evidente nelle parole chiave degli stessi topic problematici individuati in precedenza. La Tabella 7.11 mostra come le stopwords e i token non interpretabili lascino il posto a termini dotati di maggiore contenuto informativo.

Rango	Topic 16 (RAKE)	Topic 20 (RAKE)
1	scegliere	posta
2	irpef	immaginata
3	donare	plajepxdwsvgulfmj9d35tbrwqa3mrpeji
4	potrebbero	arte
5	essere	canale
6	sensibili	università
7	dichiarazione	salvadori
8	nulla	youtu
9	spesa	prime
10	ente	firenze

Tabella 7.11: Top 10 parole per i Topic 16 e 20 con rappresentazione RAKE

Il confronto tra le due rappresentazioni mostra un miglioramento parziale ma apprezzabile. Nel Topic 16 le stopwords che dominavano la rappresentazione c-TF-IDF (*non, una, te, tuo, altresì*) sono sostituite da termini lessicalmente più informativi (*scegliere, donare, sensibili, spesa, ente*), rendendo il topic chiaramente interpretabile come legato alla donazione del cinque per mille e alla dichiarazione IRPEF. Permangono tuttavia alcune parole a basso contenuto semantico (*potrebbero, essere, nulla*).

Nel Topic 20 il miglioramento è più contenuto. I token alfanumerici e gli hashtag concatenati (*museorisorgimentomilano, postalcard, e0pxjrezl4*) lasciano il posto a termini interpretabili come *posta, arte, università* e *firenze*, che delineano un topic legato alla *mail art* e al circuito artistico-postale. Resta tuttavia presente una stringa alfanumerica non interpretabile e un token troncato (*youtu*), verosimilmente residuo di un URL.

Discussione L'adozione di un modello di embedding specifico per la lingua italiana, da sola, non ha prodotto un miglioramento della qualità complessiva della partizione tematica: sul piano quantitativo le metriche restano paragonabili a quelle della baseline, ma sul piano qualitativo la rappresentazione dei topic peggiora a causa della comparsa di *stopword* e parole non informa-

7.3.2 Analisi comparativa delle architetture spaziali: Riduzione Dimensionale e Clustering

tive. L'integrazione di RAKE come metodo di estrazione delle parole chiave interviene proprio su questo aspetto, restituendo topic più leggibili e interpretabili, tuttavia non cancella del tutto il problema. Una possibile causa di questo residuo di rumore lessicale risiede nella natura linguisticamente non omogenea del corpus: la presenza di documenti redatti in lingue diverse dall'italiano e di link esterni alla piattaforma di Facebook (come ad esempio Youtube) introduce nei topic termini che né l'embedding specializzato né la procedura di estrazione delle parole chiave, entrambi calibrati sul lessico italiano, riescono a filtrare in modo soddisfacente.

7.3.2 Analisi comparativa delle architetture spaziali: Riduzione Dimensionale e Clustering

Per valutare l'impatto strutturale dei moduli di riduzione dimensionale e clustering, la pipeline è stata riorganizzata secondo un disegno fattoriale completo 3×2 . Sono state testate le interazioni tra tre metodi di riduzione dimensionale (UMAP, PCA e PCA con kernel RBF) e due algoritmi di clustering (HDBSCAN e K-Means). Per isolare l'effetto di queste modifiche, la fase di embedding testuale è stata mantenuta costante per tutte le esecuzioni usando il modello multilingua.

La Tabella 7.12 mostra i risultati quantitativi e il numero di topic individuati dalle 6 configurazioni.

Modello	N. Topic	Coerenza	FREX	Hellinger	Diversità
Baseline	27	-5.99 ± 3.35	0.73 ± 0.21	0.40 ± 0.12	0.94
PCA + HDBSCAN	20	-3.12 ± 2.71	0.68 ± 0.20	0.37 ± 0.10	0.82
KPCA + HDBSCAN	24	-3.20 ± 2.60	0.69 ± 0.21	0.37 ± 0.11	0.77
UMAP + K-Means	97	-4.78 ± 3.00	0.39 ± 0.21	0.30 ± 0.10	0.60
PCA + K-Means	20	-2.10 ± 0.50	0.14 ± 0.10	0.19 ± 0.05	0.61
KPCA + K-Means	22	-2.26 ± 0.77	0.16 ± 0.13	0.19 ± 0.06	0.59

Tabella 7.12: Confronto delle metriche al variare dell'algoritmo di riduzione dimensionale e clustering. I valori sono riportati come media $\pm$ deviazione standard, a esclusione della Diversità (metrica globale dell'intero modello).

Effetto della riduzione dimensionale a parità di clustering La sostituzione di UMAP con una proiezione lineare (PCA) o non lineare a kernel (KPCA) mantenendo HDBSCAN come algoritmo di partizionamento produce un effetto sistematico: la coerenza semantica migliora in modo netto, passando da un valore medio di -5.99 della baseline a circa -3.1 in entrambe le varianti, con una dispersione anch'essa leggermente ridotta. Il numero di topic individuati scende leggermente (da 27 a 20 e 24, rispettivamente), e la diversità decresce di circa dieci punti percentuali, mentre esclusività FREX e distanza di Hellinger restano sostanzialmente allineate alla baseline (variazioni inferiori a una deviazione standard). PCA e KPCA producono risultati pressoché sovrapponibili, con un leggero vantaggio della prima sulla diversità (0.82 contro 0.77) e della seconda sul numero di topic conservati. Una possibile lettura complessiva è che le proiezioni lineari, preservando le direzioni di massima varianza, conducono HDBSCAN a individuare cluster meno frammentati e internamente più coesi sul piano lessicale: la coerenza migliora proprio perché vengono privilegiati i grossi blocchi tematici, a discapito delle nicchie più periferiche che UMAP riusciva invece a isolare in topic dedicati.

Effetto del clustering a parità di riduzione dimensionale La sostituzione di HDBSCAN con K-Means mantenendo UMAP come modulo di riduzione modifica la struttura della partizione in modo più radicale rispetto alle variazioni precedenti. Il numero di topic individuati passa da 27 a 97: poiché K-Means non dispone di un cluster dedicato al rumore, ogni documento viene necessariamente assegnato a un topic, e la richiesta di un numero elevato di cluster si traduce in una segmentazione fine dello spazio embedding. La coerenza media risulta intermedia (-4.78), ma il dato più rilevante è il crollo dell'esclusività FREX, che passa da 0.73 a 0.39 . Una FREX così bassa indica che porzioni significative dei top words sono condivise da più topic: la partizione perde quindi capacità discriminativa nonostante l'elevato numero di cluster. Lo stesso fenomeno si riflette nella diversità globale, che scende a 0.60 , e nella distanza di Hellinger media, che si riduce a 0.30 , segnalando topic via via più simili alla distribuzione del corpus.

Combinazione di riduzione lineare e K-Means Le configurazioni PCA + K-Means e KPCA + K-Means rappresentano il caso limite delle tendenze osservate sopra. La coerenza raggiunge i valori migliori in assoluto della batteria di esperimenti (compresa tra -2.10 e -2.26), con una deviazione standard ridotta a circa un sesto di quella della baseline: tutti i topic appaiono internamente compatti e simili tra loro per livello di coerenza. Questo risultato, tuttavia, non costituisce una prova di qualità della partizione, ma piuttosto un suo artefatto. L'esclusività FREX collassa a $0.14-0.16$, la distanza di Hellinger media scende a 0.19 e la diversità si attesta attorno a 0.60 , allineandosi ai valori della configurazione UMAP + K-Means. Una distanza di Hellinger così bassa indica che le distribuzioni di parole dei topic differiscono solo in misura limitata da quella del corpus complessivo: i topic, in altri termini, riproducono il vocabolario medio dei documenti, e la loro coerenza apparentemente elevata è diretta conseguenza del fatto che le parole top sono in larga parte le più frequenti del corpus. Le due varianti, PCA e KPCA, restituiscono risultati quantitativamente molto simili, con un numero di topic finali di 20 e 22 rispettivamente.

Ispezione visiva delle parole chiave L'analisi delle parole rappresentative dei topic conferma le tendenze emerse dalle metriche. Nelle configurazioni HDBSCAN + PCA e HDBSCAN + KPCA, la maggior parte dei topic individuati nella baseline viene conservata, ma si osserva una proliferazione di cluster centrati sul medesimo nucleo semantico, in particolare nell'area astronomica (planetario, osservatorio, sistema solare, missioni spaziali): la stessa macro-area viene rappresentata da quattro o cinque topic distinti, con sovrapposizioni evidenti nei termini più caratterizzanti. Il calo della diversità è coerente con questa osservazione. Nelle configurazioni con K-Means su UMAP la frammentazione assume caratteristiche diverse: temi minori della baseline vengono spaccati in più topic (sostenibilità, sottomarini, cambiamento climatico ricorrono in più cluster separati ma lessicalmente affini), e compaiono topic generici dominati da parole molto frequenti nel corpus museale. Le configurazioni con K-Means su riduzione lineare presentano infine il fenomeno più critico: alcuni topic risultano dominati da parole funzionali (*perché*, *può*, *cosa*, *essere*) o da termini di servizio del lessico istituzionale, e

non sono interpretabili come unità tematiche distinte. Questi *topic-rumore* non sono compensati da topic specifici di nicchia, che risultano in larga misura assenti.

Sintesi Il confronto tra le sei configurazioni evidenzia che le metriche di coerenza, esclusività, diversità e distanza di Hellinger non sono allineate fra loro e non possono essere lette da sole. Le configurazioni con K-Means ottengono la migliore coerenza assoluta, ma i valori fortemente depressi di FREX e Hellinger segnalano che la partizione non distingue i topic in modo sostanziale: una coerenza elevata su topic poco specifici non costituisce un vantaggio. All'estremo opposto, la baseline sacrifica coerenza in cambio di una partizione fine, fortemente differenziata e con topic ben separati dal corpus medio, come testimoniato dai valori più alti di diversità, FREX e Hellinger. Le configurazioni HDBSCAN + PCA e HDBSCAN + KPCA si collocano in una posizione intermedia, con un compromesso ragionevole tra coerenza e separazione, ma a costo della perdita dei temi più specifici. La configurazione UMAP + K-Means, infine, illustra il limite di un algoritmo di clustering rigido in uno spazio a struttura locale complessa: l'elevato numero di topic non si traduce in maggior potere descrittivo, ma in ridondanza interna. Le evidenze raccolte indicano dunque che la combinazione UMAP + HDBSCAN rappresenta, fra le opzioni testate, la più equilibrata sul piano della distintività e dell'articolazione tematica della partizione.

7.3.3 Rappresentazione con RAKE

Per completare l'analisi spaziale, si è valutato l'effetto della sostituzione del modulo di rappresentazione c-TF-IDF con RAKE. Coerentemente con quanto osservato nella Sezione 7.3.2, il test è stato condotto su tre configurazioni rappresentative di tre regimi distinti: la baseline (UMAP + HDBSCAN), considerata di riferimento; UMAP + K-Means, caratterizzata da un'elevata frammentazione della partizione; e PCA + K-Means, in cui il c-TF-IDF restituisce topic dominati da parole funzionali a basso contenuto informativo. Poiché RAKE interviene unicamente sulla fase di rappresentazione e non mo-

7.3.3 Rappresentazione con RAKE

difica la partizione, ogni confronto isola il solo effetto della diversa estrazione delle parole chiave.

Risultati quantitativi Le Tabelle 7.13 e 7.14 riportano le metriche delle sei configurazioni. L'esclusività FREX risulta pressoché identica fra le coppie di configurazioni: la metrica è calcolata sulla distribuzione c-TF-IDF sottostante e non sulla rappresentazione mostrata, motivo per cui non riflette la sostituzione del modulo. Le altre tre metriche presentano invece variazioni significative. La coerenza migliora nelle configurazioni in cui il c-TF-IDF aveva valori più bassi, con un guadagno di circa 1.5 punti sulla baseline e di 0.9 punti su UMAP + K-Means. Nella configurazione PCA + K-Means, al contrario, la coerenza peggiora sensibilmente, passando da -2.10 a -3.85 . Questo comportamento è coerente con quanto osservato in precedenza: il valore di coerenza elevato del c-TF-IDF era diretta conseguenza del fatto che i top words coincidevano con le parole più frequenti del corpus; sostituendoli con keyphrase più specifiche estratte da RAKE, la coerenza torna a valori realistici e simili a quelli delle altre configurazioni. La distanza di Hellinger media aumenta sistematicamente di circa 0.05 in tutti e tre i casi, indicando che le distribuzioni dei topic prodotte da RAKE si allontanano leggermente da quella del corpus globale. La diversità, invece, diminuisce in tutte le configurazioni di 0.09–0.10 punti: questo calo, in apparente contrasto con l'aumento di Hellinger, è dovuto alla ricorrenza, fra le keyphrase estratte da RAKE, di token tipici dei *social post* del corpus, come *https*, *www* e *bit*, che compaiono fra le parole rappresentative di numerosi topic.

Metrica	Baseline	UMAP + K-Means	PCA + K-Means
N. Topic	27	97	20
Coerenza	-5.99 ± 3.35	-4.78 ± 3.00	-2.10 ± 0.50
FREX	0.73 ± 0.21	0.39 ± 0.21	0.14 ± 0.10
Hellinger	0.40 ± 0.12	0.30 ± 0.10	0.19 ± 0.05
Diversità	0.94	0.60	0.61

Tabella 7.13: Metriche delle tre configurazioni di clustering con rappresentazione c-TF-IDF (riferimento).

Metrica	Baseline	UMAP + K-Means	PCA + K-Means
N. Topic	27	97	22
Coerenza	-4.47 ± 3.01	-3.90 ± 2.41	-3.85 ± 1.83
FREX	0.73 ± 0.21	0.39 ± 0.21	0.15 ± 0.12
Hellinger	0.46 ± 0.14	0.35 ± 0.10	0.24 ± 0.07
Diversità	0.84	0.51	0.55

Tabella 7.14: Metriche delle tre configurazioni dopo la sostituzione del modulo di rappresentazione con RAKE. I valori sono confrontabili riga per riga con quelli della Tabella 7.13.

Ispezione visiva delle parole chiave Il confronto dei top words conferma e completa la lettura quantitativa. Sulla baseline RAKE produce una rappresentazione più ricca per alcuni topic specifici: il Topic 16, ad esempio, passa da una rappresentazione c-TF-IDF già informativa (*fiscale, contribuente, redditi, spesa, tassa, donazioni*) a una formulazione in cui emergono termini più caratteristici dell'azione specifica descritta dai documenti (*scegliere, redditi, irpef, iniziativa, fiscale*), riconducibili al cinque per mille e alla dichiarazione IRPEF. Nella maggior parte dei topic, tuttavia, RAKE introduce nelle prime posizioni token derivati dalle URL presenti nei post social (*https, http, bit, www*), che il c-TF-IDF tendeva invece a deprimere. Questi token non sono interpretabili e ricorrono trasversalmente, motivo per cui la diversità globale cala nonostante i topic restino correttamente separati sul piano semantico. Nella configurazione UMAP + K-Means il fenomeno è ancora più marcato: poiché il numero di topic è elevato (97) e molti di essi insistono sullo stesso nucleo semantico, RAKE non riesce a disambiguare i cluster vicini, e quasi un terzo dei topic presenta *https* fra le prime tre keyphrase. La ridondanza interna alla partizione, già rilevata sul c-TF-IDF, non viene quindi attenuata dal cambio di rappresentazione. Nella configurazione PCA + K-Means, infine, RAKE non risolve il problema dei topic-rumore: laddove c-TF-IDF restituiva parole funzionali (*perché, può, cosa*), RAKE le sostituisce con altre keyphrase generiche o con frammenti di URL (*https, tel, nuovo, possiamo*). I topic restano non interpretabili, e il peggioramento della coerenza riflette il fatto che, una volta rimossi gli artefatti lessicali del c-TF-IDF, la partizione

7.3.4 Introduzione del Gibbs Sampling

sottostante non possiede una vera struttura tematica.

Discussione L'introduzione di RAKE come modulo di rappresentazione ha un effetto disomogeneo, fortemente condizionato dalla qualità della partizione su cui opera. Sulla baseline il cambio di rappresentazione produce un miglioramento parziale sul piano della coerenza e una rappresentazione dei topic qualitativamente più ricca per alcuni casi specifici, ma introduce in contropartita il problema dei token web ricorrenti, che riduce la diversità globale. Sulla configurazione UMAP + K-Means l'effetto è simile, ma è insufficiente a compensare il problema strutturale della frammentazione, che resta il limite principale della configurazione. Sulla configurazione PCA + K-Means, infine, RAKE non solo non risolve il problema dei topic-rumore, ma evidenzia in modo netto la natura artefattuale della coerenza ottenuta da c-TF-IDF, riportandola su valori coerenti con la reale separazione tematica della partizione. Il confronto conferma quindi che la rappresentazione, da sola, non può compensare i limiti strutturali del clustering: il modulo di rappresentazione modifica l'aspetto dei topic, ma la loro qualità informativa dipende in ultima istanza dalla qualità della partizione prodotta a monte.

7.3.4 Introduzione del Gibbs Sampling

L'ultima modifica testata sulla pipeline consiste nell'introduzione di uno stadio di Gibbs sampling come passo di raffinamento successivo al clustering. Il modulo opera sulle assegnazioni documento–topic prodotte dalla fase di clustering iniziale e ne aggiorna iterativamente la distribuzione attraverso un campionamento collassato in stile LDA. Sull'output di Gibbs viene quindi applicata una nuova proiezione tramite UMAP con distanza di Hellinger seguita da una ri-clusterizzazione con HDBSCAN, che produce la partizione finale. La fase di Gibbs è stata configurata con 5 iterazioni e iperparametri $\alpha = 0.1$, $\beta = 0.01$. Anche in questo caso, per mantenere confrontabili i risultati con le sezioni precedenti, le configurazioni testate sono baseline (UMAP + HDBSCAN), UMAP + K-Means e PCA + K-Means.

Risultati quantitativi La Tabella 7.15 riporta le metriche delle tre configurazioni di riferimento, mentre la Tabella 7.16 riporta i valori corrispondenti dopo l'introduzione di Gibbs. L'effetto più evidente del nuovo modulo è la modifica della cardinalità della partizione. Nella baseline il numero di topic passa da 27 a 106, in PCA + K-Means da 20 a 84; la configurazione UMAP + K-Means costituisce l'unica eccezione, con un numero di topic sostanzialmente invariato (da 97 a 102). Il fenomeno è una diretta conseguenza del fatto che Gibbs ammorbidisce le assegnazioni hard prodotte dal clustering: dove la partizione di partenza era poco numerosa rispetto alla varietà tematica del corpus, lo spazio delle distribuzioni topic-documento risultante viene ri-clusterizzato in un numero molto più elevato di gruppi. Le altre metriche presentano un comportamento eterogeneo a seconda della configurazione. Sulla baseline tutte le metriche peggiorano: la coerenza scende leggermente da -5.99 a -6.15 , ma soprattutto la FREX collassa da 0.73 a 0.30 , la distanza di Hellinger media scende da 0.40 a 0.25 e la diversità da 0.94 a 0.62 . La partizione, frammentata in oltre cento topic, perde quindi la distintività che ne costituiva il principale punto di forza. Sulla configurazione UMAP + K-Means l'effetto è opposto: la coerenza migliora di oltre un punto (da -4.78 a -3.42), la diversità sale da 0.60 a 0.72 , mentre FREX e Hellinger restano sostanzialmente stabili. Il numero di topic, già molto elevato in partenza, non aumenta di molto, e l'aggiornamento delle assegnazioni risulta in una partizione più coerente. Sulla configurazione PCA + K-Means, infine, Gibbs produce un risultato misto: la coerenza si abbassa da -2.10 a -4.56 , riportandola a valori coerenti con quelli delle altre configurazioni, ma in compenso la FREX raddoppia (da 0.14 a 0.29) e la distanza di Hellinger media aumenta da 0.19 a 0.27 . La perdita di coerenza, anche in questo caso, è verosimilmente artefattuale, mentre l'aumento di FREX e Hellinger segnala una reale separazione dei topic rispetto al corpus, ottenuta però al prezzo di una quadruplicazione del numero di cluster.

Ispezione visiva delle parole chiave L'analisi dei top words conferma il quadro tracciato dalle metriche. Sulla baseline post-Gibbs sopravvivono alcuni topic ben caratterizzati (*bicicletta, zodiaco, eroi e mitologia greca, chimica*), ma una porzione significativa dei nuovi cluster risulta dominata da *stopword*

7.3.4 Introduzione del Gibbs Sampling

Metrica	Baseline	UMAP + K-Means	PCA + K-Means
N. Topic	27	97	20
Coerenza	-5.99 ± 3.35	-4.78 ± 3.00	-2.10 ± 0.50
FREX	0.73 ± 0.21	0.39 ± 0.21	0.14 ± 0.10
Hellinger	0.40 ± 0.12	0.30 ± 0.10	0.19 ± 0.05
Diversità	0.94	0.60	0.61

Tabella 7.15: Metriche delle tre configurazioni di clustering prima dell'introduzione di Gibbs (riferimento).

Metrica	Baseline	UMAP + K-Means	PCA + K-Means
N. Topic	106	102	84
Coerenza	-6.15 ± 4.93	-3.42 ± 3.05	-4.56 ± 3.65
FREX	0.30 ± 0.22	0.38 ± 0.23	0.29 ± 0.17
Hellinger	0.25 ± 0.11	0.26 ± 0.08	0.27 ± 0.10
Diversità	0.62	0.72	0.62

Tabella 7.16: Metriche delle stesse tre configurazioni dopo l'introduzione del modulo di Gibbs sampling come passo di raffinamento. I valori sono confrontabili riga per riga con quelli della Tabella 7.15.

italiane (*di, la, una, da*) o da articoli e preposizioni inglesi (*the, of, and, to*), questi ultimi verosimilmente residui dei post in lingua inglese presenti nel corpus. Compaiono inoltre alcuni topic estremamente specifici, costituiti da combinazioni di hashtag o entità non riconducibili a un tema chiaramente identificabile. Su PCA + K-Means il fenomeno è ancora più marcato: oltre ai topic-stopword (*che, un, una, di, la, in*), compaiono topic dominati da numeri o formati orari (*00, 30, alle, 14, 19*), che riflettono la presenza nel corpus di indicazioni temporali ricorrenti nei post di programmazione. Sulla configurazione UMAP + K-Means, invece, la rappresentazione dei topic appare qualitativamente più solida: pur in presenza di un numero elevato di cluster, la maggior parte dei topic risulta tematicamente leggibile (*padiglione + sottomarino + agenzia spaziale, cinque per mille, volontariato, costellazioni e mitologia*), e i topic-rumore sono in proporzione minoritari.

Discussione L'introduzione del Gibbs sampling ha un effetto fortemente dipendente dalla qualità e dalla cardinalità della partizione di partenza. Il caso più favorevole è quello di UMAP + K-Means, in cui Gibbs agisce su una partizione già fitta ma poco coerente: l'aggiornamento iterativo delle assegnazioni migliora coerenza e diversità senza aumentare ulteriormente il numero di topic, fornendo l'unico caso in cui il modulo produce un beneficio chiaro su tutte le metriche rilevanti. Al contrario, sulla baseline, dove la partizione iniziale era già equilibrata e ben separata, Gibbs introduce frammentazione senza guadagno: il numero di topic quasi quadruplica e tutte le metriche di distintività regrediscono, conducendo a una rappresentazione meno informativa di quella di partenza. Sulla configurazione PCA + K-Means Gibbs interviene sui sintomi del problema strutturale già discusso, riportando coerenza, FREX e Hellinger su valori realistici e separati, ma al prezzo di una partizione con 84 topic e un'elevata frazione di topic-rumore. Al di là dei risultati quantitativi, la quadruplicazione del numero di topic costituisce un limite pratico rilevante: oltre il centinaio di topic finali, la partizione perde di fatto interpretabilità manuale, e l'analisi qualitativa diventa difficilmente sostenibile senza ulteriori passi di aggregazione o filtraggio. Il vantaggio metrico osservato sulla configurazione UMAP + K-Means va dunque letto tenendo conto del fatto che la partizione risultante conta comunque 102 topic, valore già elevato rispetto al numero di temi distinti effettivamente riconoscibili nel corpus.

Capitolo 8

Conclusioni

Il lavoro svolto ha messo a confronto tre approcci al *topic modeling* sul corpus dei post Facebook di otto musei scientifici italiani e ha sperimentato una serie di varianti della pipeline di BERTopic, con l'obiettivo di individuarne una configurazione più adatta al testo italiano. Le evidenze raccolte non permettono di concludere che una delle modifiche proposte costituisca un miglioramento netto rispetto alla *baseline*: nessuna delle varianti testate ha prodotto una partizione complessivamente più informativa, e alcune di esse hanno reso più visibili limiti che la configurazione di riferimento mascherava soltanto in parte. Più che un esito negativo, questo risultato va letto come una mappa dei punti in cui la pipeline mostra le proprie fragilità su un corpus come quello in esame, e costituisce il principale contributo della tesi.

8.1 Il rumore lessicale e la gestione dei token sconosciuti

L'osservazione più sorprendente emersa dalle sperimentazioni è la ricorrenza, fra le parole rappresentative dei topic di BERTopic, di token a basso o nullo contenuto informativo: *stopword* italiane, articoli e preposizioni in inglese, frammenti di URL (*https*, *www*, *bit*, *youtu*), identificativi alfanumerici e hashtag concatenati. Il fenomeno è risultato particolarmente marcato nelle

varianti basate sul modello di embedding monolingue italiano e nelle configurazioni in cui interveniva il Gibbs sampling, ma è già presente, in misura minore, nella *baseline*.

Si tratta di un dato che, a un primo sguardo, contrasta con la promessa di un *topic modeling* fondato su rappresentazioni contestuali. L'intero senso del passaggio da una rappresentazione *bag-of-words* a un embedding Transformer risiede proprio nella possibilità di pesare ciascun token in funzione del contesto e di attribuire un peso semantico marginale ai termini funzionali ubiqui, che in un'analisi puramente frequentistica inquinerebbero i topic. Il fatto che parole come *non*, *una*, *tuo*, o intere sequenze in inglese, riemergano nelle top-list dei topic indica che questa promessa non è completamente mantenuta sul corpus in esame.

Una possibile spiegazione del fenomeno riguarda la fase di *tokenizzazione* a monte dell'embedding. I modelli Transformer processano il testo attraverso un vocabolario di sotto-parole (*subword tokenization*, tipicamente WordPiece o BPE), in modo che parole non presenti nel vocabolario vengano scomposte in frammenti riconoscibili. Su un corpus informale e linguisticamente eterogeneo come quello dei post museali, ricco di hashtag, URL accorciati, identificativi tecnici e termini in lingue diverse, questa procedura sembra non funzionare in modo ottimale: i token sconosciuti vengono frammentati in modo poco significativo, e i frammenti risultanti perdono ancoraggio contestuale, riemergendo poi nella fase di rappresentazione c-TF-IDF (che, va ricordato, opera sui token grezzi e non sugli embedding) come elementi distintivi dei cluster. Lo stesso meccanismo spiega la sopravvivenza delle stringhe alfanumeriche e dei frammenti di URL fra le parole rappresentative: in un embedding contestuale ben calibrato, tali token dovrebbero collassare in regioni semanticamente irrilevanti dello spazio vettoriale, ma se la tokenizzazione li tratta come unità indipendenti il loro effetto si propaga lungo l'intera pipeline.

La conseguenza pratica di questa osservazione è interessante: contrariamente a quanto la letteratura su BERTopic generalmente assume, un *preprocessing* linguistico preliminare (rimozione di URL, normalizzazione di hashtag, filtraggio aggressivo dei documenti non in italiano) potrebbe rivelarsi necessario

anche per una pipeline neurale. Il fatto stesso che ciò sia necessario costituisce una piccola sorpresa metodologica, e rappresenta una delle direzioni più immediate per il proseguimento di questo lavoro.

8.2 Coerenza semantica e interpretabilità dei topic

Un'altra considerazione emersa riguarda il rapporto fra la coerenza semantica dei topic e la loro effettiva interpretabilità da un punto di vista umano. Sebbene BERTopic abbia restituito topic semanticamente più coerenti delle controparti, l'interpretazione dei topic prodotti si è rivelata meno immediata rispetto a quella offerta da LDA e, soprattutto, da STM.

Il motivo di questa apparente contraddizione potrebbe risiedere nella natura stessa dei cluster individuati da BERTopic: i campi semantici emersi sono spesso molto coerenti, ma al contempo molto specifici, al punto da richiedere uno sforzo interpretativo non trascurabile per attribuire loro un'etichetta. In quest'ottica, l'ulteriore ispezione dei documenti più rappresentativi di ciascun topic appare quasi necessaria: le parole più rappresentative, da sole, sono raramente sufficienti a chiarire la natura dell'argomento individuato dal modello. È significativo, a tal proposito, che le versioni più recenti di BERTopic abbiano integrato nella propria pipeline la possibilità di delegare a un *Large Language Model* la generazione di etichette e descrizioni per ciascun topic a partire dalle parole e dai documenti più rappresentativi.

Dall'altra parte, i modelli bayesiani hanno restituito *top-words* lessicalmente più “pulite”. Ne consegue che, anche a fronte di una coerenza semantica complessivamente inferiore a quella di BERTopic, l'interpretazione umana dei topic risulta più immediata: le parole chiave conducono in modo più diretto ad un tema, e dunque il ricorso ai documenti rappresentativi assume il ruolo di conferma più che quello di interpretazione vera e propria.

Va sottolineato, infine, che queste valutazioni dipendono molto dalla natura del corpus analizzato. Su materiali linguisticamente più strutturati (abstract scientifici, articoli di quotidiano, documenti istituzionali) è ragionevole aspettarsi che il divario fra coerenza semantica e leggibilità si riduca, e che quindi

la maggiore precisione semantica di BERTopic possa essere un beneficio. Il corpus dei post museali, con la sua brevità, l'informalità del registro, la frequente presenza di elementi para-linguistici (hashtag, emoji, URL accorciati), costituisce per ogni tecnica di *topic modeling* un banco di prova particolarmente esigente: si ritiene che parte delle difficoltà interpretative osservate sia attribuibile a queste caratteristiche più che a limiti intrinseci dei metodi.

8.3 Le varianti e l'assenza di un miglioramento netto

Nessuna delle modifiche testate ha prodotto un miglioramento sistematico della pipeline. Il modello di embedding monolingue, che a priori ci si sarebbe attesi competitivo rispetto al multilingue di default, non ha modificato in modo apprezzabile le metriche e ha anzi accentuato il problema delle parole vuote nelle rappresentazioni. Le combinazioni alternative di riduzione dimensionale e clustering hanno restituito un panorama di compromessi senza un chiaro vincitore: le configurazioni con K-Means producono coerenze elevatissime, che però l'analisi congiunta con esclusività FREX e distanza di Hellinger ha mostrato essere in larga misura artefattuali; le combinazioni con PCA o KPCA su HDBSCAN guadagnano in coerenza ma sacrificano la capacità di isolare nicchie tematiche; la *baseline* UMAP + HDBSCAN, infine, resta la più equilibrata, pur senza eccellere su nessuna singola metrica.

L'introduzione di RAKE come modulo di rappresentazione costituisce il caso più ambivalente. Su singoli topic problematici la sostituzione del c-TF-IDF ha prodotto rappresentazioni qualitativamente più ricche, sostituendo *stop-word* e identificativi con termini lessicalmente informativi; al tempo stesso, l'estrazione basata sulla struttura interna del testo ha fatto emergere altri tipi di rumore, in particolare i token web ricorrenti nei post social. Il bilancio è quindi misto: RAKE migliora la leggibilità di alcuni topic ma non risolve in maniera strutturale il problema del rumore lessicale, e in alcune configurazioni (come PCA + K-Means) ne fa anzi emergere l'artefattualità sottostante. È uno spunto promettente, che meriterebbe un approfondimento dedicato, magari combinato con il *preprocessing* discusso sopra e con una calibrazione mirata dei parametri di estrazione.

L'integrazione del Gibbs sampling, ispirata al lavoro di Zhu e Liu [32], ha confermato un limite già segnalato nel paper di riferimento: il modulo tende a far esplodere il numero di topic finali (da 27 a 106 sulla *baseline*, da 20 a 84 su PCA + K-Means). Nel paper originale questo effetto viene gestito attraverso una successiva fase di riduzione manuale dei topic, condotta dagli autori sulla base dell'ispezione qualitativa dei cluster. Nel caso del corpus museale tale strategia non è risultata praticabile vista la numerosità raggiunta (oltre cento argomenti in più configurazioni). Il miglioramento di alcune metriche osservato in un paio di configurazioni (in particolare UMAP + K-Means) va dunque ridimensionato alla luce di questo limite operativo: la partizione risultante, anche quando migliora sulle metriche, resta troppo frammentata per essere realmente utilizzabile come strumento pratico.

8.4 Direzioni future

Le evidenze raccolte indicano alcune direzioni di ricerca naturali. La prima è l'introduzione di un *preprocessing* specifico per BERTopic su corpora informali e linguisticamente eterogenei. La domanda di ricerca sottostante (perché una pipeline costruita su rappresentazioni contestuali abbia bisogno di un *preprocessing* per ottenere risultati puliti) è forse più interessante della soluzione pratica, e meriterebbe un'indagine dedicata sul comportamento della fase di tokenizzazione su corpora di questo tipo.

La seconda direzione riguarda RAKE: i risultati ottenuti sono incoraggianti ma parziali, e una sua integrazione più sistematica nella pipeline potrebbe restituire i benefici osservati senza le problematiche legate ai token web. La terza direzione concerne il Gibbs sampling, che resta uno strumento teoricamente attraente ma necessita di un meccanismo di riduzione automatica dei topic in grado di sostituire l'aggregazione manuale impiegata nel paper di riferimento: senza tale componente, l'integrazione finisce per spostare il problema della partizione anziché risolverlo.

La quarta direzione riguarda lo sviluppo di diagnostiche basate sui documenti anziché sulle parole rappresentative. Le metriche di coerenza utilizzate in questo elaborato, così come la maggior parte di quelle proposte nella lettera-

tura di riferimento, operano quasi esclusivamente sulle *top-words* di ciascun topic. Tuttavia, come osservato nella sezione precedente, l'interpretazione dei topic prodotti da BERTopic può richiedere anche il ricorso ai documenti più rappresentativi. Di conseguenza si potrebbe valutare la qualità di un topic anche attraverso misure costruite sui documenti stessi. Una possibile linea di sviluppo consisterebbe nello sfruttare gli *embeddings* dei documenti per calcolare la loro distanza dal centroide del cluster di appartenenza, eventualmente confrontandola con la distanza dai centroidi degli altri cluster. Strumenti come la similarità del coseno o il punteggio di Silhouette potrebbero offrire una misura sintetica della coesione interna di ciascun topic e della sua separazione rispetto agli altri, costruita sugli stessi oggetti che l'analista è poi chiamato a leggere per interpretare i risultati. Una diagnostica di questo tipo si presterebbe particolarmente bene a BERTopic, la cui logica di clustering nello spazio degli *embeddings* rende il calcolo immediato, ma potrebbe in linea di principio essere estesa anche ai modelli probabilistici, calcolando gli *embeddings* dei documenti in un secondo momento.

In generale, il quadro che emerge è quello di una pipeline neurale matura e flessibile, ma non necessariamente superiore agli approcci classici quando ci si confronta con una lingua a bassa disponibilità di risorse come l'italiano e con un registro testuale non standard come quello presente nei *social media*. Le promesse del paradigma contestuale, su un corpus di questo tipo, vanno verificate empiricamente e non assunte: la maggior parte del lavoro di adattamento che ci si attenderebbe di poter delegare al modello ricade, in pratica, ancora a monte. Riconoscere questo punto, e indagarne le cause, costituisce forse il contributo più solido della presente tesi.

Appendice A

Top words dei topic

A.1 Nota introduttiva

Questa appendice riporta, per ciascuno dei tre modelli di topic modeling utilizzati nell’analisi (LDA, STM e versione di base di BERTopic), le dieci parole più rappresentative di ogni topic. Le parole sono ordinate per rappresentatività decrescente all’interno del topic, secondo il criterio specifico di ciascun modello: la probabilità parola–topic (β) per il modello LDA, la probabilità di emissione (con riferimento alla misura FREX) per il modello STM e il punteggio c-TF-IDF per BERTopic.

A.2 LDA

Tabella A.2.1: Parole più rappresentative dei topic del modello LDA ($K = 20$).

Topic	Parole più rappresentative
1	leonardo, milano, vinci, tecnologia, scienza, realizzare, galleria, disegno, san, mostra
2	prenotazione, attività, gratuito, ingresso, evento, incontro, info, settembre, cambiamento, climatico

continua nella pagina successiva

Tabella A.2.1 – *continua dalla pagina precedente*

Topic	Parole più rappresentative
3	prossimo, scoprire, numero, acquistare, parola, appuntamento, on-line, tecnico, giocare, martedì
4	strumento, storico, libro, biblioteca, immagine, antico, sala, presentare, sapere, conservare
5	stella, pianeta, terra, luce, vedere, sole, sistema, solare, osservare, metro
6	scientifico, storia, collezione, tema, interattivo, laboratorio, diverso, accompagnare, settimana, porta
7	galileo, scienza, firenze, mostra, fondazione, università, tecnica, naturale, incontro, ciclo
8	ricerca, parlare, punto, campo, umano, capire, persona, vista, ruolo, fisico
9	storia, viaggio, raccontare, scoperta, percorso, dedicare, cielo, partire, curatore, podcast
10	progetto, pubblico, realizzare, centro, ambiente, collaborazione, sostenibile, servizio, giovane, fondazione
11	astronomico, brera, visita, possibile, marte, pianeta, cupola, osservatorio, schiaparelli, trovare
12	toti, sottomarino, spazio, padiglione, spaziale, visita, italiano, europea, ferroviario, vega
13	maca, laboratorio, info, bambino, euro, partecipante, costo, prenotazione, obbligatorio, domenica
14	scoprire, colore, piccolo, oggetto, creare, acqua, materiale, forma, dare, necessario
15	italiano, italia, donna, roma, culturale, comunicazione, storico, festival, polo, giugno
16	attività, partecipare, dedicare, scuola, laboratorio, famiglia, studente, ambiente, programma, prenotazione
17	trieste, scientifico, ricercatore, studio, università, notte, comune, immaginario, internazionale, scuola
18	spazio, luna, spaziale, missione, terra, satellite, bordo, astronauta, apollo, stazione

continua nella pagina successiva

Tabella A.2.1 – *continua dalla pagina precedente*

Topic	Parole più rappresentative
19	biglietto, sabato, online, realtà, scoprire, attività, virtuale, domenica, programma, digitale
20	novembre, dicembre, marzo, domenica, gennaio, venerdì, ottobre, aprire, aprile, occasione

A.3 STM

Tabella A.3.2: Parole più rappresentative dei topic del modello STM ($K = 17$).

Topic	Parole più rappresentative
1	leonardo, vinci, milano, galleria, tecnologia, scienza, san, disegno, ingegnere, realizzare
2	brera, astronomico, visita, schiaparelli, cupola, minuto, osservatorio, pianeta, marte, biglietto
3	laboratorio, maca, euro, partecipante, costo, info, domenica, bambino, prenotazione, durata
4	tecnologia, umano, scienza, ricerca, parlare, salute, punto, incontro, cambiamento, sviluppo
5	trieste, ricercatore, scientifico, università, studio, notte, scuola, internazionale, istituto, comune
6	spaziale, spazio, luna, missione, terra, satellite, viaggio, bordo, biglietto, dedicare
7	numero, giocare, prossimo, online, parola, acquistare, commento, scoprire, biglietto, rubrica
8	mostra, opera, realizzare, scientifico, biblioteca, arte, strumento, restauro, dedicare, collezione
9	maca, ambiente, visita, laboratorio, attività, dedicare, prenotazione, appuntamento, info, settimana

Tabella A.3.2 – *continua dalla pagina precedente*

Topic	Parole più rappresentative
10	cielo, stella, prenotazione, planetario, info, astronomia, attività, corso, planetariofst, pagamento
11	storia, italiano, storico, raccontare, roma, donna, culturale, comunicazione, libro, polo
12	galileo, firenze, scienza, storia, università, fondazione, scientifico, incontro, studio, tecnica
13	attività, biglietto, scoprire, online, programma, sabato, visita, realtà, virtuale, domenica
14	terra, pianeta, sole, luce, stella, solare, vedere, sistema, trovare, osservare
15	domenica, scoprire, bambino, scientifico, attività, prenotazione, gennaio, aprire, scienza, sabato
16	storia, scientifico, collezione, padiglione, sottomarino, toti, spazio, storie, ferroviario, attualità
17	progetto, scuola, studente, collaborazione, fondazione, realizzare, design, san, tema, promuovere

A.4 BERTopic

Tabella A.4.3: Parole più rappresentative dei topic del modello BERTopic ($K = 27$).

Topic	Parole più rappresentative
1	museoscienza, museo, visita, gallerie, prenotazione, 2023, online, www, sabato, org
2	calendario, gregoriano, marzo, periodo, giorni, day, primavera, summer, 1582, anni
3	francobollo, francobolli, carta, centenario, poligrafico, postale, filatelia, stampa, patrimonio, nelle

Tabella A.4.3 – *continua dalla pagina precedente*

Topic	Parole più rappresentative
4	sounds, sound, sonora, suono, sonore, musicali, orchestra, suoni, musica, vibrare
5	videogiochi, videogame, gaming, game, gyga, play, evento, biglietti, bandowonder, paolo
6	storico, storia, fascismo, nobel, sterminio, ebrei, liberazione, popoli, morte, 1945
7	mitologia, principesse, greca, leggende, planetario, planetariofst, regine, stelle, personaggi, mostri
8	zodiaco, mitologiche, zodiacale, lucertole, stella, scorpione, leone, erigone, mito, _____
9	autrice, comitato, libro, bibliocaffèsuigeneris, invitano, donne, decreto, costituzione, nella, storia
10	bicicletta, bici, pedali, triciclo, acomeambiente, sicurezza, ruote, metteremo, propositi, trasporto
11	chimiche, chimica, scienza, chemioluminescenza, elettroni, sostanze, sostanza, sperimentare, innovativo, creatività
12	anni, 1969, internet, elettronici, elettronico, digitali, trasmettitore, anniversario, pubblica, immaginarioscientifico
13	bubbles, bolle, bolla, 2023, anni, iridescenti, magico, storieaportechiuse, possono, magia
14	climatici, università, climatica, campus, climatico, museum, scientifica, progetto, ambiente, studi
15	neuroscienze, olimpiadi, cervelli, cervello, olimpico, studentesse, partecipano, studenti, competizione, gara
16	agenda, 2030, 2023, futuro, obiettivi, calendario, obiettivo, proporlo, 2024, trasformarti
17	fiscale, contribuente, redditi, spesa, tassa, donazioni, museo, tasse, gratuito, iniziativa
18	volare, aerea, paracadute, paracadutista, uccelli, sperimentarne, spiraliforme, elicottero, prototipo, natura
19	gift, museoscienza, museo, regali, card, ingresso, sconto, sottomarinò, gratuito, gallerie

Tabella A.4.3 – *continua dalla pagina precedente*

Topic	Parole più rappresentative
20	roma, romarama2020, viaggineiparaggi, turismo, luoghi, viaggio, della, viaggi, partecipare, bossanova
21	com, http, mailartista, postale, posta, creativa, mailart, arte, link, museorisorgimentomilano
22	museiwelcome, museiwelcomefirenze, inclusività, accessibile, ciechi, consapevolezza, animato, celebre, persone, inclusive
23	algoritmi, algoritmo, intelligenza, tecnologie, pubblica, consigliano, artificiale, online, interattiva, app
24	pesci, pesce, marini, delfini, sea, oceano, fish, squali, ocean, tartarughe
25	biotecnologie, genoma, ricercatore, biologico, mutazioni, mutazione, genetiche, molecolare, genetica, molecola
26	orologio, mezzanotte, mezzogiorno, notturno, secolo, museumatnight, museumatnightchallenge, luna, notte, solare
27	animatrici, animatori, biotecnologie, museoscienza, laboratori, inventors, laboratorio, museo, scienze, accademia

Appendice B

Valutazione qualitativa dei topic

B.1 Nota introduttiva

Questa appendice riporta, per ciascun modello, i punteggi grezzi assegnati a ogni topic nel corso della valutazione qualitativa umana effettuata dal professor Crabu descritta nella Sezione 6, da cui sono stati ricavati gli indicatori aggregati presentati nel capitolo dei risultati. Per ogni topic sono indicati: la *Coerenza* semantica (scala 0–5), l'*Interpretabilità* (scala 0–5), il numero di *Parole intruse* sulle dieci parole principali (conteggio 0–10), la *Distintività* rispetto agli altri topic (scala 0–5) e l'elenco dei *Topic sovrapposti*, ossia gli identificativi dei topic giudicati ridondanti rispetto a quello considerato (*nessuno* se non ne sono stati rilevati). Viene inoltre fornito un grafo per ogni modello che mostra i gruppi di topic distinti individuati a partire dalle segnalazioni del professor Crabu. Ogni grafo è stato costruito tenendo isolati i topic che presentavano una distintività pari a 4 o 5 e creando archi in base alle sovrapposizioni indicate per i topic con distintività inferiore o uguale a 3.

B.2 LDA

Tabella B.2.1: Valutazione qualitativa dei topic del modello LDA ($K = 20$).

Topic	Coer.	Interp.	Intrusi	Distint.	Topic sovrapposti
1	5	5	1	5	nessuno
2	4	4	2	3	2, 3, 13, 15, 16, 19, 20
3	2	2	5	3	2, 3, 13, 15, 16, 19, 20
4	3	2	4	5	nessuno
5	5	5	1	3	11, 18
6	3	2	3	4	10
7	5	3	0	5	nessuno
8	2	0	4	5	nessuno
9	3	0	1	5	nessuno
10	3	0	1	4	6
11	5	4	1	3	5, 18
12	5	4	1	5	nessuno
13	3	0	0	3	2, 3, 13, 15, 16, 19, 20
14	3	2	2	5	nessuno
15	3	0	4	3	2, 3, 13, 15, 16, 19, 20
16	4	3	0	3	2, 3, 13, 15, 16, 19, 20
17	4	5	0	5	nessuno
18	5	5	0	3	5, 11
19	3	2	2	3	2, 3, 13, 15, 16, 19, 20
20	5	0	0	3	2, 3, 13, 15, 16, 19, 20

Numero di topic distinti effettivi indicato dal valutatore: **11**.

B.3 STM

Tabella B.3.2: Valutazione qualitativa dei topic del modello STM ($K = 17$).

Topic	Coer.	Interp.	Intrusi	Distint.	Topic sovrapposti
1	5	5	1	4	4, 12
2	4	5	0	3	6, 10, 14
3	4	2	0	3	9, 13, 15
4	3	3	0	3	5, 11, 17

continua nella pagina successiva

Tabella B.3.2 – *continua dalla pagina precedente*

Topic	Coer.	Interp.	Intrusi	Distint.	Topic sovrapposti
5	3	3	0	4	4, 17
6	5	4	1	3	10, 14, 16
7	1	0	6	2	3, 13, 15
8	3	0	1	4	11, 16
9	2	2	0	3	3, 13, 15
10	5	3	1	3	2, 6, 14
11	1	0	1	4	4, 8
12	5	4	0	4	1, 4
13	3	5	0	2	3, 7, 9, 15
14	5	3	0	3	2, 6, 10
15	3	4	0	3	3, 9, 13
16	5	5	1	3	6, 8
17	3	3	2	4	4, 5

Numero di topic distinti effettivi indicato dal valutatore: **8**.

B.4 BERTopic

Tabella B.4.3: Valutazione qualitativa dei topic del modello BERTopic ($K = 27$).

Topic	Coer.	Interp.	Intrusi	Distint.	Topic sovrapposti
1	5	5	3	2	19, 27
2	5	0	2	4	16, 26
3	5	1	1	4	21
4	5	2	1	5	nessuno
5	5	0	2	4	23
6	5	4	1	4	9
7	5	2	1	3	8
8	5	0	1	3	7, 26
9	1	0	5	3	6
10	5	0	1	4	14, 19
11	5	2	1	3	15, 25
12	4	0	2	3	5, 23

Tabella B.4.3 – *continua dalla pagina precedente*

Topic	Coer.	Interp.	Intrusi	Distint.	Topic sovrapposti
13	2	0	4	4	27
14	5	5	0	3	16, 24
15	2	2	1	4	11, 25
16	0	0	6	3	2, 14
17	5	3	0	4	19
18	4	1	1	4	10
19	1	0	5	2	1, 17, 27
20	5	2	2	4	22
21	1	0	6	3	3, 22
22	0	0	6	4	20, 21
23	3	1	1	3	5, 12
24	5	3	1	5	nessuno
25	5	4	0	4	11, 15, 27
26	5	4	1	3	2, 8
27	3	1	1	3	1, 19, 25

Numero di topic distinti effettivi indicato dal valutatore: **20**.

B.5 Grafi di sovrapposizione dei topic

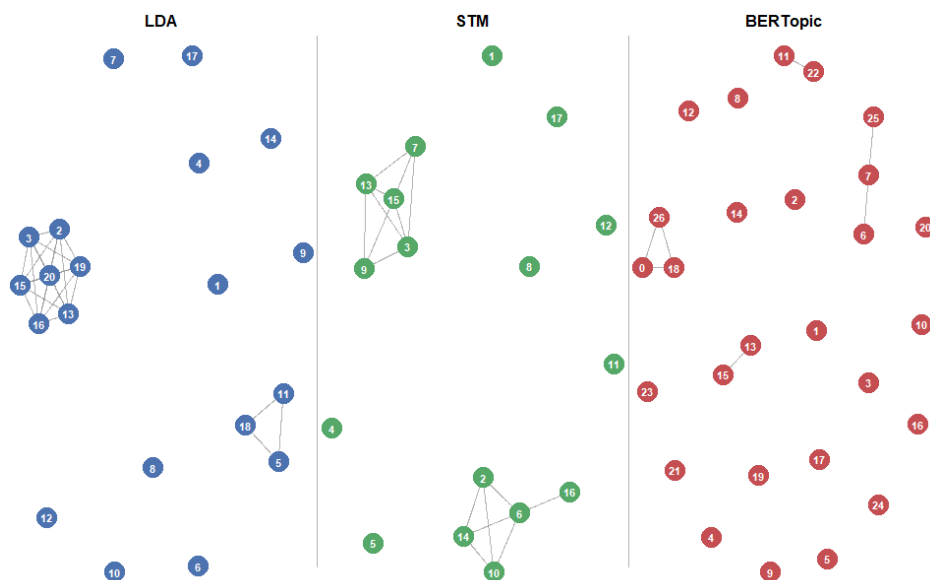


Figura B.1: Grafi di sovrapposizione tematica tra i topic dei tre modelli (LDA, STM, BERTopic). Ciascun nodo rappresenta un topic e ciascun arco una sovrapposizione semantica rilevata dal valutatore. I nodi con indice di distintività pari a 4 o superiore sono stati tenuti isolati. I gruppi di nodi connessi corrispondono ai cluster tematici.

Bibliografia

- [1] R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy, *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations*, in Advances in Knowledge Discovery and Data Mining (PAKDD), Lecture Notes in Computer Science, vol. 6118, pp. 391–402, Springer, 2010.
- [2] J. M. Bischof, E. M. Airolidi, *Summarizing Topical Content with Word Frequency and Exclusivity*, in Proceedings of the 29th International Conference on Machine Learning (ICML), pp. 201–208, 2012.
- [3] D. M. Blei, A. Y. Ng, M. I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
- [4] D. M. Blei and J. D. Lafferty, “Correlated Topic Models,” in *Advances in Neural Information Processing Systems 18*, pp. 147–154, 2006.
- [5] D. M. Blei, J. D. Lafferty, *A Correlated Topic Model of Science*, The Annals of Applied Statistics, vol. 1, no. 1, pp. 17–35, 2007.
- [6] J. Boyd-Graber, Y. Hu, D. Mimno, *Applications of Topic Models*, Foundations and Trends in Information Retrieval, vol. 11, no. 2–3, pp. 143–296, 2017.
- [7] R. J. G. B. Campello, D. Moulavi, J. Sander, *Density-Based Clustering Based on Hierarchical Density Estimates*, in Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Lecture Notes in Computer Science, vol. 7819, pp. 160–172, 2013.

- [8] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, *A Density-Based Method for Adaptive LDA Model Selection*, Neurocomputing, vol. 72, no. 7–9, pp. 1775–1781, 2009.
- [9] CrowdTangle, *Academics & Researchers*, CrowdTangle Help Center, 2024. <https://web.archive.org/web/20240315004509/https://help.crowdtangle.com/en/collections/2981118-academics-and-researchers>.
- [10] dbmdz, *BERT base Italian uncased*, Hugging Face, 2024. Disponibile all'indirizzo: <https://huggingface.co/dbmdz/bert-base-italian-uncased>.
- [11] R. Deveaud, E. SanJuan, P. Bellot, *Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval*, Document Numérique, vol. 17, no. 1, pp. 61–84, 2014.
- [12] A. B. Dieng, F. J. R. Ruiz, D. M. Blei, *Topic Modeling in Embedding Spaces*, Transactions of the Association for Computational Linguistics, vol. 8, pp. 439–453, 2020.
- [13] J. Eisenstein, A. Ahmed, E. P. Xing, *Sparse Additive Generative Models of Text*, in Proceedings of the 28th International Conference on Machine Learning (ICML), pp. 1041–1048, 2011.
- [14] M. Garmur, G. King, Z. Mukerjee, N. Persily, B. Silverman, *CrowdTangle Platform and API*, Harvard Dataverse, V2, 2019. <https://doi.org/10.7910/DVN/SCCQYD>.
- [15] T. L. Griffiths, M. Steyvers, *Finding Scientific Topics*, Proceedings of the National Academy of Sciences, vol. 101, suppl. 1, pp. 5228–5235, 2004.
- [16] M. Grootendorst, *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*, arXiv preprint arXiv:2203.05794, 2022.
- [17] E. Hellinger, *Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen*, Journal für die reine und angewandte Mathematik, vol. 136, pp. 210–271, 1909.

BIBLIOGRAFIA

- [18] T. Hofmann, *Probabilistic Latent Semantic Indexing*, in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57, 1999.
- [19] Hotelling, H. (1933). *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 24(6), 417–441; 24(7), 498–520.
- [20] L. McInnes, J. Healy, J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv preprint arXiv:1802.03426, 2018.
- [21] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, A. McCallum, *Optimizing Semantic Coherence in Topic Models*, in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 262–272, 2011.
- [22] N. Murzintcev, *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*, R package version 1.0.2, 2016. Disponibile all’indirizzo: <https://CRAN.R-project.org/package=ldatuning>
- [23] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, *Automatic Evaluation of Topic Coherence*, in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pp. 100–108, 2010.
- [24] N. Reimers, I. Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982–3992, 2019.
- [25] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, D. G. Rand, *Structural Topic Models for Open-Ended Survey Responses*, American Journal of Political Science, vol. 58, no. 4, pp. 1064–1082, 2014.
- [26] M. E. Roberts, B. M. Stewart, D. Tingley, *stm: An R Package for Structural Topic Models*, Journal of Statistical Software, vol. 91, no. 2, pp. 1–40, 2019.

- [27] M. Röder, A. Both, A. Hinneburg, *Exploring the Space of Topic Coherence Measures*, in Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM), pp. 399–408, 2015.
- [28] A. Sciandra, M. Trevisani, A. Tuzzi, *Temporal trends and presidential traits in the Italian end-of-year addresses: comparing and contrasting KBS and STM results*, *Quality & Quantity*, vol. 59, pp. 905–935, 2025. DOI: <https://doi.org/10.1007/s11135-024-01959-x>
- [29] M. Straka, J. Hajič, J. Straková, *UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing*, in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), pp. 4290–4297, 2016.
- [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Attention Is All You Need*, in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
- [31] X. Wu, X. Dong, T. Nguyen, A. T. Luu, *Effective Neural Topic Modeling with Embedding Clustering Regularization*, in Proceedings of the 40th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol. 202, pp. 37335–37357, 2023.
- [32] Y. Zhu, Y. Liu, *Gibbs-BERTopic: A Hybrid Approach for Short Text Topic Modeling*, IEEE Access, DOI: 10.1109/ACCESS.2025.3552221, 2025.