POLITECNICO DI TORINO

Dipartimento di Ingegneria Gestionale e della Produzione

Corso di Laurea Magistrale in Ingegneria Gestionale

Gestione e progettazione dei servizi digitali

Politecnico di Torino

Modellazione e analisi dinamica dei topic nella Digital Voice of Customer con BERTopic: applicazione al caso studio Spotify

Relatore:

Prof. Federico Barravecchia

Candidata:

Alice Severi

1. Indice

1. Indice 2
1.1 Indice delle figure 3
1.2 Indice delle tabelle 4
2. Introduzione 6
3. Revisione della letteratura 8
3.1. Digital Voice of Customer 8
3.1.1. Le implicazioni manageriali della Digital VoC 10
3.2. Introduzione al Topic Modeling 12
3.3. Dynamic Topic Modeling 19
3.4. BERTopic 23
3.4.1. Applicazioni di BERTopic 29
4. Metodologia di ricerca 32
4.1. Descrizione generale del processo di analisi 32
4.2. Setup ambiente, raccolta e pre-processing dei dati 34
4.3. Implementazione del modello 38
4.3.1. Allenamento modello statico 38
4.3.2. Labeling e raffinamento dei topic 43
4.3.3. Visualizzazione dei topic 47
4.4. Validazione del modello statico 50
4.4.1. Validazione non supervisionata del modello statico 50
4.4.2. Validazione supervisionata del modello statico 54
4.5. Analisi dinamica dei topic 55
4.5.1. Validazione supervisionata dell'analisi dinamica 60
5. Applicazione del modello alle recensioni della piattaforma Spotify 62
5.1. Descrizione del dataset 62
5.2. Applicazione del modello al dataset 64
5.3. Labeling e merging dei topic 67
5.4. Visualizzazioni del modello statico 69
5.5. Metriche di validazione 74
5.6. Applicazione dell'analisi dinamica ex-post 75
5.7. Risultati dell'analisi dinamica 76
6. Discussione dei risultati potenziali e implicazioni manageriali 85
6.1. Dal monitoraggio reattivo alla diagnostica proattiva 85

6.2 Esempi di discontinuità e trend 86
6.3 Visualizzazioni per il decision-making 86
7. Conclusioni 89
8. Bibliografia 91
Appendice A 96
Appendice B 102

1.1 Indice delle figure

Figura 1 - Framework concettuale della relazione tra Digital VoC e tre aspetti aziendali (Sharma et al., 2025). 11
Figura 2 - Processo generale di Topic Modeling (Hankar et al., 2025) 13
Figura 3 - Schema gerarchico delle relazioni tra text mining e varie tecniche di Topic Modeling 14
Figura 4 - Flusso di attività del STM (Barravecchia et al., 2020). 18
Figura 5 - Rappresentazione grafica di un DTM (per tre fette temporali). I parametri naturali βt,k e αt evolvono nel tempo (Blei and Lafferty, 2006). 21
Figura 6 - Componenti del BERTopic (Hananto, 2023). 27
Figura 7 - Matrice di Similarità (Alamsyah and Girawan, 2023). 30
Figura 8 - Flowchart della metodologia di ricerca 33
Figura 9 - Diagramma di flusso di pre-processing e modello statico 50
Figura 10 - Diagramma di flusso della validazione del modello 54
Figura 11 - Diagramma di flusso dell'analisi dinamica ex-post 60
Figura 12 - Flowchart generale degli step del processo 61
Figura 13 - Frequenza delle recensioni per anno nel dataset di Spotify 62
Figura 14 - Distribuzione cumulata delle recensioni in base alla lunghezza conteggiata in parole nel dataset di Spotify 63
Figura 15 - Distribuzione delle recensioni per topic post-merge nel dataset di Spotify 69
Figura 16 - Intertopic Distance Map del dataset di Spotify 70
Figura 17 - Similarity Matrix del dataset di Spotify 71
Figura 18 - Punteggi c-TF-IDF per le parole chiave dei topic 73
Figura 19 - Percentuale di recensioni appartenenti al topic "Limitazioni della versione free" negli anni 76
Figura 20 - Parole rappresentative del topic “Limitazioni della versione free” per anno 77

Figura 21 - Percentuale di recensioni appartenenti al topic "Affidabilità e stabilità dell'app" negli anni 77
Figura 22 - Parole rappresentative del topic “Affidabilità e stabilità dell’app” per anno 78
Figura 23 - Percentuale di recensioni appartenenti al topic "Riproduzione di podcast" negli anni 79
Figura 24 - Parole rappresentative del topic “Riproduzione di podcast” per anno 79
Figura 25 - Percentuale di recensioni appartenenti al topic "Accesso ai lyrics" negli anni 80
Figura 26 - Parole rappresentative del topic “Accesso ai lyrics” per anno 81
Figura 27 - Percentuale di recensioni appartenenti al topic "Interfaccia utente" negli anni 81
Figura 28 - Parole rappresentative del topic “Interfaccia utente” per anno 82
Figura 29 - Percentuale di recensioni appartenenti al topic "Presenza del widget" negli anni 82
Figura 30 - Parole rappresentative del topic "Presenza del widget” per anno 83
Figura 31 - Percentuale di recensioni appartenenti al topic "Limitazione sugli audiolibri" negli anni 83
Figura 32 - Parole rappresentative del topic "Limitazione sugli audiolibri” per anno 83
Figura 33 - Percentuale di recensioni appartenenti al topic "Presenza di annunci" negli anni 102
Figura 34 - Parole rappresentative del topic "Presenza di annunci” per anno 102
Figura 35 - Percentuale di recensioni appartenenti al topic "Motivazione del rating" negli anni 103
Figura 36 - Parole rappresentative del topic "Motivazione del rating” per anno 103
Figura 37 - Percentuale di recensioni appartenenti al topic "Aggiornamenti dell'app" negli anni 103
Figura 38 - Parole rappresentative del topic "Aggiornamenti dell'app” per anno 104
Figura 39 - Percentuale di recensioni appartenenti al topic "Esperienza complessiva" negli anni 104
Figura 40 - Parole rappresentative del topic "Esperienza complessiva” per anno 104
Figura 41 - Percentuale di recensioni appartenenti al topic "Disponibilità del catalogo musicale" negli anni 105
Figura 42 - Parole rappresentative del topic "Disponibilità del catalogo musicale” per anno 105

1.2 Indice delle tabelle

Tabella 1 - Pre-processing: input, operazioni e output 37
Tabella 2 - Modello statico: input, operazioni e output 42
Tabella 3 - Labeling: input, operazioni e output 45
Tabella 4 - Visualizzazione: input, operazioni e output 49
Tabella 5 – Validazione del modello: input, operazioni e output 53
Tabella 6 – Modello dinamico: input, operazioni e output 58

2. Introduzione

Nel contesto odierno, la Digital Voice of Customer (Digital VoC) raccoglie opinioni, commenti e valutazioni spontanee degli utenti su blog, social network, forum e portali di recensioni online, come TripAdvisor, Yelp.com, Amazon, ecc. Questa evoluzione del feedback del cliente è resa possibile dall'avvento del Web 2.0, che ha trasformato l'utente da semplice fruitore ad autore di contenuti, rendendolo promotore o detrattore dei servizi o dei prodotti offerti dalle aziende. La Digital Voice of Customer ha ottenuto un grande successo come fonte di informazione poiché è gratuita e facilmente accessibile. Essa si è così affermata come valida alternativa alle tradizionali tecniche di customer research, offrendo insight spontanei e in tempo reale sui bisogni dei consumatori. Inoltre, la VoC digitale non è soggetta a bias perché si basa su feedback spontanei degli utenti.

La Digital Voice of Customer offre un flusso continuo di informazioni non strutturate e caratterizzate dal linguaggio naturale, che possono rivelare aspetti critici o opportunità per migliorare il prodotto o il servizio offerto. La possibilità di cogliere insight in tempo reale sulle percezioni dei clienti riduce il tempo necessario a interventi correttivi e migliora la customer satisfaction, identificando tempestivamente bisogni non soddisfatti, problematiche ricorrenti e opportunità di cross-selling o up-selling.

Per le organizzazioni che ricevono enormi volumi di feedback da fonti eterogenee, è diventato cruciale tradurre quei testi in insight azionabili. Tuttavia, la natura stessa dei dati testuali, spesso colloquiali, frammentati e in continua espansione, rende l'analisi manuale difficilmente praticabile. Per rispondere a questa necessità, sono nati gli algoritmi di Topic Modeling, una branca dell'elaborazione del linguaggio naturale che permette di estrarre automaticamente i temi latenti da grandi corpora di documenti. Modelli classici, come il Latent Dirichlet Allocation (LDA) o lo Structural Topic Model (STM), si sono dimostrati estremamente efficaci nell'organizzare i feedback in categorie tematiche, permettendo di mappare le dimensioni della qualità percepita.

Il limite di questi approcci tradizionali risiede però nell'incapacità di catturare la dinamicità dei feedback. Le recensioni online riflettono infatti un contesto in perenne mutamento. Le percezioni degli utenti rispondono a variazioni stagionali, aggiornamenti del servizio, eventi critici o modifiche strutturali dell'offerta.

Per affrontare questa esigenza, la presente ricerca adotta un approccio basato su BERTopic, un modello di Topic Modeling neurale, combinato con un'analisi temporale *ex-post* dei topic individuati.

L’obiettivo è ricostruire l’evoluzione dei temi per comprendere non solo cosa dicono gli utenti, ma anche come il loro linguaggio e le loro percezioni del servizio o del prodotto cambiano nel tempo. A differenza dei Dynamic Topic Model classici, che impongono meccanismi di continuità potenzialmente in grado di attenuare shock improvvisi, l’approccio proposto consente di preservare le discontinuità tipiche della Digital Voice of Customer, mantenendo al contempo uno spazio semantico comune che rende i topic confrontabili tra epoche diverse.

Tale scelta metodologica è finalizzata a massimizzare la stabilità e la coerenza semantica dei topic identificati. Affidarsi esclusivamente all'automazione dinamica del modello può, in alcuni contesti, introdurre inesattezze legate a fluttuazioni lessicali temporanee. Al contrario, isolare prima i temi fondamentali tramite il robusto clustering neurale permette di operare una validazione manuale e un filtraggio dei cluster più rigoroso. Questo garantisce che l'analisi dell'evoluzione del linguaggio della Voice of Customer poggi su basi semantiche verificate, mantenendo una trasparenza totale nel passaggio dalla scoperta del tema alla sua distribuzione cronologica.

L’obiettivo principale della tesi è quindi verificare l’efficacia del framework proposto nel monitorare l’evoluzione dei topic nel tempo e nel trasformare la Digital Voice of Customer in informazioni utili a supportare decisioni di gestione del servizio.

Il lavoro è strutturato come segue: il secondo capitolo offre una revisione della letteratura sulla Digital VoC e sullo stato dell’arte delle tecniche di Topic Modeling, con un focus sui modelli dinamici e sul funzionamento di BERTopic. Il terzo capitolo descrive la metodologia della ricerca, illustrando la costruzione del dataset, le fasi di pre-processing, il flusso di analisi per l’estrazione dei topic e l’impostazione dell’analisi dinamica. Nel quarto capitolo viene presentato il caso di studio applicato alle recensioni degli utenti di Spotify, analizzando i risultati sia dal punto di vista statico che dinamico. Infine, le conclusioni discutono le implicazioni manageriali, i limiti dello studio e le possibili traiettorie di ricerca futura.

3. Revisione della letteratura

3.1. Digital Voice of Customer

La Voice of Customer (VoC) rappresenta l'insieme dei pareri, dei commenti, delle aspettative e delle percezioni dei clienti nei confronti di un servizio, di un prodotto o dell'azienda stessa. Costituisce uno strumento fondamentale per comprendere il grado di soddisfazione del cliente e per individuare i determinanti che influenzano la percezione della qualità (Aguwa et al., 2012). La VoC può essere raccolta tramite diverse modalità, sia tradizionali che digitali, e viene utilizzata per orientare le decisioni strategiche e operative, contribuendo così a migliorare l'efficacia delle politiche aziendali e la qualità dei prodotti o servizi (Griffin and Hauser, 1993).

Tradizionalmente, la raccolta delle informazioni relative alla VoC si è basata su metodi che prevedono un'interazione diretta con i clienti. Tra le principali tecniche utilizzate vi sono interviste, questionari, sondaggi e focus group. Questi strumenti, seppur efficaci, presentano alcune restrizioni, come la limitatezza dei campioni analizzati e la periodicità di aggiornamento delle informazioni, oltre alla spesa elevata in termini di personale e tempo (Palese and Usai, 2018). Sono inoltre presenti ulteriori limitazioni, tra cui la possibile influenza della soggettività degli esperti nella selezione iniziale degli item, con il rischio di trascurare aspetti rilevanti, e la presenza di potenziali errori nelle risposte, spesso difficili da rilevare (Mastrogiacomo et al., 2021). Queste limitazioni, unite alla crescente digitalizzazione, hanno aperto la strada a nuove modalità di raccolta dati, più rapide e più rappresentative, tra cui la Digital Voice of Customer (Digital VoC).

La Digital VoC supera le limitazioni legate alla raccolta dati classica grazie agli User-Generated Content (UGC). Quest'ultimi sono contenuti generati spontaneamente dagli utenti, provenienti da social media, recensioni online, forum e chat, riconosciuti come fonti autentiche dei feedback dei clienti (Barravecchia et al., 2022). Gli User-Generated Content si caratterizzano per tre aspetti fondamentali. Il primo è il contributo personale e lo sforzo creativo: i consumatori partecipano attivamente nel condividere le proprie opinioni ed esperienze, apportando un contributo diretto e originale al contenuto generato. In secondo luogo, il contenuto prodotto deve essere accessibile al pubblico, o quantomeno a un gruppo rilevante di persone, consentendo così la condivisione di informazioni e opinioni in un contesto più ampio e contribuendo al dibattito collettivo. Infine, per essere considerato User-Generated Content, il contenuto deve essere creato al di fuori di contesti professionali o obblighi lavorativi, garantendo che il feedback sia spontaneo e autentico, privo di influenze dirette da parte di aziende o entità professionali (Baier et al., 2025; Santos, 2022).

L’ampio volume di dati disponibili consente di analizzare campioni molto estesi, garantendo una copertura rappresentativa del pubblico di riferimento e una solida base statistica per gli insight. Un ulteriore vantaggio della Digital VoC è la disponibilità continua di feedback: il flusso costante di nuove opinioni assicura dati sempre aggiornati, eliminando i vincoli temporali tipici delle indagini periodiche (Donald et al., 2024). L’affidabilità del feedback digitale è maggiore rispetto ai metodi tradizionali poiché è meno soggetta a bias, visto che gli UGC riflettono esperienze genuine non orientate da domande strutturate dall’azienda. È, tuttavia, importante sottolineare che la Digital VoC introduce altri bias, come quelli legati alla rappresentatività spontanea del campione: il fenomeno per il quale scrivono recensioni solo gli utenti che sono particolarmente soddisfatti o particolarmente insoddisfatti del prodotto o servizio (Park et al., 2018). L’automazione offerta dalle tecniche di text mining e Natural Language Processing (NLP) riduce sensibilmente i tempi e i costi di analisi: Sentiment Analysis e Topic Modeling, per esempio, permettono di estrarre in modo efficiente il sentiment, le opinioni, i temi emergenti e le strutture tematiche latenti all’interno di grandi volumi di testo (Hankar et al., 2025; Mustak et al., 2024).

Negli ultimi anni, la Digital VoC, con la crescita costante del volume delle recensioni online, ha dato origine a grandi moli di dati testuali non strutturati. Basti pensare alle ricerche che evidenziano come città turistiche di rilievo presentino un numero medio di recensioni per ristorante molto elevato: Budapest raggiunge una media di 685 recensioni per locale, seguita da Roma con 645. Anche in città con un minore afflusso turistico, come Lussemburgo, i numeri rimangono significativi, con una media di circa 270 recensioni per ristorante (Lupșa-Tătaru et al., 2023).

Questa esplosione di dati è resa possibile dalla facilità di accesso e dalla spontaneità dei feedback digitali. Tuttavia, comporta anche importanti sfide legate alla scalabilità, alla gestione del “rumore” e alla velocità di aggiornamento. Le tecniche manuali non sono più sostenibili per elaborare milioni di documenti; pertanto, è necessario ricorrere a tecniche come il Natural Language Processing e il text mining che supportino la gestione di grandi volumi di dati (Li et al., 2022; Mustak et al., 2024). Inoltre, non tutto l’UGC risulta rilevante o di qualità uniforme, rendendo cruciale l’implementazione di filtri e metriche di validazione per distinguere il segnale dal rumore. La pulizia del dataset e la successiva verifica tramite le metriche di validazione sono necessarie a causa della sempre maggiore presenza di recensioni false e di contenuti generati da bot, che potrebbero inquinare i risultati. Infine, la rapida generazione di nuovi feedback richiede modelli flessibili in grado di adattarsi dinamicamente ai

cambiamenti (Wani et al., 2024). Questo contesto conferma la necessità di modelli che consentano di analizzare grandi volumi di UGC e di interpretare le variazioni dei contenuti nel tempo.

3.1.1. Le implicazioni manageriali della Digital VoC

Dal punto di vista manageriale, la Digital Voice of Customer rappresenta una leva strategica fondamentale per affrontare le sfide poste dall'attuale contesto competitivo, caratterizzato da rapidi cambiamenti nei gusti dei consumatori, nelle tecnologie, nell'economia e nei fattori macro-ambientali. I dati derivanti dalla Digital VoC non solo influenzano i processi decisionali di acquisto dei singoli utenti ma guidano anche i manager delle organizzazioni nelle decisioni strategiche, in quanto rappresentano una fonte importante di informazioni non filtrate sul cliente (Palese and Usai, 2018; Srinivas and Ramachandiran, 2020).

L'integrazione della Digital VoC nei processi aziendali consente di raggiungere l'eccellenza operativa attraverso l'identificazione mirata di inefficienze e la gestione proattiva di problemi operativi, come dimostrato dal caso di una grande azienda del retail riportato dall'*Harvard Business Review*. L'azienda ha ridotto del 25% i tempi di consegna e del 30% i costi operativi grazie a insights generati tramite la Digital VoC (Sharma et al., 2025). Si sottolinea, tuttavia, che per trarre valore da queste informazioni, i manager devono anche valutare la qualità e l'autenticità delle fonti.

Dal punto di vista dell'innovazione, la Digital VoC consente un approccio agile allo sviluppo di nuovi prodotti e servizi, basato sull'ascolto continuo del cliente. L'analisi dei feedback consente di individuare i punti di forza e le criticità di prodotti o servizi, orientando gli investimenti in ricerca e sviluppo verso le caratteristiche più apprezzate e intervenendo tempestivamente per correggere eventuali difetti (Subhashini et al., 2021). Un esempio emblematico di sviluppo prodotto è quello del gruppo LEGO, che ha coinvolto attivamente la propria community nel processo di co-creazione, portando alla nascita di nuove linee di prodotto allineate ai desideri dei propri clienti. In linea con questo approccio, Accenture ha mostrato che le aziende che integrano la Digital VoC nella strategia di innovazione hanno il doppio delle probabilità di successo nel lancio di nuovi prodotti (Sharma et al., 2025). Questo tipo di coinvolgimento diretto dei clienti promuove la customer-centricity e il cliente viene considerato un collaboratore attivo nel processo di creazione di valore.

L'impiego della Digital VoC favorisce l'adattamento a contesti dinamici, grazie alla capacità di fornire segnali precoci sui cambiamenti di preferenze e sulle nuove tendenze emergenti. Deloitte ha dimostrato che le aziende che sfruttano la Digital VoC sono più agili e flessibili nelle loro operazioni, potendo modificare rapidamente le strategie in risposta a segnali di mercato (Melzner et al., 2023; Sharma et

al., 2025). L'adattamento strategico e la personalizzazione dinamica dell'offerta sono resi possibili dall'analisi continua del sentiment, dei comportamenti e delle aspettative dei clienti.

In letteratura, è stato dimostrato tramite modelli di equazioni strutturali l'esistenza di una relazione causale positiva e significativa tra la Digital VoC e i tre aspetti chiave della performance aziendale appena discussi: favorisce l'efficienza operativa, stimola la capacità di guidare l'innovazione e migliora l'adattabilità dell'organizzazione a contesti dinamici (Sharma et al., 2025), come mostrato in Figura 1.

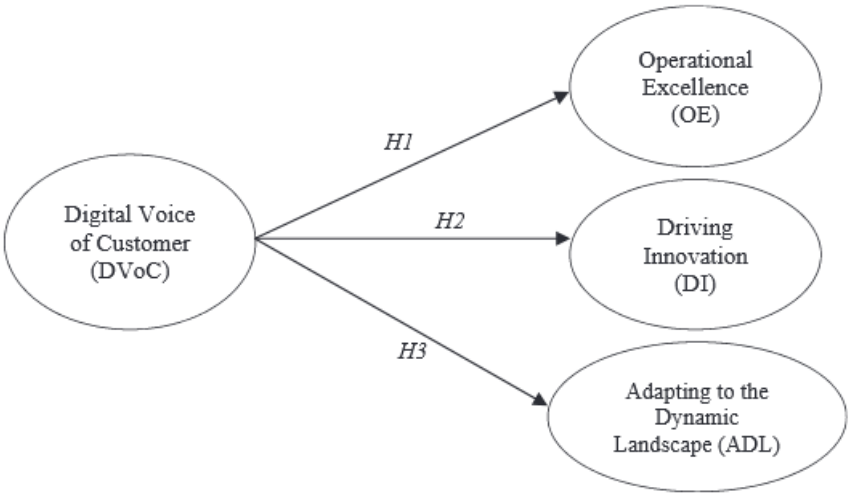

Figura 1 - Framework concettuale della relazione tra Digital VoC e tre aspetti aziendali (Sharma et al., 2025).

Gli insight derivati dall'analisi degli UGC supportano anche le strategie di marketing e di comunicazione, permettendo di segmentare il mercato in base ai bisogni emergenti, personalizzare i messaggi promozionali e progettare campagne più efficaci, con un conseguente aumento del tasso di engagement e della fedeltà dei clienti. In tal senso, la Digital VoC permette di superare la segmentazione demografica tradizionale, favorendo una segmentazione comportamentale e psicologica basata sui reali bisogni espressi spontaneamente dall'audience. Inoltre, la Digital VoC facilita il monitoraggio della reputazione online: grazie al costante flusso di opinioni, le aziende possono rilevare tempestivamente eventuali crisi o trend negativi e attivare prontamente piani di risposta e gestione delle crisi sui social media e sulle piattaforme digitali (Mastrogiacomo et al., 2021). Tuttavia, l'efficacia della Digital VoC può essere limitata da bias nei dati, scarsa rappresentatività delle piattaforme utilizzate o un'eccessiva focalizzazione su insight operativi a scapito di una visione strategica di lungo termine (Sharma et al., 2025).

Queste applicazioni strategiche dimostrano come la Digital VoC non sia semplicemente uno strumento di misurazione della soddisfazione, ma un fattore chiave per l'innovazione, la competitività e la capacità di risposta rapida alle evoluzioni del mercato.

3.2. Introduzione al Topic Modeling

Visti i volumi sempre più consistenti di dati generati dalla Digital VoC e vista la composizione non strutturata dei testi che la compongono, la ricerca si è orientata verso lo sviluppo e l'impiego di metodi capaci di analizzare efficacemente i contenuti, con particolare attenzione all'individuazione dei temi discussi nei dataset, chiamati topic (Mastrogiacomo et al., 2021).

Tra gli approcci maggiormente adottati in questo ambito c'è il Topic Modeling (Barravecchia et al., 2024). Il Topic Modeling è una tecnica di text mining che permette di individuare, all'interno dei dataset, temi latenti utili a valutare in modo accurato la qualità del servizio (Blei, 2012). È stato, infatti, dimostrato che attraverso il Topic Modeling guidato si riesce ad estrarre dal corpus testuale le cinque determinanti del SERVQUAL (Aspetti tangibili, Affidabilità, Capacità di risposta, Capacità di rassicurazione, Empatia) (Palese and Usai, 2018; Parasurman et al., 1988). Questo collegamento tra tecniche di data science e framework classici permette alle organizzazioni di mappare i feedback digitali direttamente sulle dimensioni chiave della soddisfazione del cliente (Korfiatis et al., 2019a).

Questi algoritmi di machine learning, essendo di tipo non supervisionato, non richiedono una preventiva etichettatura dei dati. Questo approccio data-driven garantisce una maggiore oggettività, in quanto permette ai temi di emergere spontaneamente dai dati, rivelando talvolta criticità o opportunità che l'azienda non aveva inizialmente previsto. I metodi di Topic Modeling sono in grado di analizzare automaticamente grandi volumi di documenti non strutturati, raggruppandoli in cluster semantici. La natura di tali temi viene poi definita attraverso le parole più rappresentative che l'algoritmo associa a ciascun topic (Abdelrazek et al., 2023; Barravecchia et al., 2021).

Nonostante l'ampia varietà di algoritmi di Topic Modeling esistenti, la maggior parte di essi segue un framework comune, sintetizzato negli step illustrati in Figura 2 (Hankar et al., 2025). Questo processo prevede l'estrazione del corpus testuale e il pre-processing del testo tramite tokenizzazione, lemmatizzazione e stemming. Segue poi la fase di rappresentazione del testo, un passaggio cruciale che trasforma i documenti in vettori matematici trattabili dagli algoritmi; tale operazione può avvenire tramite logiche di tipo bag-of-words o mediante rappresentazioni più moderne basate su embedding (Devlin et al., 2019). Infine, si giunge al cuore del processo: l'individuazione dei k topic, ai quali sono associate n parole, ognuna caratterizzata da una specifica probabilità di appartenenza al k-esimo tema, definendo così la struttura probabilistica dell'intero corpus (Blei et al., 2003).

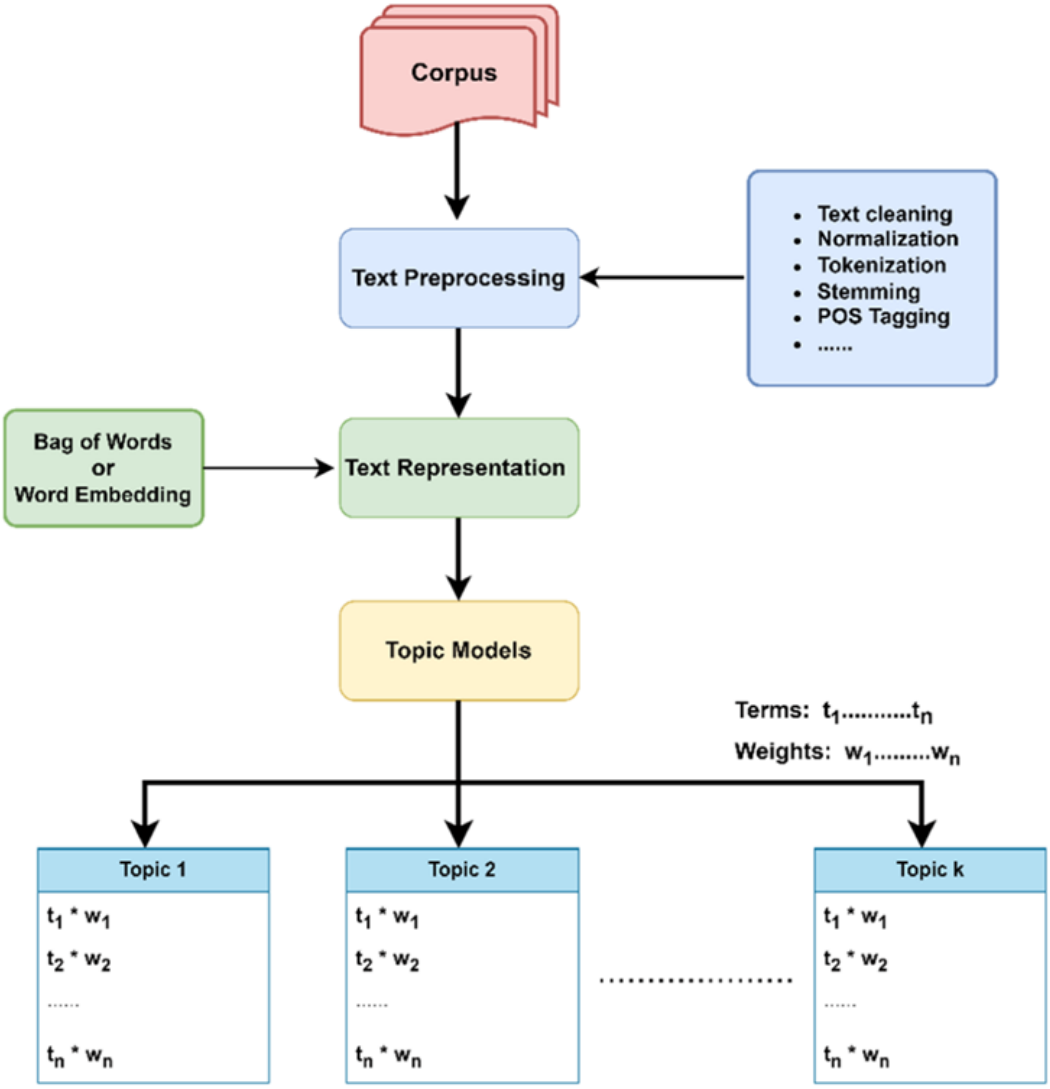

Figura 2 - Processo generale di Topic Modeling (Hankar et al., 2025)

Tra i modelli più utilizzati, come illustrato in Figura 3, figurano il Latent Dirichlet Allocation (LDA) e lo Structural Topic Model (STM). Questi due algoritmi rappresentano le fondamenta dell'approccio probabilistico e saranno discussi in questo capitolo. Successivamente, la trattazione si sposterà verso estensioni più avanzate: il Dynamic Topic Model (DTM), necessario per catturare l'evoluzione temporale dei temi, e BERTopic, che introduce il paradigma dei modelli neurali basati su Transformer.

È importante sottolineare, tuttavia, che il panorama del Topic Modeling è estremamente vasto e in costante evoluzione. I modelli citati in questa tesi rappresentano le soluzioni più consolidate in letteratura per l'analisi della Digital VoC ma esistono numerose altre varianti specializzate per contesti particolari o dataset con caratteristiche peculiari (Vayansky and Kumar, 2020).

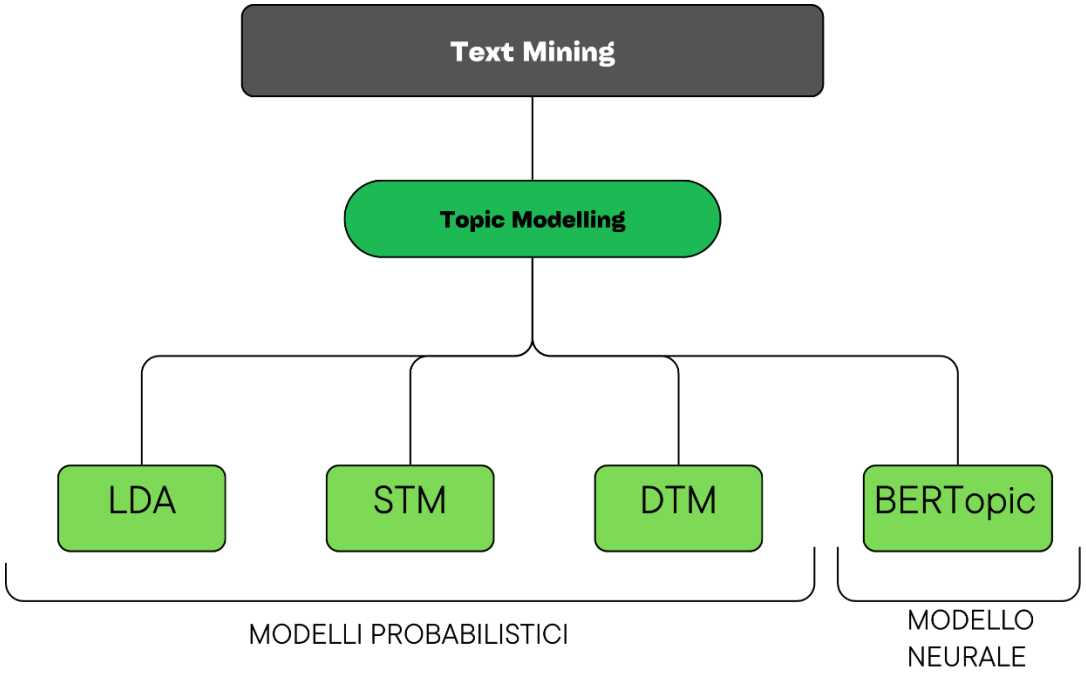

Figura 3 - Schema gerarchico delle relazioni tra text mining e varie tecniche di Topic Modeling

I modelli di Topic Modeling possono essere categorizzati in due famiglie metodologiche principali, che si differenziano sia per la modalità di rappresentazione del testo sia per il meccanismo attraverso cui i topic vengono individuati (Abdelrazek et al., 2023). È doveroso citare un'ulteriore famiglia di Topic Modeling, quella dei metodi non probabilistici, che si basano su tecniche algebriche o di fattorizzazione. In questa tesi non sono affrontati e per maggiori informazioni si rimanda alla letteratura sul tema (Hankar et al., 2025).

I modelli probabilistici come LDA, STM e DTM si basano su assunzioni generative classiche: ogni documento è concepito come una combinazione di topic latenti e ciascun topic come una distribuzione probabilistica di parole. Tale impostazione si appoggia solitamente sulla rappresentazione bag-of-words, un metodo che trasforma il testo in dati quantitativi ignorando l'ordine delle parole e la struttura sintattica per concentrarsi esclusivamente sulla frequenza di occorrenza dei termini all'interno del topic (Zhang et al., 2010). In questa famiglia di modelli, infatti, il documento viene ridotto a un 'sacchetto di parole' (bag-of-words): il significato non viene estratto dalla grammatica o dall'ordine delle frasi, ma semplicemente osservando quali parole tendono a comparire insieme con maggiore frequenza. L'assenza di contesto sequenziale viene compensata in fase di scelta delle parole chiave più rappresentative di ogni topic, dall'utilizzo di metriche come il TF-IDF (Term Frequency - Inverse Document Frequency) (Aizawa, 2003). Questo indicatore è fondamentale per l'interpretabilità dei risultati: permette di pesare l'importanza dei termini all'interno dei topic individuati, isolando i vocaboli più distintivi e rari rispetto a quelli troppo comuni e generici. Tale processo di pesatura facilita

l'etichettatura umana dei topic, garantendo che la solida base probabilistica del modello si traduca in temi coerenti e descrittivi (Hankar et al., 2025).

Successivamente, con l'introduzione delle rappresentazioni dense derivate dai Transformer, si è sviluppata una seconda famiglia di modelli, definibili come neurali o embedding-based. Gli embeddings possono essere descritti come traduzioni numeriche delle parole: ogni termine viene convertito in un vettore di numeri reali che ne cattura il significato grammaticale e semantico (Thapa et al., 2024). In questi modelli gli embeddings contestuali non si limitano a contare le parole, ma le proiettano in uno spazio vettoriale multidimensionale dove la vicinanza fisica corrisponde alla somiglianza semantica. A differenza dei modelli probabilistici, dove i topic emergono da co-occorrenze statistiche, qui la scoperta dei temi avviene tramite algoritmi di clustering applicati a questi vettori. Questo permette di cogliere sinonimie, polisemie e relazioni contestuali profonde che sfuggono ai modelli classici. Tuttavia, tali vantaggi introducono nuove criticità: la "scatola nera" dei Transformer rende il processo meno trasparente rispetto al rigore matematico di LDA, e i risultati diventano strettamente dipendenti dall'architettura e dal corpus di pre-addestramento del modello di embedding prescelto (Ajinaja et al., 2025). A queste problematiche si aggiunge una limitazione intrinseca legata all'interpretabilità dei risultati: mentre i modelli probabilistici restituiscono nativamente liste di parole chiave, i modelli basati esclusivamente su embedding producono cluster di vettori numerici che non sono immediatamente comprensibili per l'operatore umano (Opitz et al., 2025). Senza l'integrazione di tecniche di estrazione testuale a posteriori, tali modelli rischiano di fornire raggruppamenti semanticamente accurati ma difficili da etichettare e tradurre in insight manageriali concreti.

Recenti studi comparano le due famiglie di tecniche di Topic Modeling: i modelli embedding-based superano quelli probabilistici in termini di coerenza dei topic e interpretabilità. Mentre i modelli classici sono vincolati a rappresentazioni bag-of-words e richiedono la pre-definizione del numero di topic, gli approcci neurali estraggono temi latenti attraverso cluster vettoriali densi, adattando automaticamente il numero di topic alla struttura semantica del corpus (Ajinaja et al., 2025).

LDA, come detto, è un modello probabilistico generativo non supervisionato che consiste nell'identificare gli argomenti che descrivono il corpus testuale, nell'associare a ciascun argomento un insieme di parole (contenuto tematico) e, infine, nel definire una combinazione specifica di questi argomenti per ciascun documento (prevalenza tematica) (Barravecchia et al., 2023, 2021). Ogni documento è rappresentato come una combinazione di più topic, e ogni topic è descritto come una distribuzione di parole. L'algoritmo, attraverso un processo inferenziale, cerca di risalire ai topic più probabili che hanno generato i testi osservati, assegnando a ciascun documento una "miscela" di temi

e a ciascun tema un insieme di parole chiave rappresentative (Blei, 2012; Jelodar et al., 2019).

La semplicità di questo modello si porta dietro delle assunzioni fondamentali. La prima è dovuta al fatto che LDA lavori con rappresentazioni bag-of-words. Sebbene questa ipotesi semplifichi molto la struttura linguistica, è accettabile se l'obiettivo è solo individuare la struttura semantica generale del testo (Vayansky and Kumar, 2020). Un'altra assunzione è che l'ordine dei documenti nella collezione non abbia importanza. Anche questa è problematica per collezioni che si sviluppano nel tempo e saranno altri modelli ad affrontare questa limitazione, studiando l'evoluzione dei topic nel tempo. LDA assume anche che il numero di topic sia noto e fisso e che quindi sia un parametro da fissare in fase preliminare (Blei, 2012).

STM è un'evoluzione del modello LDA, che presenta un vantaggio significativo rispetto ai modelli di base: permette di includere informazioni aggiuntive, come le valutazioni dei clienti, la data, il luogo di pubblicazione, l'autore e il suo genere (Barravecchia et al., 2020; Roberts et al., 2014). Questo avviene attraverso l'inclusione di covariate di interesse sia nelle distribuzioni a priori delle proporzioni documento-topic sia in quelle topic-parola. A differenza dei modelli tradizionali, che assumono la prevalenza tematica (cioè la frequenza con cui un argomento appare) e il contenuto tematico (cioè le parole associate a un argomento) come invarianti tra i partecipanti, STM consente di integrare covariate che riflettono possibili variazioni di questi aspetti (Korfiatis et al., 2019b; Roberts et al., 2014). Questo rende STM particolarmente indicato per analizzare la VoC dal momento che non analizza solo il testo della recensione ma anche altri metadati quali titolo, autore, data e valutazione (Korfiatis et al., 2019b). In definitiva, STM si differenzia da LDA in quanto consente ai topic di essere correlati tra loro, permette a ciascun documento di avere una distribuzione a priori sulle proporzioni dei topic personalizzata in base alle covariate, e fa sì che l'uso delle parole all'interno di un topic possa variare in funzione delle covariate stesse.

In letteratura, l'applicazione dei modelli probabilistici di Topic Modeling a collezioni di dati non strutturati avviene generalmente attraverso una sequenza di passaggi standardizzati (Barravecchia et al., 2020). I principali step, sintetizzati in Figura 4, sono:

1. Estrazione del dataset e analisi preliminare dei dati: i dati testuali vengono raccolti tramite web scraping da diverse fonti online. Oltre ai contenuti delle recensioni, si possono estrarre metadati utili per le fasi successive (Mastrogiacomo et al., 2021).

2. Pre-processing testuale: il testo viene normalizzato, avviene cioè la conversione in minuscolo, la rimozione di punteggiatura, numeri, parole troppo frequenti o troppo rare, parole troppo

brevi o troppo lunghe, la rimozione delle stopwords nella lingua scelta. Il testo viene poi segmentato in token, e sottoposto a stemming per ricondurre le parole alla loro radice (es. “likes”, “liked” e “likely” vengono tutte ricondotte alla parola “like”). Questo processo di riduzione della dimensionalità è cruciale per concentrare il contenuto informativo del corpus, permettendo all'algoritmo di ignorare le variazioni morfologiche irrilevanti e focalizzarsi sul nucleo semantico dei termini. Successivamente, si eliminano le parole non relative al topic, come “ another”, “mean”, “review”, etc. Vengono, infine, sostituiti gli n-grammi più comuni (es. “customer service” → “customerservice”) per preservare il significato delle espressioni composte (Barravecchia et al., 2020; Korfiatis et al., 2019b; Mastrogiacomo et al., 2021).

3. Identificazione del numero ottimale di topic da estrarre: il parametro *K* (numero di topic) viene definito in modo iterativo, valutando l’held-out likelihood (metrica che quantifica la somiglianza del modello al sottoinsieme di UGC) o la coerenza semantica di ciascuna parola con il topic o l’esclusività delle parole associate al topic o i residui, ovvero la connessione tra numero di argomenti e adattamento del modello. Questi metodi aiutano a selezionare il numero di topic più rappresentativo del corpus (Barravecchia et al., 2020; Korfiatis et al., 2019b; Mastrogiacomo et al., 2021).

4. Labeling, ovvero assegnazione di etichette interpretative a ciascun topic sulla base delle parole chiave emerse: in questa fase, gli algoritmi probabilistici restituiscono, per ogni topic, una lista di parole chiave ottenute secondo diversi criteri statistici. Tra i più utilizzati figurano FREX (parole frequenti nel topic ed esclusive), utilizzato prevalentemente in STM, e Highest Probability (parole con probabilità più alta all’interno di ciascun topic), ma anche Lift (cioè un punteggio calcolato dividendo la distribuzione parola-topic per la distribuzione empirica della frequenza delle parole) e Score (cioè le parole con il punteggio più alto secondo il ranking LDA). La combinazione di questi diversi criteri può rendere l’etichettatura più efficace. Per una maggiore affidabilità, invece, è possibile supportare il labeling analizzando anche le recensioni più rappresentative (cioè quelle con il maggior peso per ciascun topic) e, se possibile, coinvolgere esperti per validare le etichette assegnate (Mastrogiacomo et al., 2021).

5. Validazione dei risultati: si confronta l’assegnazione automatica dei topic con un’assegnazione manuale effettuata da valutatori umani al fine di valutare in modo oggettivo l’efficacia dell’algoritmo. Per ogni recensione e topic, il confronto può portare a quattro possibili esiti: Vero Positivo (TP), algoritmo e valutatori concordano sull’assegnazione di un topic; Vero Negativo (TN), algoritmo e valutatori concordano nel non assegnare un topic; Falso Positivo

(FP), l'algoritmo assegna un topic che i valutatori non avevano previsto (errore di tipo I); Falso Negativo (FN), l'algoritmo non assegna un topic che era stato invece indicato dai valutatori (errore di tipo II). Sulla base di questi casi, è possibile calcolare alcuni indicatori di performance: l'*accuracy* che misura la percentuale di assegnazioni corrette sul totale, il *recall* che rappresenta la capacità dell'algoritmo di individuare i topic effettivamente presenti, il *precision rate* che indica la proporzione di assegnazioni corrette tra tutte quelle positive effettuate dall'algoritmo (Barravecchia et al., 2020) e l'*F-measure* che è la media armonica tra precision e recall, e fornisce una misura bilanciata che considera sia la correttezza che la completezza delle assegnazioni (Mastrogiacomo et al., 2021). Per una descrizione più approfondita delle modalità di calcolo di questi indicatori, si rimanda alla letteratura esistente sul tema. L'impiego di queste metriche permette di quantificare l'affidabilità del modello, garantendo che le conclusioni manageriali riflettano realmente la struttura informativa dei dati originali.

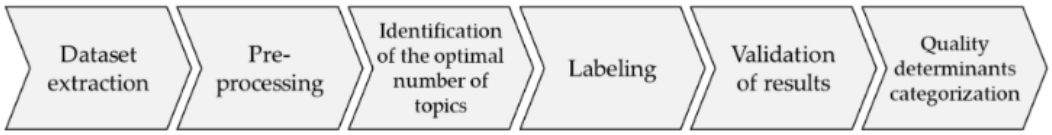

Figura 4 - Flusso di attività del STM (Barravecchia et al., 2020).

La fase algoritmica vera e propria prevede la costruzione del vocabolario di parole presenti nei documenti da analizzare, la rappresentazione dei testi come bag-of-words, l'inizializzazione delle distribuzioni di probabilità sia di topic-documento sia di topic-parole, che vengono poi affinate iterativamente finché non si raggiunge la convergenza (Barravecchia et al., 2023).

Nonostante l'efficacia dei modelli probabilistici statici come LDA e STM nell'estrazione di temi latenti dai dati testuali, essi presentano alcune limitazioni intrinseche che ne riducono l'applicabilità in contesti dinamici. In primo luogo, tali modelli assumono che la distribuzione dei topic e delle parole all'interno del corpus sia stazionaria, ovvero invariata nel tempo. Questa ipotesi risulta poco realistica nel caso di collezioni testuali che si evolvono, come le recensioni online, dove i termini dei temi emergenti possono variare rapidamente in funzione di eventi esterni, cambiamenti nel servizio o mutamenti nelle aspettative degli utenti (Korfiatis et al., 2019b). Inoltre, i modelli statici non tengono conto della sequenzialità temporale dei documenti, trattando ogni recensione come indipendente dalle altre, e non sono in grado di cogliere trend, ciclicità o transizioni (Murshed et al., 2023). Per rispondere a queste criticità e comprendere meglio l'evoluzione dei temi trattati nel tempo, è stato sviluppato un approccio alternativo, anch'esso probabilistico: il Dynamic Topic Modeling.

3.3. Dynamic Topic Modeling

Il Dynamic Topic Model (DTM), introdotto da Blei e Lafferty nel 2006, estende l'LDA (Blei et al., 2003; Blei and Lafferty, 2006). Mentre quest'ultimo modella le distribuzioni dei topic e delle parole in un corpus statico, il DTM estende questa capacità permettendo di analizzare come questi pattern tematici evolvono nel tempo. Specificamente, il DTM consente di modellare due aspetti fondamentali dell'evoluzione di un corpus documentale:

- l'evoluzione del contenuto tematico: ovvero come cambia la distribuzione delle parole all'interno di un singolo topic nel tempo;
- l'evoluzione della prevalenza tematica nel dataset: ovvero come cambia l'importanza di ciascun topic all'interno dell'intero corpus documentale nei diversi intervalli temporali, permettendo di identificare temi emergenti o in declino.

Per raggiungere questo scopo, il DTM opera una suddivisione del corpus in "fette" temporali discrete (ad esempio, anni, trimestri o mesi). In ciascuna fetta temporale t, i topic e le loro prevalenze tematiche non sono indipendenti, ma evolvono da quelli della fetta precedente $t-1$ (Blei and Lafferty, 2006).

Ogni topic k è descritto al tempo t da un vettore $\beta_{t,k}$ di parametri naturali, che rappresentano i pesi delle parole nel topic. Il vettore $\beta_{t,k} \in \mathbb{R}^V$ ha una lunghezza pari a V, ovvero il numero di termini nel vocabolario. Si assume che il vettore $\beta_{t,k}$ per il topic k a tempo t dipenda da quello al tempo $t-1$ ($\beta_{t-1,k}$) più una perturbazione casuale distribuita normalmente. La formula che descrive l'evoluzione dei pesi delle parole nel topic k nel tempo è l'equazione (1).

$$\beta_{t,k} \mid \beta_{t-1,k} \sim N\left(\beta_{t-1,k}, \sigma^2 I\right) \tag{1}$$

Dove $\sigma^2 I$ è la matrice di covarianza diagonale, con σ^2 varianza identica su ogni componente e nessuna correlazione tra le componenti. Questo implica che l'evoluzione del peso di una parola in un topic è indipendente dall'evoluzione di un'altra parola nello stesso topic.

Per l'intervallo t, si definisce un vettore $\alpha_t \in \mathbb{R}^K$ che rappresenta le tendenze generali dei *K* topic in quell'arco di tempo, influenzando la probabilità che un documento parli di un certo topic. La formula che descrive l'evoluzione nel tempo delle tendenze dei topic è l'equazione (2).

$$\alpha_t \mid \alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I) \tag{2}$$

Dove $\delta^2 I$ è la matrice di covarianza diagonale, con δ^2 varianza identica su ogni componente e nessuna

correlazione tra le componenti.

Mentre $\beta_{t,k}$ modella il contenuto del topic, il parametro α_t agisce come un regolatore della prevalenza del topic stesso nel tempo, determinando quanto spazio quel tema occupa nel dibattito collettivo in un dato periodo.

Per ogni periodo t, il processo generativo del DTM è il seguente:

1. Si estraggono i $\beta_{t,k}$ a partire da $\beta_{t-1,k}$, secondo (1);
2. Si estraggono gli α_t a partire da α_{t-1}, secondo (2);
3. Per ogni documento *d* nel periodo *t*:
    - Si campiona la variabile latente secondo (3)

(3) $$\eta \sim N(\alpha_t, a^2 I)$$

Dove $\eta \in \mathbb{R}^K$ rappresenta le propensioni tematiche di un documento;

- Per ogni parola $w_{t,d,n}$ del documento *d* al tempo *t*:
    a) Si estrae la variabile latente $z_{t,d,n}$, che indica quale topic tra i *K* possibili è stato selezionato per la parola $w_{t,d,n}$, campionandola da una distribuzione multinomiale con probabilità definite da η, come si vede in (4). η è trasformato tramite softmax esplicitata in (5).

(4) $$Z \sim Multinomial(\pi(\eta))$$

(5) $$\pi(\eta) = \theta_{t,d} = \frac{\exp(\eta_{d,k})}{\sum_{j=1}^{K} \exp(\eta_{d,j})}$$

Dove $\theta_{t,d}$ descrive il mix di topic per il documento d al tempo *t;*

b) Si estrae la parola $w_{t,d,n}$ da una distribuzione multinomiale con probabilità ottenute applicando la softmax (6) al vettore $\beta_{t,k}$, che rappresenta i parametri associati al topic *k* estratto per la parola *n* nel documento *d* al tempo *t*, secondo quanto descritto in (7).

(6) $$\pi(\beta_{t,k})_w = \frac{\exp(\beta_{t,k,w})}{\sum_w \exp(\beta_{t,k,w})}$$

(7) $$W_{t,d,n} \sim Multinomial\left(\pi(\beta_{t,z})\right).$$

Nella Figura 5 è rappresentato graficamente il processo generativo di questo modello. È interessante notare che, se le frecce orizzontali (che indicano la dipendenza temporale) non fossero presenti, il modello si ridurrebbe a un insieme di Topic Modeling indipendenti per ogni periodo temporale.

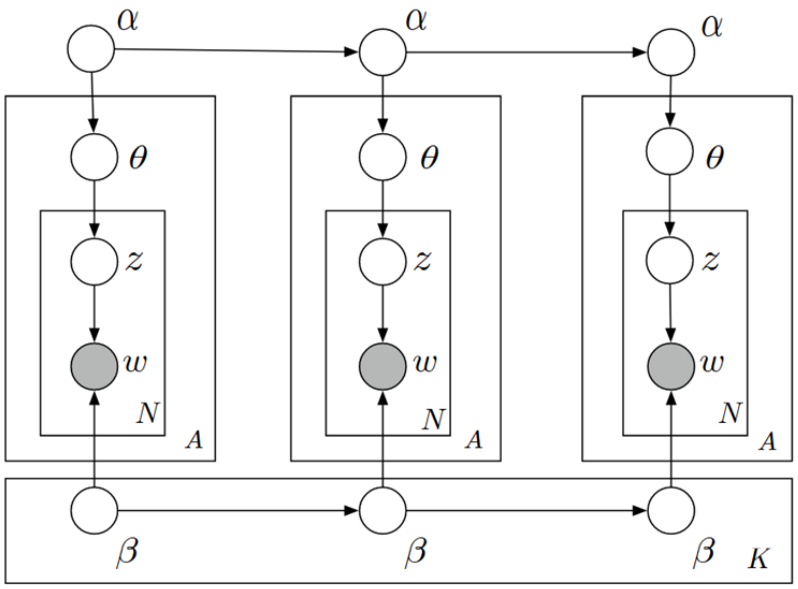

Figura 5 - Rappresentazione grafica di un DTM (per tre fette temporali). I parametri naturali βt,k e αt evolvono nel tempo (Blei and Lafferty, 2006).

Dunque, nel DTM i parametri dei topic ($\beta_{t,k}$ e α_t) cambiano nel tempo secondo una distribuzione Gaussiana, mentre le parole nei documenti ($w_{t,d,n}$) sono generate da una distribuzione Multinomiale. Queste due distribuzioni non sono "coniugate": in termini bayesiani, ciò significa che la distribuzione a posteriori dei parametri latenti (i topic) non ha una forma chiusa e analiticamente trattabile dopo aver osservato i dati (le parole nei documenti). Questo rende l'inferenza *intractable*, ovvero matematicamente troppo complicata da risolvere esattamente. Per superare questa difficoltà gli autori suggeriscono di usare l'inferenza variazionale (Blei et al., 2017). Questo approccio consiste nel costruire una distribuzione più semplice, detta distribuzione variazionale q, che approssimi al meglio la vera distribuzione a posteriori. La distribuzione q ha dei "parametri liberi" che vengono ottimizzati per minimizzare la divergenza tra q e la vera distribuzione a posteriori (Blei and Lafferty, 2006).

Le variabili latenti, quelle che si vogliono inferire nel DTM, sono: i parametri dei topic, che descrivono

come cambiano le distribuzioni delle parole di ciascun topic *k* nel tempo *t* $\{\beta_{t,k}\}$; le proporzioni di topic nel documento *d* al tempo *t*, che indicano quanto un documento parla di ciascun topic $\{\theta_{t,d}\}$; gli indicatori di topic per ogni parola nel documento, che specificano da quale topic si ritiene che una parola nel documento *d* al tempo *t* sia stata generata $\{z_{t,d,n}\}$. Per approssimare la distribuzione a posteriori *p* di queste variabili latenti, si costruisce una distribuzione variazionale *q* fattorizzata in (8).

$$(8) \quad q(\beta, \theta, z) = \prod_{k=1}^{K} q\left(\beta_{k,1}, \dots, \beta_{k,T} \mid \widehat{\beta_{k,1}}, \dots, \widehat{\beta_{k,T}}\right) \times \prod_{t=1}^{T} \left(\prod_{d=1}^{D_t} q(\theta_{t,d} \mid \gamma_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n} \mid \varphi_{t,d,n})\right)$$

Nella fattorizzazione di *q* viene applicata l'approssimazione mean-field: si trattano indipendentemente i gruppi di variabili latenti (*β, θ, z*), tranne che per la sequenza $\{\beta_{1:T,k}\}$ che mantiene la dipendenza temporale.

I parametri variazionali liberi che approssimano le variabili latenti sono: $\widehat{\beta_{k,t}}$, $\gamma_{t,d}$ e $\varphi_{t,d,n}$.

$\widehat{\beta_{k,t}}$ è il parametro variazionale che approssima $\beta_{k,t}$. Come precedentemente visto, $\beta_{k,t}$ si distribuisce secondo una Normale Multivariata. L'evoluzione temporale di questi parametri può essere stimata tramite il Filtro di Kalman variazionale, che sfrutta le proprietà di linearità e gaussianità per un forward–backward smoothing continuo, oppure la Regressione Wavelet variazionale, che cattura sia tendenze lente sia picchi rapidi usando basi wavelet. Per maggiori informazioni tecniche su questi due metodi si rimanda al paper di Blei e Lafferty (Blei and Lafferty, 2006).

$\gamma_{t,d}$ è il parametro variazionale libero della distribuzione Dirichlet su $\theta_{t,d}$ in (9).

$$(9) \quad q(\theta_{t,d}) = Dirichlet(\gamma_{t,d})$$

Infine, $\varphi_{t,d,n}$ è il parametro variazionale libero della distribuzione Multinomiale su $z_{t,d,n}$ in (10).

$$(10) \quad q(z_{t,d,n}) = Multinomial(\varphi_{t,d,n})$$

I parametri variazionali $\{\widehat{\beta_{k,t}}, \gamma_{t,d}, \varphi_{t,d,n}\}$ vengono ottimizzati massimizzando l'Evidence Lower BOund (ELBO), una funzione che fornisce un limite inferiore alla verosimiglianza marginale dei dati e misura quanto la distribuzione variazionale *q* sia vicina alla vera distribuzione a posteriori. Il problema di ottimizzazione si risolve alternando i seguenti passaggi:

- Aggiornamenti "in forma chiusa" (cioè espressioni analitiche) per $\gamma_{t,d}$ e $\varphi_{t,d,n}$: proprio come in LDA, le formule di aggiornamento per $\gamma_{t,d}$ e $\varphi_{t,d,n}$ si ottengono derivando

l'ELBO rispetto a questi parametri e imponendo che il gradiente sia zero. Questo permette di trovare le soluzioni ottimali direttamente.

- Ottimizzazione numerica per $\widehat{\beta_{k,t}}$ (tramite conjugate gradient o simili): ad ogni iterazione, si calcola il gradiente dell'ELBO rispetto a $\widehat{\beta_{k,t}}$ e si aggiorna nella direzione che massimizza l'ELBO, ripetendo il processo finché l'ELBO converge a un massimo locale.

La prima applicazione del DTM avviene a cura di coloro che hanno sviluppato il modello, Blei e Lafferty. Selezionando 30.000 articoli pubblicati in 118 anni sulla rivista *Science,* hanno ottenuto un corpus composto da circa 16.000 parole, a seguito di una fase di pre-processing (Blei and Lafferty, 2006). Per esplorare il corpus e i suoi temi, hanno stimato un modello di topic dinamico a 20 componenti. Attraverso l'inferenza a posteriori, sono riusciti ad analizzare l'andamento dei topic nel tempo. Per validare quantitativamente il modello, gli autori hanno condotto un'analisi predittiva, confrontando la capacità del DTM di prevedere gli articoli dell'anno successivo di *Science* rispetto a due modelli di topic statici: uno addestrato su tutti gli anni precedenti e uno addestrato solo sull'anno precedente. Il DTM ha dimostrato una performance superiore, assegnando costantemente una maggiore probabilità agli articoli dell'anno seguente rispetto ai modelli statici (Blei and Lafferty, 2006).

In definitiva, il Dynamic Topic Model di Blei e Lafferty rappresenta il modello tradizionale e il riferimento fondamentale per l'analisi dell'evoluzione dei topic nel tempo. La sua architettura matematica, basata sull'inferenza variazionale con Filtro di Kalman o Regressione Wavelet, sebbene robusta a livello concettuale, rende la sua implementazione diretta particolarmente complessa e computazionalmente onerosa.

3.4. BERTopic

Nell'ambito dei modelli neurali precedentemente definiti, BERTopic si distingue come uno dei framework più avanzati e versatili, grazie alla sua architettura modulare che ottimizza l'uso delle reti neurali per l'estrazione tematica (Grootendorst, 2022).

A differenza di altri modelli embedding-based che si limitano al clustering di vettori statici, dove una parola ha un unico valore numerico indipendentemente dal contesto, la natura neurale di BERTopic risiede nell'integrazione di Transformer pre-addestrati. Il capostipite di questa tecnologia è BERT (Devlin et al., 2019), il quale permette di generare rappresentazioni bidirezionali profonde. Su questa architettura si basa Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Mentre i primi modelli

di embedding rappresentano un anello di congiunzione capace di riconoscere la somiglianza tra parole superando il limite del BoW, essi soffrono ancora dell'incapacità di distinguere i diversi significati di un termine in base al contesto. L'architettura neurale di BERTopic supera anche questa difficoltà e genera rappresentazioni numeriche del testo che non sono solo dense, ma contestualizzate: la stessa parola assume coordinate vettoriali differenti a seconda del senso che acquisisce nella frase (Tripodi, 2021). Questo permette al modello di catturare la semantica profonda della Digital VoC, risolvendo con precisione i problemi di polisemia e sinonimia che affliggono i modelli probabilistici e gli approcci vettoriali più rudimentali.

BERTopic, inoltre, supera un altro svantaggio tipico dei modelli embedding-based, adottando un approccio ibrido. Sebbene il cuore del modello sia neurale, la descrizione dei topic non rimane confinata nello spazio vettoriale, che spesso è di difficile lettura, come già detto, ma viene tradotta in linguaggio naturale attraverso il c-TF-IDF (Grootendorst, 2022). Questa tecnica permette di estrarre le parole chiave più distintive per ogni cluster, garantendo che l'elevata capacità di astrazione neurale si traduca in risultati concreti, leggibili e coerenti con le aspettative umane dell'analisi della qualità del servizio o del prodotto (Bianchi et al., 2021).

I quattro componenti che costituiscono l'algoritmo BERTopic (Grootendorst, 2022) sono i seguenti, come illustrato in Figura 6:

1. Embedding dei documenti: BERTopic trasforma i documenti di testo pre-processati in rappresentazioni numeriche, cioè in embedding vettoriali, che consentono di confrontare semanticamente i testi nello spazio vettoriale. Questi embedding vengono poi utilizzati principalmente per raggruppare documenti semanticamente simili tramite tecniche di clustering. Si ricorda che BERTopic utilizza il framework Sentence Transformers, che implementa l'approccio Sentence-BERT (SBERT), che converte le frasi o i paragrafi in rappresentazioni vettoriali, utilizzando modelli linguistici pre-addestrati (Reimers and Gurevych, 2019);

2. Riduzione della Dimensionalità: i vettori di embedding generati sono tipicamente ad alta dimensionalità. Questo rappresenta un problema in quanto in spazi con molte dimensioni le distanze tra i punti tendono a diventare simili tra loro, rendendo difficile distinguere tra dati realmente vicini e dati lontani (Verleysen and François, 2005). Di conseguenza, le tecniche di clustering perdono efficacia e affidabilità. Per ovviare a questo problema, è necessario ridurre la dimensionalità degli embedding, mantenendo però le caratteristiche più importanti dei dati. Per rendere il successivo passaggio di clustering più efficiente e maneggevole, quindi,

BERTopic utilizza UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2020). UMAP riduce questi vettori ad alta dimensionalità, proiettandoli in uno spazio con un numero di dimensioni inferiore, mantenendo le relazioni significative tra i documenti. Rispetto a tecniche più datate come la PCA (Principal Component Analysis), che si limitano a proiezioni lineari, UMAP è in grado di preservare sia la struttura locale che quella globale dei dati. Questo significa che i documenti molto simili rimangono vicini, ma viene mantenuta anche la separazione tra gruppi tematici molto diversi, garantendo una mappa dello spazio vettoriale più fedele alla complessità del linguaggio umano;

3. Clustering: una volta che gli embedding sono stati ridotti dimensionalmente, BERTopic applica HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) per raggruppare i documenti simili in cluster, che diventeranno i topic. HDBSCAN è un metodo di clustering gerarchico basato sulla densità che modella i cluster attraverso un approccio *soft-clustering*, che consente di trattare il rumore come outlier (Campello et al., 2015). Questo impedisce l'assegnazione di documenti non correlati a cluster forzati e, dunque, migliora la qualità della rappresentazione degli argomenti rispetto ad altri modelli, come per esempio LDA, nel quale ogni documento deve appartenere per forza a un mix di topic. La capacità di identificare il rumore è particolarmente preziosa nell'analisi delle recensioni online, dove spesso sono presenti testi irrilevanti, spam o commenti privi di contenuto informativo. Escludendo questi outlier dal processo di formazione dei topic, HDBSCAN garantisce che i nuclei tematici finali siano 'puliti' e composti solo da feedback realmente significativi, migliorando molto la precisione del clustering;

4. Etichettatura dei cluster: dopo che ogni documento viene assegnato a un cluster, il passaggio finale consiste nell'associare ad ogni cluster un topic. Per fare ciò, BERTopic utilizza una variante del TF-IDF, chiamata c-TF-IDF (class-based TF-IDF), per estrarre le parole più rappresentative di ciascun topic. Il TF-IDF classico viene impiegato per identificare le parole significative all'interno di un singolo documento; il c-TF-IDF, invece, individua le parole più rilevanti a livello di cluster, ovvero per ciascun topic (Paltoglou and Thelwall, 2010). Per applicare questa tecnica, è necessario concatenare tutti i documenti appartenenti a un cluster, trattandoli come un unico documento. L'obiettivo è sia identificare la rilevanza locale, ovvero premiare le parole che compaiono con alta frequenza nel cluster specifico $tf_{n,k}$, sia identificare la distintività globale, penalizzando le parole che compaiono frequentemente in tutti gli altri cluster (tramite il logaritmo del rapporto tra il volume totale delle parole e la frequenza globale del termine tf_n), assicurando che le parole scelte siano esclusive di quel topic e non termini

generici del corpus. In pratica, il c-TF-IDF agisce come un filtro: tiene solo le parole che descrivono bene quel gruppo specifico e scarta quelle che si trovano ovunque nel dataset. Questo è il passaggio che permette di passare dai calcoli matematici a parole vere e proprie che permettono di generare insight.

Questo passaggio consente di generare una distribuzione parole-topic, ovvero un elenco di parole chiave ordinate in base alla loro rilevanza per ciascun argomento. La formula per determinare i punteggi per ogni cluster è quella presentata in (11).

(11) $$W_{n,k} = tf_{n,k} \times \log\left(1 + \frac{\mathrm{A}}{tf_n}\right)$$

Dove $W_{n,k}$ rappresenta il peso della parola *n* all'interno del cluster *k*, ovvero quanto è rappresentativa la parola *n* per quel topic. $tf_{n,k}$ è la frequenza della parola *n* all'interno del cluster *k*. A è la media del numero totale di parole per ciascun cluster: questo parametro funge da fattore di normalizzazione. Senza di esso, i cluster più voluminosi avrebbero punteggi naturalmente più alti rispetto a quelli piccoli, distorcendo il confronto. A riporta virtualmente tutti i cluster sullo stesso piano, permettendo di confrontare correttamente l'importanza delle parole indipendentemente dalla dimensione del gruppo. Infine, tf_n è la frequenza totale della parola n in tutti i cluster.

Nel caso in cui il numero di topic generati sia troppo elevato, BERTopic prevede un meccanismo per ridurli automaticamente: i topic meno rappresentati vengono progressivamente fusi con quelli più simili, basandosi sulle loro rappresentazioni c-TF-IDF, fino a raggiungere il numero di topic desiderato dall'utente (Grootendorst, 2022).

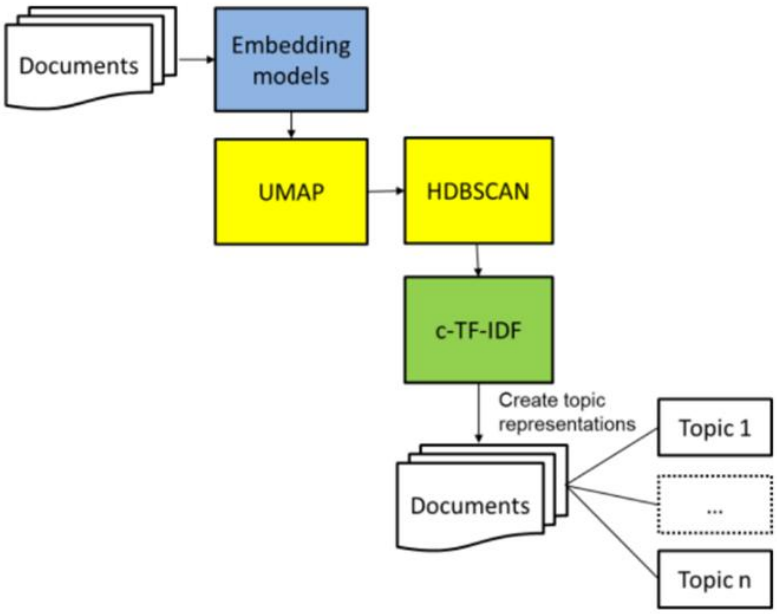

Figura 6 - Componenti del BERTopic (Hananto, 2023).

La modularità di questo framework risiede nella possibilità di sostituire o configurare indipendentemente ciascuno di questi quattro moduli, permettendo di adattare l'algoritmo alle specificità del linguaggio naturale e del corpus testuale analizzato (Grootendorst, 2022).

BERTopic consente di analizzare l'evoluzione dei temi nel tempo. Per farlo, si procede inizialmente addestrando BERTopic sull'intero corpus ignorando la componente temporale dei dati, così da generare una rappresentazione globale dei topic (Wang and McCallum, 2006). Una volta ottenuta la rappresentazione globale, BERTopic calcola la rappresentazione locale di ciascun topic in ogni timestamp moltiplicando la frequenza del termine nei documenti del tempo *t* per i valori IDF globali precedentemente calcolati, come si vede in (12).

(12) $$W_{n,k,t} = tf_{n,k,t} \times \log\left(1 + \frac{\mathrm{A}}{tf_n}\right)$$

In questa configurazione, mentre l'IDF rimane globale per garantire una base di confronto comune, la frequenza del termine tf viene calcolata solo sui documenti appartenenti al tempo t. Questa tecnica permette di osservare come il lessico di un medesimo topic evolva, catturando ad esempio l'emergere di nuovi termini specifici in un determinato anno.

Un vantaggio di questo approccio è che non è necessario ripetere la fase di embedding e clustering: una volta costruiti i cluster iniziali, è possibile aggiornare dinamicamente la rappresentazione dei topic nei diversi intervalli temporali. Inoltre, questa tecnica può essere estesa anche ad altri metadati, come ad esempio l'autore o la fonte dei documenti (Grootendorst, 2022).

Infine, per ottenere una rappresentazione dei topic più stabile nel tempo, è possibile applicare una tecnica di smoothing che combina la rappresentazione del topic all'istante *t* con quella all'istante *t-1*, attenuando variazioni brusche e favorendo una transizione più lineare dei temi nel tempo (Grootendorst, 2022).

Sebbene il framework di BERTopic integri una funzionalità nativa per il Topic Modeling dinamico, ovvero *topics over time*, la sua natura modulare consente di estrarre topic statici di alta qualità per poi analizzarli in relazione a metadati esterni tramite procedure personalizzate.

Nel presente lavoro, si è scelto di sfruttare la robustezza del clustering semantico di BERTopic per identificare i nuclei tematici fondamentali, procedendo successivamente a una caratterizzazione temporale *ex-post*. Tale scelta risponde alla necessità di garantire la massima coerenza qualitativa dei risultati. Definendo i topic sull'intero corpus prima di osservarne la dinamica, è stato possibile effettuare un'operazione di validazione e fusione (merge) dei temi ridondanti. Senza questo passaggio preventivo, si correrebbe il rischio di analizzare l'evoluzione di cluster frammentati o troppo simili tra loro, rendendo difficile l'interpretazione dei trend. In questo modo, invece, la dimensione temporale viene applicata solo a nuclei tematici già verificati e consolidati, assicurando che ogni variazione osservata nel tempo rifletta un reale cambiamento nel sentiment dei consumatori e non un'incertezza statistica dell'algoritmo.

In sintesi, BERTopic si distingue per la sua flessibilità e robustezza (Krishnan, 2023), grazie alla capacità di sfruttare modelli linguistici avanzati e di separare il processo di embedding dei documenti dalla generazione delle rappresentazioni dei topic. Questo permette di adattarsi facilmente a diversi contesti e di ottenere risultati competitivi anche con risorse limitate. A conferma della qualità dei topic generati, BERTopic consente la valutazione mediante metriche come il *coherence score*, che misura la coerenza semantica tra le parole chiave di ciascun argomento, mostrando in diversi contesti prestazioni migliori rispetto ad approcci più tradizionali come LDA (Grootendorst, 2022).

Tuttavia, il modello presenta alcune semplificazioni: in primo luogo, assume tipicamente che ogni documento sia associato a un singolo topic prevalente, il che può risultare limitante per testi estremamente complessi o ambivalenti (Blei, 2012; Blei et al., 2003). Inoltre, sebbene la fase di scoperta dei temi sia neurale, la loro descrizione finale si basa su una pesatura di termini c-TF-IDF che, pur essendo efficace, eredita alcune caratteristiche della logica bag-of-words, come il rischio di ridondanze tra parole chiave simili (Chang et al., 2009). Nonostante queste limitazioni, BERTopic rappresenta un significativo passo avanti nel campo del Topic Modeling, combinando l'efficacia dei Transformer con la praticità d'uso necessaria per l'analisi di grandi volumi di dati come la Digital VoC.

3.4.1. Applicazioni di BERTopic

Recentemente BERTopic è stato molto utilizzato grazie alla sua capacità di gestire con estrema precisione grandi volumi di dati non strutturati, tipici delle piattaforme digitali e non solo.

Un esempio è il paper intitolato "AI Techniques and Applications for Online Social Networks and Media: Insights From BERTopic Modeling" nel quale BERTopic è stato applicato per analizzare un ampio corpus di pubblicazioni accademiche relative all'uso dell'intelligenza artificiale nei social network e nei media online, con l'obiettivo di individuare i principali temi di ricerca e le relative tendenze (Nedungadi et al., 2025). I topic emersi comprendono temi come la rilevazione di fake news, l'analisi del sentiment, la rilevazione di discorsi di odio, l'analisi di big data, il rilevamento di bot, la sorveglianza della salute pubblica e il monitoraggio della salute mentale tramite social media. Ogni topic è stato descritto attraverso le 20 parole chiave più rappresentative e una selezione di documenti centrali. Lo studio ha affrontato quattro domande principali sull'uso dell'AI nei social network: come proteggere gli utenti, come analizzare la diffusione delle informazioni, come gestire grandi volumi di dati in tempo reale e come rilevare contenuti dannosi. Le risposte evidenziano i benefici dell'AI nella personalizzazione e nel monitoraggio, ma anche i rischi legati a bias, privacy e disinformazione. Tra le soluzioni proposte vi sono tecniche di apprendimento equo, sistemi di moderazione automatica e metodi scalabili per l'analisi dei dati. L'impiego di BERTopic in un contesto così formale e specialistico ne dimostra la grande flessibilità: il modello, infatti, non è utile solo per analizzare testi brevi come le recensioni, ma si rivela efficace anche per mappare la letteratura scientifica. Questa capacità di gestire linguaggi e dataset così diversi conferma la validità di BERTopic come strumento ideale per ottenere una mappatura precisa e profonda dei temi.

Un altro studio applica l'algoritmo BERTopic nell'ambito delle recensioni nel settore dell'abbigliamento. Viene utilizzato il modello RoBERTa (una versione ottimizzata e più robusta del modello BERT originale, addestrata su un volume maggiore di dati per migliorare la comprensione del linguaggio naturale) per la classificazione delle recensioni secondo cinque dimensioni qualitative predefinite (Materiali, Costruzione, Colore, Finitura, Durabilità). BERTopic è stato impiegato per individuare automaticamente temi ricorrenti nelle recensioni degli utenti, che gli autori non avevano precedentemente predetto. L'analisi ha coinvolto un corpus composto da oltre 51.000 recensioni di prodotti di abbigliamento raccolte da Amazon e Kaggle, sottoposte a un processo di pre-processing linguistico (tokenizzazione, rimozione di stopword, lemmatizzazione). I risultati includono l'identificazione di dieci temi centrali, quali la morbidezza e il comfort nei maglioni, la fragilità delle cerniere, la sottigliezza dei tessuti economici e problematiche legate alla vestibilità o alla durabilità dei

capi. Inoltre, la matrice di similarità tra i topic ha permesso di evidenziare correlazioni utili alla progettazione di soluzioni sostenibili condivise tra diversi argomenti. Per esempio, il Topic 0 e il Topic 8 mostrano una forte somiglianza nella matrice di similarità, in Figura 7; per questo, le aziende potrebbero usare gli stessi coloranti naturali per entrambe le categorie di abbigliamento. Le intuizioni derivate da BERTopic si sono rivelate particolarmente preziose per le aziende, offrendo strumenti per comprendere meglio le esigenze dei consumatori, migliorare la progettazione di capi più duraturi e ridurre lo spreco e l'impatto ambientale legato agli scarti tessili. Nel complesso, l'approccio proposto dimostra la validità di BERTopic nel trasformare semplici recensioni in decisioni aziendali concrete (come la scelta di materiali o coloranti), confermando il valore del modello come ponte tra il linguaggio spontaneo dei consumatori e l'ottimizzazione dei processi industriali (Alamsyah and Girawan, 2023).

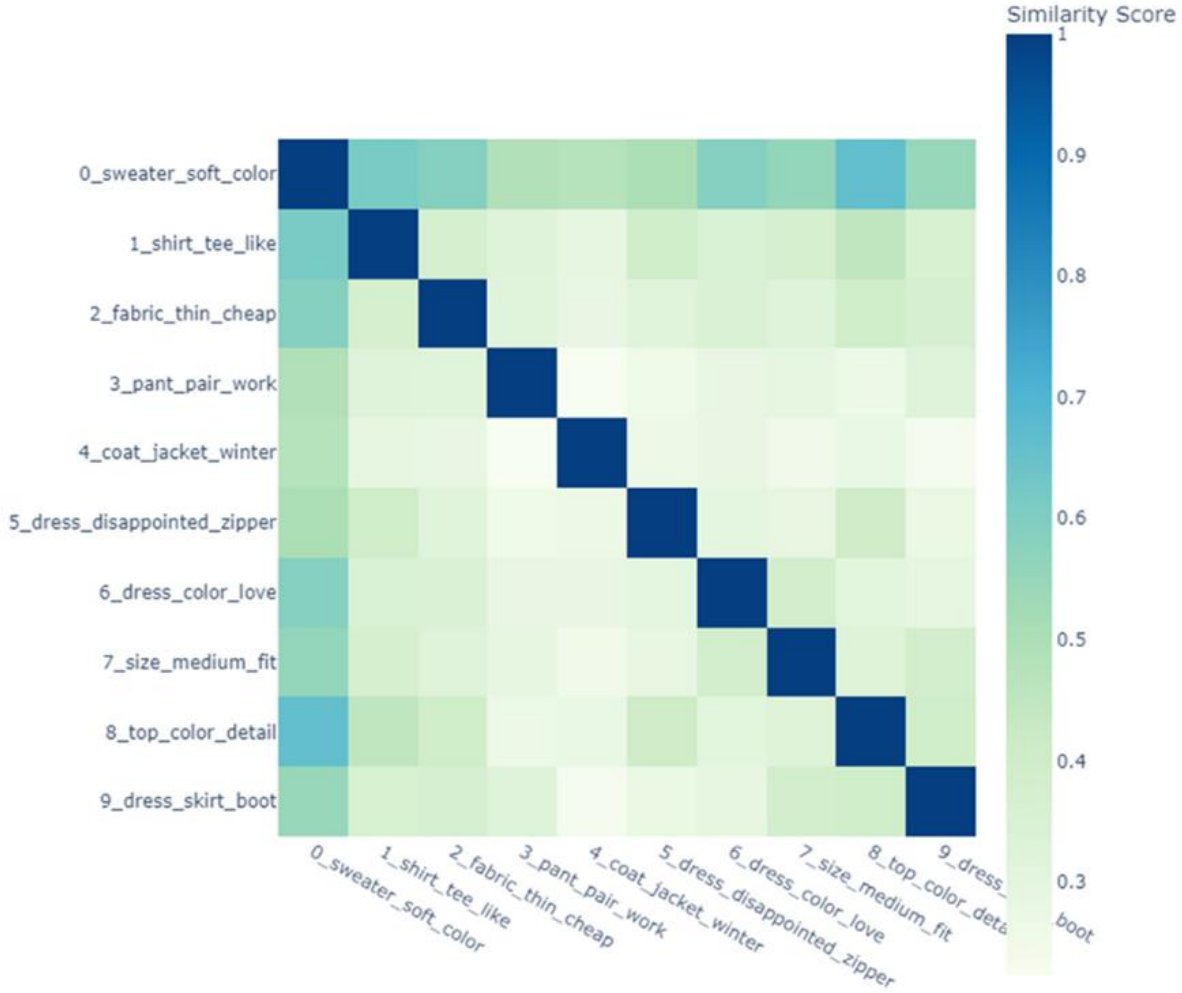

Figura 7 - Matrice di Similarità (Alamsyah and Girawan, 2023).

Un'altra interessante applicazione di BERTopic è stata condotta da Yi, Oh e Kim, con l'obiettivo di identificare i fattori che influenzano la soddisfazione degli investitori retail nei confronti dei servizi di mobile trading. Analizzando oltre 270.000 recensioni di utenti su quattro app leader del settore, gli

autori hanno utilizzato BERTopic per estrarre 33 topic distinti, poi interpretati e classificati con l'ausilio di un Large Language Model (ChatGPT-4). I topic sono stati mappati su cinque dimensioni di qualità del servizio ispirate al SERVQUAL: funzionalità, usabilità, qualità dell'informazione, servizio clienti e qualità del sistema. I risultati evidenziano che la soddisfazione degli utenti è fortemente influenzata da feedback positivi sull'usabilità, sull'informazione disponibile e sul supporto clienti, mentre le principali fonti di insoddisfazione riguardano problemi di servizio, instabilità del sistema e difficoltà funzionali. Le implicazioni manageriali sono significative: i broker tradizionali dovrebbero investire nel potenziamento del servizio personalizzato, mentre quelli online devono focalizzarsi sul miglioramento della stabilità tecnica e sull'ottimizzazione dell'esperienza utente, senza trascurare la qualità del supporto, fattore critico in caso di insoddisfazione. L'aspetto innovativo di questo lavoro risiede nell'integrazione tra la capacità di clustering di BERTopic e la potenza interpretativa dell'LLM: mentre il primo organizza i dati in gruppi coerenti, il secondo permette di superare la semplice lista di parole chiave, fornendo descrizioni dei topic più ricche e sfumate. Tale approccio riduce drasticamente il tempo necessario per la revisione manuale dei risultati e aumenta l'oggettività dell'analisi, confermando come l'evoluzione del Topic Modeling verso sistemi ibridi rappresenti oggi la soluzione ottimale per estrarre valore strategico dalla Digital VoC. Questo studio dimostra il potenziale dell'integrazione tra BERTopic e LLM per trasformare grandi volumi di recensioni in insight strategici a supporto del miglioramento continuo dei servizi digitali (Yi et al., 2025).

4. Metodologia di ricerca

4.1. Descrizione generale del processo di analisi

In questa tesi viene sviluppata una metodologia di analisi testuale basata sull'algoritmo BERTopic (Grootendorst, 2022), finalizzata all'identificazione dei topic e all'analisi della loro dinamica temporale all'interno di un corpus testuale strutturato temporalmente. L'obiettivo principale non è l'interpretazione dei risultati specifici del dataset analizzato, bensì la definizione di una procedura metodologica riproducibile e generalizzabile, in grado di individuare tendenze emergenti, variazioni improvvise e fenomeni di stagionalità in diversi contesti applicativi.

Nel presente lavoro, i topic vengono stimati in modo statico sull'intero insieme di recensioni, tramite l'applicazione di BERTopic; successivamente, l'analisi viene estesa alla dimensione temporale attraverso una procedura *ex-post*, che consente di osservare come il contenuto lessicale dei topic individuati si distribuiscano ed evolvano nel tempo. Questo approccio permette di analizzare variazioni temporali, tendenze emergenti e fenomeni di stagionalità, senza ricorrere a un modello di Topic Modeling dinamico propriamente detto.

Il processo di analisi sviluppato in questa tesi si articola in un modello strutturato e riproducibile. In una prima fase viene configurato l'ambiente di lavoro in R, integrando un ambiente virtuale Python per consentire l'esecuzione dell'algoritmo BERTopic. Questa scelta tecnica permette di unire la potenza di calcolo dei modelli Transformer di Python con la flessibilità di manipolazione dati e visualizzazione tipica di R. Successivamente, una volta selezionato il dataset, viene eseguita una fase di pre-processing dei testi che include il filtraggio per lingua, la rimozione delle recensioni non informative, la pulizia del rumore linguistico (stopword, punteggiatura e caratteri speciali) e la normalizzazione delle forme lessicali.

Completata la fase preliminare, viene addestrato il modello di Topic Modeling statico, indipendente dalla dimensione temporale, dal quale si ottengono le rappresentazioni globali dei topic e l'assegnazione dei documenti ai cluster tematici. I topic individuati sono quindi sottoposti a una fase di raffinamento e labeling, che combina informazioni quantitative fornite dal modello e ispezione qualitativa dei contenuti.

In una fase successiva, l'analisi viene estesa alla dimensione temporale attraverso una procedura *ex-post*: i documenti vengono aggregati in base ai rispettivi timestamp e, per ciascun intervallo temporale, vengono calcolate rappresentazioni specifiche dei topic tramite il metodo class-based TF-IDF. Questo

consente di osservare come la rilevanza semantica delle parole associate a ciascun topic vari nel tempo.

L'intero workflow, dalla preparazione dei dati alla costruzione del modello statico, dalla fase di labeling alla derivazione delle rappresentazioni dinamiche, definisce una procedura strutturata e replicabile per l'analisi dell'evoluzione temporale dei temi all'interno di un corpus testuale.

Le diverse fasi dell'indagine sono riassunte nel flowchart riportato in Figura 8 e verranno discusse con maggiore dettaglio nei prossimi capitoli.

Il codice completo sviluppato in ambiente R è consultabile in Appendice A.

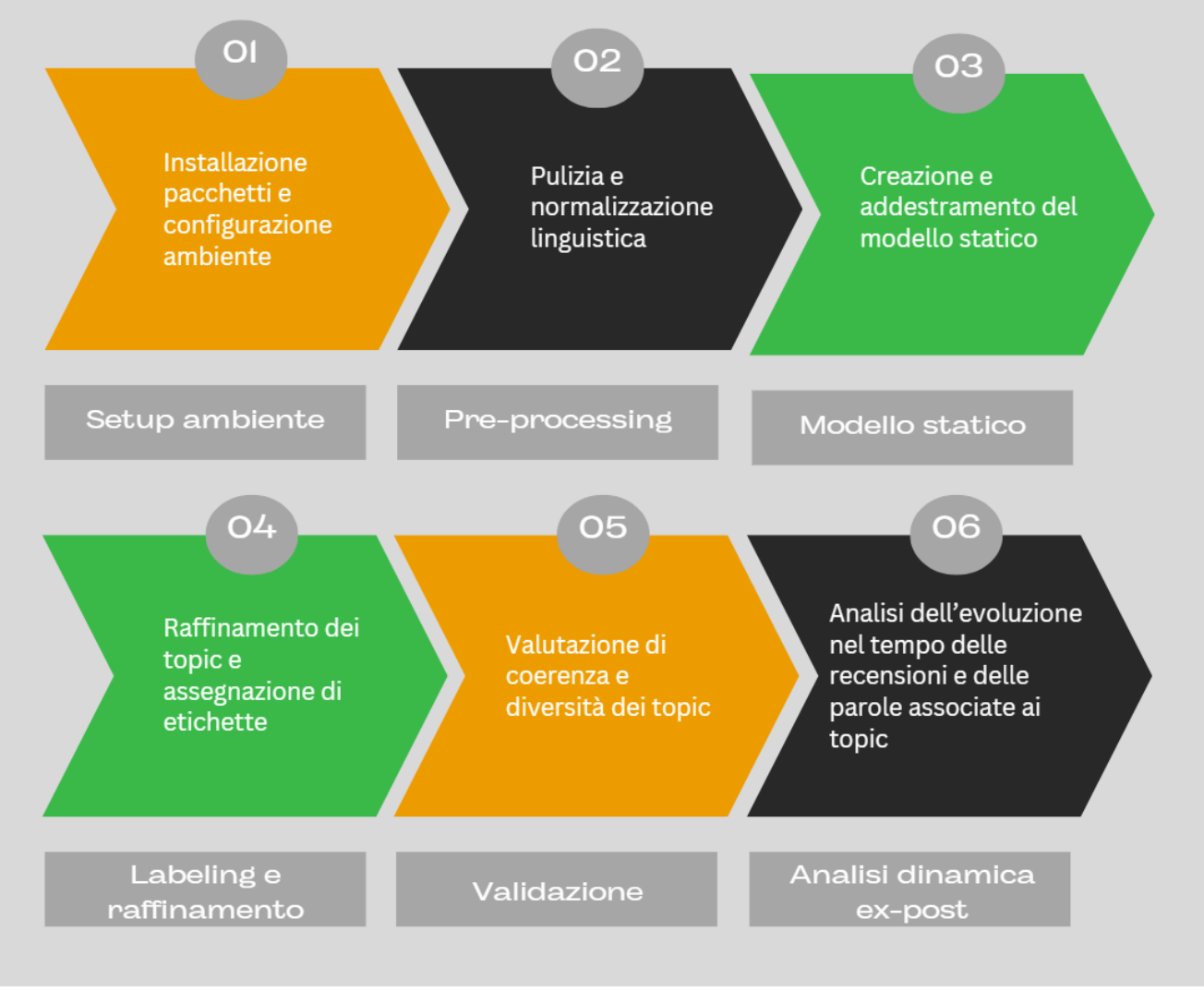

Figura 8 - Flowchart della metodologia di ricerca

4.2. Setup ambiente, raccolta e pre-processing dei dati

BERTopic è stato sviluppato originariamente per l'ambiente Python, che ne rappresenta il contesto di utilizzo più diffuso . L'architettura metodologica di questa tesi, tuttavia, prevede l'utilizzo di R come ambiente principale di orchestrazione. Questa scelta è motivata dalla superiore flessibilità di R nelle operazioni di manipolazione dei dati, nell'analisi descrittiva e nella produzione di output grafici replicabili.

Per consentire l'interoperabilità tra i due linguaggi, è stato utilizzato il pacchetto *reticulate*, che permette a R di richiamare Python come *backend* computazionale. In questa architettura, la suddivisione dei compiti è netta: Python viene impiegato esclusivamente per le componenti di modellazione Natural Language Processing, gestendo le fasi matematicamente più onerose dell'algoritmo BERTopic, quali la generazione degli embedding, la riduzione della dimensionalità e il clustering. R, invece, funge da ambiente principale di orchestrazione e analisi. In R vengono gestite la preparazione iniziale del corpus, la gestione dei metadati temporali e, soprattutto, l'interpretazione dei risultati. Tale approccio consente di sfruttare i modelli Transformer residenti in Python, mantenendo però in R la flessibilità necessaria per l'analisi, la validazione qualitativa dei topic e la produzione di visualizzazioni grafiche avanzate.

Poiché *reticulate* richiede una distribuzione locale di Python, è stato configurato un ambiente virtuale dedicato tramite la funzione *virtualenv_create()*. All'interno di R sono state installate le dipendenze critiche per il funzionamento dei moduli di BERTopic: *sentence-transformers* per la creazione degli embeddings semantici, *umap-learn* per la proiezione dei dati, *hdbscan* per l'individuazione dei cluster e la libreria *bertopic* per il coordinamento delle diverse fasi di modellazione. Questo assetto garantisce che l'intero processo sia stabile, isolato da altre installazioni di sistema e perfettamente replicabile.

Parallelamente, in ambiente R devono essere caricati i principali pacchetti necessari alle fasi di manipolazione dei dati e di pre-processing testuale. In particolare, risulta opportuno utilizzare strumenti per la gestione e il filtraggio dei dati tabellari (*dplyr*), la normalizzazione e pulizia delle stringhe testuali (*stringr*), il rilevamento automatico della lingua delle recensioni (*cld3*), la tokenizzazione dei testi (*tokenizers*) e le operazioni di normalizzazione linguistica, comprendenti lemmatizzazione e stemming (*textstem*, *SnowballC*). La gestione della dimensione temporale, funzionale all'analisi dinamica dei topic, può essere supportata dal pacchetto *lubridate*. Ulteriori pacchetti devono essere impiegati per la visualizzazione interattiva dei risultati (*plotly*), l'esportazione dei dati in formato tabellare (*openxlsx*) e il supporto alle attività di text mining secondo il paradigma

tidy (*tidytext*).

Una volta completata la configurazione dell'ambiente di lavoro, il workflow metodologico prevede la raccolta e il pre-processing dei dati, seguiti dall'implementazione del modello di Topic Modeling.

Per la raccolta dei dati, è possibile fare riferimento a piattaforme di open data, come la piattaforma Kaggle[1], ampiamente utilizzata in ambito accademico e industriale. La piattaforma fornisce strumenti per esplorare, filtrare e confrontare i dataset, facilitando la selezione dei dati più adatti agli scopi di ricerca. È consigliabile privilegiare dataset in formato strutturato, ad esempio file CSV, per facilitare l'importazione e la gestione dei dati. Inoltre, i dataset selezionati dovrebbero presentare dimensioni sufficienti a fornire una rappresentazione significativa dei temi trattati, pur rimanendo gestibili in termini di memoria e tempi di calcolo. Infine, è necessario scegliere dataset con informazioni temporali associate alle recensioni, necessarie per condurre l'analisi dinamica dei topic nel tempo. Kaggle permette di accedere a dataset eterogenei e di grandi dimensioni, riducendo significativamente i tempi necessari per la raccolta e la preparazione dei dati.

Il dataset, una volta acquisito, deve essere importato nell'ambiente di analisi per la fase di pre-processing. Questo step ha l'obiettivo di rendere i testi uniformi e idonei all'analisi tramite BERTopic. In primo luogo, è necessario convertire i testi in formato UTF-8 (un sistema di codifica che garantisce la corretta rappresentazione di simboli e accenti provenienti da diverse lingue) e rimuovere eventuali caratteri invisibili o residui di formattazione, ad esempio ritorni a capo o tabulazioni. Le osservazioni prive di contenuto dopo questi primi filtri devono essere eliminate. Successivamente, può essere opportuno filtrare i documenti per lingua, utilizzando la libreria *cld3*, al fine di garantire coerenza con gli strumenti di elaborazione del linguaggio naturale impiegati. Infine, per garantire una densità informativa sufficiente, è consigliabile rimuovere i testi eccessivamente brevi.

Una volta definito il sottoinsieme rilevante di documenti, ciascun testo può essere sottoposto a una funzione di pulizia che includa:

- la conversione delle lettere in minuscolo;
- la rimozione di URL e caratteri non alfabetici;
- la normalizzazione degli spazi;

[1] https://www.kaggle.com/

- la tokenizzazione del testo;
- l'eliminazione dei token troppo brevi;
- l'applicazione della lemmatizzazione;
- la ricomposizione dei token normalizzati in una stringa testuale pulita.

Al termine di questa fase, i documenti che risultano troppo brevi o scarsamente informativi possono essere esclusi, mantenendo solo testi che soddisfano una soglia minima di contenuto.

A seguito delle operazioni di filtraggio e pulizia, è buona pratica monitorare la numerosità del corpus risultante, al fine di valutare l'impatto delle scelte di pre-processing sulla dimensione del campione analizzato.

La rimozione esplicita delle stopwords non è necessariamente richiesta nella fase di pre-processing per BERTopic. Infatti, l'utilizzo di embedding semantici basati su Sentence-BERT riduce naturalmente l'impatto dei termini funzionali nella rappresentazione dei documenti, concentrandosi sulla semantica dei termini significativi. Inoltre, il processo di clustering si basa su similarità semantiche e non su frequenze lessicali. Infine, il calcolo del c-TF-IDF penalizza automaticamente i termini ad alta frequenza e a basso potere discriminante.

Nel caso di analisi temporali, è necessario procedere alla standardizzazione della variabile temporale associata a ciascun documento. La data originale deve essere convertita in un formato coerente e successivamente aggregata secondo la granularità temporale prescelta (ad esempio annuale o mensile). La scelta del livello di aggregazione temporale non è intrinseca al modello, ma dipende dagli obiettivi dell'analisi e dalla distribuzione dei dati.

È fondamentale che l'informazione temporale rimanga associata a ciascun documento lungo tutto il workflow. Solo mantenendo questo legame è infatti possibile, una volta identificati i topic, analizzarli in ogni specifico intervallo di tempo.

Il dataset finale, contenente sia i testi puliti sia le informazioni temporali associate, deve essere salvato in un formato che ne garantisce la riproducibilità e il riutilizzo nelle successive fasi di analisi.

La Tabella 1 riassume in modo schematizzato il workflow della fase di pre-processing, indicando per ciascuna fase gli input utilizzati, le operazioni svolte e gli output generati.

Tabella 1 - Pre-processing: input, operazioni e output

STEP	INPUT	OPERAZIONI	OUTPUT
Setup ambiente	Ambiente vuoto	Installazione pacchetti; creazione e configurazione ambiente virtuale tramite *reticulate*	Ambiente integrato R–Python configurato e pronto
Importazione dataset	File CSV	Lettura del CSV	Dataframe *recensioni*
Conversione encoding	Colonna Review del dataframe *recensioni*	Conversione in UTF-8, rimozione caratteri non validi	Testi della colonna Review uniformati
Rimozione caratteri invisibili	Colonna Review del dataframe *recensioni*	Rimozione tabulazioni e ritorni a capo	Testi della colonna Review privi di caratteri invisibili
Rimozione recensioni vuote	Colonna Review del dataframe *recensioni*	Eliminazione righe vuote	Dataframe senza righe con celle della colonna Review vuote
Filtro lingua target	Colonna Review del dataframe *recensioni*	Rilevamento automatico della lingua tramite *cld3* e selezione delle sole recensioni nella lingua target	Dataframe *recensioni_en* monolingua
Filtro lunghezza minima	Colonna Review del dataframe *recensioni_en*	Rimozione dei testi al di sotto di una soglia minima di parole	Dataframe *recensioni_en* adeguato
Pulizia e normalizzazione testo (*clean_text*)	Colonna Review del dataframe *recensioni_en*	Conversione in minuscolo; rimozione URL e caratteri non alfabetici; normalizzazione spazi	Testi normalizzati
Tokenizzazione e lemmatizzazione (*clean_text*)	Colonna Review del dataframe *recensioni_en*	Tokenizzazione; rimozione token brevi; lemmatizzazione; ricostruzione testo	Colonna CleanReview
Filtro recensioni troppo brevi	Colonna CleanReview del dataframe *recensioni_en*	Filtro su lunghezza minima (≥ 5 parole)	Colonna CleanReview senza recensioni troppo brevi
Conversione data	Colonna Date del dataframe *recensioni_en*	Conversione in formato *Date*	Colonna Date_clean del dataframe *recensioni_en*
Creazione variabile temporale	Colonna Date_clean del dataframe *recensioni_en*	Estrazione dell'unità temporale dalla data	Colonna Year
Salvataggio dati	Dataframe *recensioni_en*	Salvataggio del dataset preprocessato	File *recensioni_pulite.rds* contenente il corpus preprocessato
Estrazione testi finali	Colonna CleanReview del dataframe *recensioni_en*	Salvataggio della colonna CleanReview in *testi*	Vettore *testi*

4.3. Implementazione del modello

4.3.1. Allenamento modello statico

Una volta completata la fase di pre-processing, il passo successivo riguarda la costruzione del modello BERTopic, che costituisce la base metodologica per le successive analisi e per l'analisi temporale *ex-post* dell'evoluzione lessicale dei topic nel tempo.

L'addestramento del modello viene effettuato utilizzando la libreria Python BERTopic, integrata nell'ambiente R tramite il pacchetto *reticulate*, insieme alle principali dipendenze necessarie al suo funzionamento, come citato nel precedente capitolo. Per la rappresentazione semantica dei testi viene impiegato un modello di embedding pre-addestrato, la scelta del quale dipende dalle caratteristiche del corpus analizzato, in particolare dalla lunghezza e complessità dei testi.

Sulla base della scelta del modello di embedding, i testi vengono proiettati in uno spazio vettoriale ad alta dimensionalità tramite l'utilizzo di embedding densi. Questa tecnica permette di trasformare ciascun documento in un vettore numerico capace di catturare le relazioni semantiche latenti, superando i limiti della semplice corrispondenza lessicale. Per ottimizzare le prestazioni del backend computazionale, il calcolo di tali rappresentazioni vettoriali viene eseguito mediante una procedura batch-oriented, che consente una gestione efficiente delle risorse di memoria e una riduzione dei tempi di elaborazione, particolarmente rilevante in presenza di dataset voluminosi.

Per quanto riguarda la fase di riduzione dimensionale, viene utilizzato l'algoritmo UMAP (Uniform Manifold Approximation and Projection), al fine di rendere più efficiente il successivo clustering preservando al contempo la struttura semantica dei dati (McInnes et al., 2020). L'applicazione di UMAP richiede la definizione di specifici parametri di funzionamento, di cui di seguito viene indicato il ruolo e l'impatto.

Il parametro *n_neighbors* controlla il numero di punti utilizzati per stimare la struttura locale dei dati e assume tipicamente valori compresi tra 5 e 50. Valori più bassi enfatizzano le relazioni locali, favorendo la formazione di cluster più piccoli e semanticamente specifici, mentre valori più elevati privilegiano la struttura globale dello spazio, producendo topic più ampi e meno granulari.

Il parametro *n_components* definisce la dimensionalità dello spazio ridotto. In letteratura, tale parametro assume generalmente valori compresi tra 2 e 15: valori troppo bassi possono comportare una perdita di informazione semantica rilevante, mentre valori più elevati aumentano la complessità del clustering senza apportare benefici significativi in termini di separazione dei topic.

Il parametro *min_dist*, che può variare nell'intervallo [0, 1], regola la distanza minima consentita tra i punti nello spazio proiettato: valori prossimi a zero favoriscono la formazione di cluster più compatti e ben separati, mentre valori più elevati determinano una distribuzione più uniforme dei punti e una minore distinzione tra i cluster. Il valore selezionato deve consentire un equilibrio tra compattezza dei cluster e stabilità della proiezione.

Al fine di garantire la riproducibilità dei risultati, è possibile fissare il parametro *random_state*, che inizializza in modo deterministico le componenti di UMAP, assicurando che a parità di input e parametri il processo di riduzione dimensionale produca risultati coerenti tra esecuzioni successive.

Come metrica di distanza si consiglia di utilizzare la *cosine distance*, particolarmente adatta agli embedding testuali generati tramite modelli Sentence-Transformers.

Il clustering dei documenti viene effettuato mediante HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), algoritmo di clustering basato sulla densità particolarmente adatto all'analisi di testi non strutturati, caratterizzati dalla presenza di rumore e da cluster di forma irregolare (Campello et al., 2015). A differenza di altri algoritmi, HDBSCAN consente di identificare automaticamente il numero di cluster e di isolare i documenti meno informativi, classificandoli come rumore. Anche HDBSCAN richiede la definizione di parametri specifici.

Il parametro *min_cluster_size* definisce il numero minimo di documenti necessari affinché un cluster venga considerato valido e assume generalmente valori compresi tra 30 e alcune centinaia, in funzione della dimensione e della variabilità del dataset. Valori più elevati riducono il numero di topic individuati, migliorandone la robustezza e l'interpretabilità, mentre valori più bassi consentono di identificare anche temi meno frequenti, ma potenzialmente più instabili.

Il parametro *min_samples* controlla il livello di conservatività del clustering. Questo parametro influenza la densità minima richiesta affinché un punto venga considerato appartenente a un cluster e assume tipicamente valori compresi tra 1 e 50, in funzione del grado di rumore presente nei dati. Valori più elevati rendono il modello più restrittivo, aumentando il numero di documenti classificati come rumore, mentre valori più bassi favoriscono l'assegnazione dei documenti incerti a cluster esistenti.

Come metrica di distanza si consiglia la distanza euclidea, coerente con lo spazio ridotto generato da UMAP. È possibile, inoltre, abilitare l'opzione *prediction_data = TRUE*, necessaria per il calcolo delle distribuzioni documento–topic e per l'ottenimento delle misure di affinità utilizzate successivamente da BERTopic.

Per la rappresentazione lessicale dei topic viene utilizzato un CountVectorizer, impiegato da BERTopic per la costruzione del vocabolario e della matrice termini–documenti necessaria alla fase di estrazione delle parole rappresentative dei topic. Il vectorizer definisce le regole di tokenizzazione e selezione dei termini che verranno applicati ai documenti già raggruppati in cluster.

Il parametro *min_df* può essere definito sia in termini assoluti che proporzionali rispetto alla dimensione del corpus. Serve ad includere nel vocabolario esclusivamente i termini che compaiono in una quota minima di documenti. L'utilizzo di un criterio proporzionale consente di adattare automaticamente il filtraggio lessicale alla dimensione del dataset, mantenendo una rappresentazione coerente anche in presenza di cluster di dimensione eterogenea. Valori più restrittivi riducono l'inclusione di termini rari e potenzialmente rumorosi, mentre soglie più permissive favoriscono una maggiore copertura lessicale, a rischio di introdurre rumore.

In aggiunta, viene introdotto il parametro *max_df* al fine di escludere i termini che compaiono in una quota eccessivamente elevata di documenti e che risultano quindi poco discriminanti per la caratterizzazione semantica dei topic. Si fa ciò per evitare che termini altamente frequenti ma poco informativi dominino la rappresentazione lessicale dei topic. L'uso combinato di *min_df* e *max_df* consente di bilanciare la rimozione del rumore lessicale con il mantenimento di termini informativi, migliorando la qualità e l'interpretabilità delle parole chiave dei topic individuati.

Nel vectorizer è possibile considerare sia unigrammi sia bigrammi (*ngram_range = (1, 2)*), così da catturare non solo singoli concetti ma anche espressioni multi-parola semanticamente rilevanti, particolarmente frequenti in alcuni contesti, come quello delle recensioni online. Si rimuovono le stopword della lingua target selezionata ed eventualmente delle stopword di dominio,

A questo punto il modello BERTopic può essere inizializzato integrando in modo esplicito il modello di embedding, il modulo di riduzione dimensionale UMAP, l'algoritmo di clustering HDBSCAN e il CountVectorizer, garantendo una gestione coerente e integrata delle diverse componenti del modello BERTopic.

La dimensione minima di un topic finale (*min_topic_size*) deve essere fissata in coerenza con il parametro *min_cluster_size* utilizzato nella fase di clustering. Tale parametro assume generalmente valori compresi tra 30 e 200 e consente di filtrare cluster poco rappresentativi o instabili, migliorando la robustezza e la leggibilità complessiva dei risultati.

Infine, è richiesto di settare il numero di parole rappresentative per ciascun topic (*top_n_words*), valore tipicamente compreso tra 5 e 15. È, inoltre, possibile abilitare l'opzione *calculate_probabilities*, che consente di ottenere, oltre all'assegnazione deterministica dei documenti ai topic, una matrice densa in cui, per ogni documento, viene calcolata la probabilità di appartenenza di un documento a ciascuno dei topic individuati. Questa rappresenta la base quantitativa per la fase successiva di *reduce_outliers.*

L'addestramento del modello viene eseguito tramite la funzione *fit_transform*, sfruttando un backend multithread basato sulla libreria *joblib*, con impiego di tutti i core disponibili (n_jobs = −1), al fine di ridurre i tempi di calcolo. Il metodo restituisce sia il vettore di assegnazione dei topic sia la matrice delle affinità documento–topic, che devono essere conservate per le analisi successive e per l'estensione temporale dei risultati.

Al termine dell'addestramento è prassi calcolare alcune statistiche descrittive di controllo, tra cui la percentuale di documenti classificati come rumore, il numero complessivo di topic individuati e la distribuzione dei documenti all'interno dei cluster, al fine di valutare la stabilità e la qualità del clustering.

Sebbene BERTopic restituisca una matrice denominata "probabilità", tali valori non rappresentano vere probabilità di appartenenza in senso statistico (Grootendorst, 2022). Poiché il clustering è effettuato tramite HDBSCAN, ogni documento viene assegnato in modo deterministico a un singolo topic (o al rumore). I valori presenti in questa matrice di "probabilità" forniti da BERTopic esprimono invece una misura di affinità semantica tra ciascun documento e i diversi topic, calcolata sulla base della distanza nello spazio degli embedding e successivamente normalizzata. Essi possono pertanto essere interpretati come indicatori di similarità relativa, utili per il ranking e l'analisi dei documenti, ma non come probabilità nel senso statistico del termine.

La Tabella 2 riassume i passaggi principali dello script, evidenziando per ciascuno gli input, le operazioni da svolgere e gli output.

Tabella 2 - Modello statico: input, operazioni e output

STEP	INPUT	OPERAZIONI	OUTPUT
Importazione librerie Python	Ambiente R configurato	Importazione delle librerie Python necessarie (bertopic, sentence_transformers, umap, hdbscan, joblib, sklearn) tramite *reticulate*	Librerie Python necessarie caricate
Caricamento modello di embedding	Modello di embedding scelto	Inizializzazione del modello di embedding tramite SentenceTransformer	Oggetto *encoder*
Creazione embeddings	Vettore *testi*	Encoding dei testi in rappresentazioni vettoriali dense	Matrice *embeddings*
Configurazione UMAP	Matrice *embeddings*	Riduzione dimensionale tramite UMAP	Oggetto *umap_model*
Configurazione HDBSCAN	Matrice *embeddings*	Clustering tramite HDBSCAN	Oggetto *hdbscan_model*
Configurazione CountVectorizer	Corpus testuale	Definizione del vocabolario e delle regole di tokenizzazione tramite CountVectorizer	Oggetto *vectorizer*
Configurazione BERTopic	*encoder* , *umap* , *hdbscan, vectorizer*	Inizializzazione del modello BERTopic con parametri personalizzati	Oggetto *model* (modello statico pronto per l'addestramento)
Addestramento modello	Vettore *testi,* matrice *embeddings*	Addestramento del modello tramite fit_transform() con eventuale parallelizzazione	Vettori *topics* e matrice *probs*
Salvataggio risultati	Modello, topics, probs, embeddings, testi, date	Salvataggio degli oggetti R in formato RDS	File .rds contenenti modello e risultati dell'analisi
Statistiche sui topic	Vettore *topics*	Calcolo delle statistiche descrittive (percentuale di rumore, numero di topic, distribuzione dei documenti)	Indicatori descrittivi sul risultato del clustering
Statistiche sulle probabilità	Matrice *probs*	Analisi della distribuzione delle probabilità di appartenenza dei documenti ai topic	Indicatori sulla struttura probabilistica dei topic

4.3.2. Labeling e raffinamento dei topic

Dopo la fase di training del modello BERTopic, è necessaria una fase di labeling e raffinamento dei topic, con l'obiettivo di migliorare l'interpretabilità semantica dei risultati e garantire una nomenclatura coerente e utilizzabile nella successiva analisi dinamica.

In una fase preliminare, si applica un filtraggio del rumore, finalizzato a migliorare la qualità dell'assegnazione dei documenti ai topic. In particolare, utilizzando la funzione *reduce_outliers()* di BERTopic con strategia basata sulle probabilità di appartenenza, risulta possibile riesaminare i documenti inizialmente classificati come rumore, cioè associati al topic −1. La procedura consente di riassegnare ai topic validi quei documenti che presentano una probabilità di appartenenza sufficientemente elevata, riducendo la quota di rumore senza alterare la struttura del clustering originale. Approcci analoghi di filtraggio e raffinamento dei topic basati sulle probabilità di appartenenza o sulla distanza semantica sono ampiamente adottati nella letteratura sul Topic Modeling e sui modelli basati su embedding, in diversi contesti applicativi (Ali et al., 2025; López et al., 2024; Tangherlini and Chen, 2024). Tale approccio consente di migliorare la copertura dei topic senza compromettere la qualità complessiva dei risultati. I criteri specifici adottati per la riassegnazione dei documenti e la calibrazione della procedura sono da definire in funzione delle caratteristiche del dataset analizzato e delle prove empiriche di rumore prima e dopo l'applicazione della funzione *reduce_outliers()*.

Per supportare il processo di interpretazione e labeling dei topic, viene condotta un'analisi di similarità semantica tra i documenti e i topic appresi dal modello. In primo luogo, si definisce una funzione per il calcolo della similarità coseno tra due vettori di embedding, misura ampiamente utilizzata in ambito NLP per valutare la vicinanza semantica tra rappresentazioni vettoriali in spazi ad alta dimensionalità (Reimers and Gurevych, 2019). Successivamente, si estraggono gli embedding dei topic appresi dal modello BERTopic e gli embedding delle recensioni, entrambi nello stesso spazio semantico. L'analisi è da limitare ai soli topic validi, escludendo il rumore, al fine di garantire coerenza interpretativa.

Per ciascun topic valido, si calcola quindi la similarità coseno tra il vettore rappresentativo del topic e tutti i vettori dei documenti. Questo passaggio produce una matrice di similarità documento–topic, in cui ogni colonna rappresenta un topic e ogni riga una recensione, con valori che indicano il grado di allineamento semantico tra documento e topic.

A partire da questa matrice, per ogni topic si selezionano esclusivamente le recensioni che il modello assegna a quel topic. Tali recensioni vengono poi ordinate in modo decrescente in base al valore di

similarità, ottenendo così una graduatoria dei documenti più rappresentativi rispetto alla semantica del cluster.

Il risultato finale viene organizzato in una struttura tabellare contenente, per ciascun topic, l'identificativo del topic, il valore di similarità coseno e il testo originale della recensione. Infine, è consigliato esportare questo output in un file CSV, contenente l'elenco completo delle recensioni ordinate per similarità all'interno di ciascun topic.

Questa procedura rappresenta uno strumento chiave per l'ispezione qualitativa dei risultati, consentendo di analizzare in modo sistematico i documenti più rappresentativi di ciascun topic, validarne il contenuto semantico e supportare il processo di assegnazione delle etichette interpretative.

In parallelo all'analisi di similarità semantica, è possibile estrarre dal modello le informazioni descrittive associate a ciascun cluster. In particolare, BERTopic produce per ogni topic una rappresentazione basata sulle parole statisticamente più rappresentative, calcolate tramite la misura c-TF-IDF, insieme alla numerosità dei documenti assegnati.

Sulla base di tali informazioni, per ciascun topic viene generata un'auto-etichetta preliminare ottenuta combinando le parole chiave maggiormente rilevanti. Questa etichettatura automatica ha esclusivamente una funzione di supporto iniziale all'analisi, consentendo una prima sintesi del contenuto dei topic, ma non è da considerarsi sufficiente né immediatamente interpretabile in ottica manageriale.

Per questo motivo, la fase di topic labeling è consigliabile condurla seguendo un approccio human-in-the-loop. Le informazioni di sintesi sui topic (identificativo, numerosità dei documenti, parole chiave e auto-etichetta) esportate in formato Excel, devono essere analizzate congiuntamente al file contenente, per ciascun topic, l'elenco completo delle recensioni assegnate, ordinate in base alla similarità coseno rispetto all'embedding del topic.

Questo approccio consente di concentrare l'analisi qualitativa sulle recensioni semanticamente più rappresentative di ciascun cluster, evitando la lettura indiscriminata dell'intero corpus. L'ispezione sistematica dei documenti con valori di similarità più elevati permette di comprendere in modo approfondito il contenuto semantico dominante di ciascun topic e di identificarne il significato latente.

Sulla base di tale analisi, si assegna a ciascun topic un'etichetta descrittiva finale (HumanLabel), formulata come una breve denominazione testuale in grado di sintetizzare in modo chiaro, coerente e

neutro il tema trattato. Una volta completata la fase di etichettatura manuale, le etichette devono essere reimportate nell'ambiente di analisi e integrate con le informazioni originali sui topic. Infine, è buona pratica effettuare controlli automatici per verificare la presenza dell'etichetta umana per ciascun topic e l'assenza di valori mancanti, garantendo la coerenza e la completezza del processo di labeling.

Nel corso di questa fase, alcuni topic distinti dal modello possono risultare semanticamente sovrapposti. Tale fenomeno è spesso riconducibile alla capacità dell'algoritmo di cogliere sfumature lessicali molto sottili che, tuttavia, afferiscono al medesimo nucleo tematico dal punto di vista interpretativo. Per risolvere queste ridondanze, si procede al merging dei topic, un'operazione che aggrega i cluster. Sulla base della similarità semantica delle parole rappresentative e dell'ispezione qualitativa delle recensioni con elevata similarità documento–topic, tali topic possono essere aggregati e associati a una medesima etichetta interpretativa.

I topic secondari derivanti dal merging devono essere successivamente rimossi dal dataset finale, così da evitare duplicazioni e garantire una rappresentazione univoca, stabile e interpretabile dei temi analizzati. Il set finale di topic ottenuto deve essere utilizzato come riferimento per le successive analisi e visualizzazioni, nonché per l'analisi dinamica dell'evoluzione dei topic nel tempo.

La Tabella 3 riassume le operazioni dedicate alla fase di labeling e analisi dei topic, indicando per ciascun passaggio gli input necessari, le operazioni da eseguire e gli output.

Tabella 3 - Labeling: input, operazioni e output

STEP	INPUT	OPERAZIONI	OUTPUT
Filtraggio del rumore	Modello BERTopic, testi, topics, probs	Riassegnazione dei documenti rumore tramite reduce_outliers() basata sulle probabilità di appartenenza (soglia = 0,2)	Vettore topics_refined
Valutazione impatto filtraggio	topics, topics_refined	Confronto della quota di documenti rumore prima e dopo il filtraggio	Tabelle di allocazione e metriche di recupero
Estrazione embedding dei topic	Modello BERTopic addestrato	Estrazione di model$topic_embeddings_	Matrice di embedding dei topic
Selezione dei topic validi	topics_refined	Esclusione del topic di rumore (ID = −1)	Vettore valid_topics
Definizione metrica di similarità	Embedding documenti e topic	Definizione della funzione di similarità coseno	Funzione cosine_sim()
Calcolo similarità documento–topic	Embedding documenti e topic validi	Calcolo della similarità coseno tra ogni documento e ciascun topic	Matrice di similarità documento-topic (similarity_matrix)
Selezione documenti per topic	similarity_matrix, topics_refined	Selezione dei documenti assegnati dal modello a ciascun topic	Sottoinsiemi documento–topic
Ordinamento recensioni	Similarità documento–topic	Ordinamento decrescente dei documenti in base alla similarità coseno	Dataframe topic_examples_df
Esportazione recensioni rappresentative	topic_examples_df	Scrittura file CSV	all_topics_all_reviews_ordered_similarity.csv
Estrazione informazioni sui topic	Modello BERTopic	Esecuzione di model$get_topic_info()	Dataframe topic_info
Rimozione topic di rumore	topic_info	Eliminazione del topic con ID = −1	Dataframe topic_info filtrato
Creazione rappresentazione testuale	Colonna Representation	Conversione della lista di parole in stringa leggibile	Colonna Words
Generazione AutoLabel	Colonna Representation	Creazione etichetta automatica con le prime tre parole chiave	Colonna AutoLabel
Preparazione file per labeling manuale	topic_info filtrato	Selezione delle colonne informative rilevanti	topic_info_for_labeling
Esportazione per labeling umano	topic_info_for_labeling	Esportazione in formato Excel	topic_info_for_labeling.xlsx
Reimportazione etichette manuali	File Excel	Lettura colonna HumanLabel	Dataframe topic_labels
Validazione etichette	topic_labels	Controllo di completezza e assenza di valori mancanti	Etichette validate
Integrazione etichette finali	topic_info, topic_labels	Merge delle etichette umane sui topic	topic_info_labeled
Merging dei topic simili	topic_info_labeled, topics_refined	Aggregazione manuale dei topic semanticamente sovrapposti e rimozione topic assorbiti	topics_merged, topic_info_merged
Salvataggio output intermedi e finali	Topic etichettati e aggregati	Salvataggio in formato RDS	topic_info_labeled_final.rds, topic_info_final.rds, topic_info_merged.rds topics_merged_vector.rds

4.3.3. Visualizzazione dei topic

Dopo l'addestramento del modello statico, la fase di labeling manuale e l'aggregazione dei topic semanticamente sovrapposti, vengono realizzate una serie di visualizzazioni con l'obiettivo di esaminare in modo sistematico la struttura tematica emersa dall'analisi. Prima di procedere all'analisi dinamica, risulta infatti utile verificare la distribuzione dei topic, la loro coerenza interna e le relazioni di similarità tra i cluster individuati.

In primo luogo, è possibile analizzare la distribuzione dimensionale dei cluster finali al fine di rappresentare in modo sintetico la struttura definitiva utilizzata nelle analisi successive. A partire dal vettore di assegnazione dei documenti ai topic aggregati (*topics_merged*), si calcola il numero di recensioni associate a ciascun topic, ottenendo una misura quantitativa della rilevanza relativa dei temi individuati all'interno del corpus. Questa rappresenta l'unica visualizzazione costruita direttamente sui topic filtrati dal rumore e successivamente aggregati, in quanto basata su informazioni derivate esternamente al modello.

I conteggi vengono quindi integrati con le informazioni descrittive dei topic finali, includendo le etichette interpretative assegnate manualmente (HumanLabel) e le parole statisticamente più rappresentative derivate dalla misura c-TF-IDF. Il dataset risultante viene ordinato in base alla numerosità decrescente dei documenti, così da facilitare l'identificazione dei topic predominanti. Sulla base di tali informazioni si costruisce una visualizzazione tramite grafico a barre utilizzando la libreria *plotly*. Il grafico rappresenta, per ciascun topic finale, il numero di recensioni associate, consentendo di valutare sia la distribuzione della rilevanza tematica all'interno del corpus sia l'eventuale presenza di squilibri dimensionali tra i cluster. A supporto dell'interpretazione, per ogni barra è possibile rendere disponibili informazioni di dettaglio consultabili tramite interazione. Infine, è opportuno esportare la visualizzazione in formato HTML, al fine di consentirne un'esplorazione interattiva.

Inoltre, è opportuno analizzare le caratteristiche semantiche dei topic prima dell'aggregazione dei topic semanticamente sovrapposti, utilizzando le visualizzazioni standard messe a disposizione da BERTopic. In particolare, tramite la funzione *model$visualize_barchart(),* viene generato un grafico che riporta, per ciascun topic individuato dal modello (escludendo il rumore), le cinque parole più rappresentative secondo la misura c-TF-IDF. Tali visualizzazioni non possono essere applicate direttamente ai topic aggregati, in quanto si basano sulle strutture interne del modello addestrato, che non vengono aggiornate a seguito di operazioni di merging manuale effettuate *ex-post*.

Il barchart viene pertanto utilizzato come strumento di supporto all'analisi qualitativa dei topic originari, permettendo di ispezionarne la composizione lessicale e valutarne la coerenza interna, senza finalità di confronto quantitativo diretto tra i pesi delle parole appartenenti a topic diversi.

Per esplorare le relazioni di similarità tra i topic nello spazio semantico, viene inoltre generata una mappa bidimensionale dei cluster tramite la funzione *model$visualize_topics()*. Questa rappresentazione si basa su una tecnica di riduzione dimensionale applicata agli embedding dei topic e consente di visualizzare le distanze relative tra i cluster, evidenziando eventuali prossimità o sovrapposizioni nello spazio semantico.

È tuttavia importante sottolineare che questa visualizzazione ha una finalità prevalentemente esplorativa. La vicinanza spaziale tra due topic nella mappa non implica necessariamente una sovrapposizione semantica completa, in quanto la riduzione dimensionale può introdurre distorsioni e perdita di informazione. Per questo motivo, anche in presenza di topic apparentemente molto prossimi o quasi coincidenti nella mappa, la valutazione finale circa la loro effettiva similarità deve essere condotta tramite un'analisi qualitativa dei contenuti testuali. Per questo motivo, la decisione di mantenere distinti o aggregare topic semanticamente vicini deve essere supportata dall'ispezione delle recensioni maggiormente rappresentative di ciascun cluster, consentendo di verificare se le affinità osservate nello spazio degli embedding si riflettessero effettivamente in una sovrapposizione tematica a livello di contenuto.

A completamento dell'analisi, si può visualizzare la medesima informazione rappresentata in forma matriciale tramite la funzione *model$visualize_heatmap()*, che restituisce una matrice di similarità tra topic basata sulla vicinanza semantica degli embedding. Mentre la mappa bidimensionale fornisce una visione intuitiva e qualitativa delle relazioni tra i cluster, la matrice di similarità permette una valutazione più analitica e comparativa del grado di affinità tra ciascuna coppia di topic. Insieme, le due visualizzazioni offrono una lettura complementare della struttura semantica dei topic appresi dal modello.

Per facilitare la consultazione dei risultati e l'utilizzo dei dati in analisi successive, è opportuno raccogliere le informazioni finali sui topic aggregati in un file CSV. In particolare, per ciascun topic vengono riportati l'identificativo numerico, l'etichetta descrittiva finale (HumanLabel), il numero di recensioni associate e le parole rappresentative, convertite in stringhe testuali separate da virgole per garantire la compatibilità con strumenti esterni, come fogli di calcolo.

La Tabella 4 riassume i principali passaggi dedicati alla fase di visualizzazione e analisi esplorativa dei topic.

Tabella 4 - Visualizzazione: input, operazioni e output

STEP	INPUT	OPERAZIONI	OUTPUT
Calcolo frequenze topic post-merge	Vettore topics_merged	Conteggio delle recensioni per ciascun topic tramite table()	Dataframe topic_counts con frequenze per topic
Preparazione dati per barchart post-merge	topic_info_merged, topic_counts	Join sulle variabili Topic e ordinamento per frequenza decrescente	Dataframe topic_info_plot
Barchart distribuzione topic (post-merge)	Dataframe topic_info_plot	Visualizzazione interattiva con plot_ly() delle recensioni per topic	Grafico HTML barchart_topics_merged.html
Estrazione informazioni topic pre-merge	Modello BERTopic addestrato	Chiamata a model$get_topic_info() ed esclusione del topic di rumore (Topic = −1)	Dataframe topic_info
Barchart parole rappresentative (pre-merge)	Modello BERTopic addestrato	Chiamata a model$visualize_barchart()	File HTML barchart_topics.html
Mappa bidimensionale dei topic (pre-merge)	Modello BERTopic addestrato	Chiamata a model$visualize_topics()	File HTML topics_map.html
Heatmap di similarità tra topic (pre-merge)	Modello BERTopic addestrato	Chiamata a model$visualize_heatmap(n_clusters = 10)	File HTML topics_heatmap.html
Costruzione tabella finale dei topic	topic_info_merged, topic_counts	Join e selezione colonne Topic, HumanLabel, Count_refined, Words, AutoLabel	Dataframe topic_info_df
Esportazione risultati finali	Dataframe topic_info_df	Scrittura su file CSV	File topic_info_df.csv

La Figura 9 rappresenta lo schema complessivo del processo di pre-processing e di costruzione del modello statico basato su BERTopic, mostrando le relazioni tra dataset, fasi di modellazione, aggregazione dei topic e visualizzazioni finali.

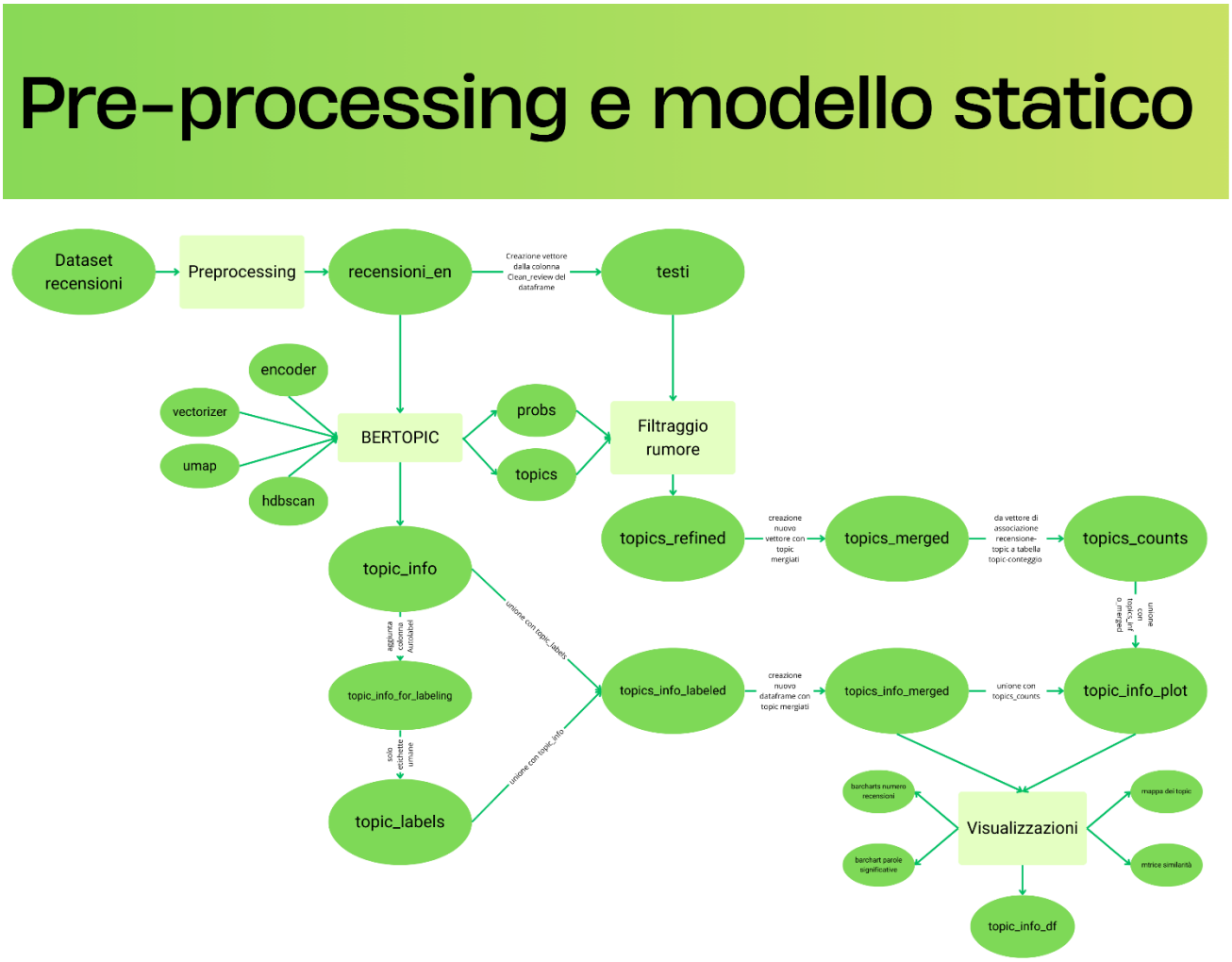

Figura 9 - Diagramma di flusso di pre-processing e modello statico

4.4. Validazione del modello statico

4.4.1. Validazione non supervisionata del modello statico

La validazione del modello BERTopic si conduce, in una prima fase, attraverso procedure non supervisionate, con l'obiettivo di valutare la qualità intrinseca dei topic generati. In linea con la letteratura sul Topic Modeling, sono adottate due metriche complementari: la topic coherence e la topic diversity, che consentono di valutare rispettivamente la coerenza semantica interna dei topic e il grado di distinzione terminologica tra di essi, fornendo una valutazione complessiva della qualità e dell'informatività del modello individuato.

4.4.1.1. Topic coherence

La coerenza dei topic è valutata tramite la metrica *U_mass*, una misura intrinseca basata sulle co-occorrenze dei termini all'interno del corpus, introdotta da (Mimno et al., 2011). Sebbene per altri dataset o contesti di ricerca possa risultare opportuno l'impiego di metriche basate su riferimenti esterni

(come Cv, che utilizza Wikipedia come corpus di confronto), nel contesto della Digital VoC è preferibile un approccio intrinseco.

La natura gergale, l'uso di acronimi e la specificità terminologica delle recensioni online rendono infatti necessario validare i topic rispetto al reale contesto d'uso dei consumatori. L'utilizzo di *U_mass* permette di evitare le distorsioni derivanti dal confronto con linguaggi generalisti o enciclopedici, garantendo una valutazione coerente con il lessico specifico del dominio analizzato.

La metrica *U_mass* quantifica il grado di associazione semantica tra le parole più rappresentative di ciascun topic e assume valori nel dominio $(-\infty, 0]$, dove valori più prossimi a zero indicano una maggiore coerenza semantica tra le parole che compongono ciascun topic. *U_mass* risulta particolarmente adatta in assenza di un corpus di riferimento esterno, in quanto, a differenza di altre misure di coerenza, consente una valutazione intrinseca basata esclusivamente sul corpus di analisi.

Dal punto di vista implementativo, il calcolo della *U_mass* viene effettuato utilizzando la libreria *gensim* tramite integrazione R–Python. In primo luogo, vengono estratte per ciascun topic le parole rappresentative finali, ottenute dal modello BERTopic dopo il processo di merging dei topic semanticamente simili. Le parole, inizialmente rappresentate come stringhe separate da virgole, sono trasformate in liste di token, procedendo alla scomposizione dei bigrammi in unigrammi. Tale scelta è coerente con l'impostazione della metrica *U_mass*, che si basa sul calcolo delle co-occorrenze tra termini all'interno di una rappresentazione bag-of-words del corpus, garantendo coerenza tra la forma dei token dei topic e quella utilizzata per la costruzione del corpus.

È opportuno precisare che, sebbene l'architettura di BERTopic sfrutti embedding contestuali per la fase di clustering, la validazione della coerenza avviene sulla rappresentazione lessicale dei topic (output del c-TF-IDF). Pertanto, il ricorso al modello bag-of-words in questa fase non è in contrasto con l'approccio neurale del modello, ma è necessario per quantificare statisticamente quanto spesso le parole chiave identificate tendano a comparire insieme nei documenti originali, garantendo così l'effettiva interpretabilità dei temi estratti.

Successivamente, i documenti del corpus sono tokenizzati e utilizzati per costruire un dizionario e una rappresentazione bag-of-words, ovvero una codifica dei testi come vettori di frequenza dei termini, che prescinde dall'ordine delle parole all'interno dei documenti. Tale rappresentazione è necessaria per il calcolo delle co-occorrenze tra termini all'interno del corpus. La coerenza *U_mass* è quindi calcolata confrontando, per ciascun topic, la frequenza congiunta delle parole rappresentative nei

documenti del corpus. Il risultato finale è un valore medio di coerenza, ottenuto aggregando le misure calcolate sull'insieme dei topic.

4.4.1.2. Topic diversity

Oltre alla coerenza interna, la validazione dei risultati include il calcolo della topic diversity. Sebbene questa metrica sia meno frequente nella letteratura tradizionale rispetto ai punteggi di coerenza, la sua integrazione è fondamentale per garantire la qualità di un modello basato su architetture neurali.

La topic diversity misura la percentuale di parole chiave univoche tra tutti i topic estratti: un valore elevato indica che il modello ha identificato temi distinti e non ridondanti. In un contesto di Digital VoC, dove i consumatori possono esprimere concetti simili con sfumature diverse, questa metrica assicura che il modello non soffra di sovrapposizioni semantiche, permettendo una mappatura dei temi che sia al contempo varia e informativa.

In sintesi, se la coerenza garantisce l'interpretabilità di ogni singolo tema, la diversity ne assicura l'eterogeneità complessiva, fornendo una visione d'insieme più completa e meno ripetitiva del dataset.

Seguendo la definizione proposta da (Lau et al., 2020), la topic diversity è misurata come il rapporto di termini unici sul totale delle parole chiave associate ai topic e assume valori nell'intervallo (0,1].

Dal punto di vista operativo, la metrica è calcolata considerando l'insieme delle parole rappresentative di tutti i topic finali. In modo analogo alla topic coherence, i bigrammi sono scomposti in unigrammi al fine di adottare una rappresentazione uniforme del vocabolario e rendere il calcolo della diversità più rigoroso e confrontabile. La topic diversity è quindi ottenuta come rapporto tra il numero di parole distinte e il numero totale di parole considerate.

Le metriche di validazione quantitativa adottate, topic coherence e topic diversity, consentono di valutare la qualità del modello nella sua formulazione statica, fornendo indicazioni robuste sulla coerenza interna dei topic e sulla loro non ridondanza. È tuttavia importante sottolineare che tali metriche non sono progettate per valutare direttamente la dimensione temporale dell'analisi: la validazione dell'evoluzione dei topic nel tempo è pertanto affrontata principalmente attraverso un'analisi qualitativa della coerenza e plausibilità dell'evoluzione lessicale osservata, come discusso nelle sezioni successive.

La Tabella 5 sintetizza la procedura adottata per la validazione quantitativa non supervisionata del modello di Topic Modeling. In particolare, vengono esplicitati in forma strutturata gli step operativi, gli input da utilizzare, le principali trasformazioni e gli output per il calcolo delle metriche di topic coherence e topic diversity.

Tabella 5 – Validazione del modello: input, operazioni e output

STEP	INPUT	OPERAZIONI	OUTPUT
Inizializzazione ambiente di validazione	Ambiente R configurato con *reticulate*	Installazione e importazione della libreria Python *gensim* e dei relativi moduli (*corpora* , *models*)	Ambiente R–Python pronto per il calcolo delle metriche di validazione
Estrazione delle parole rappresentative dei topic	*topic_info_merged$ Words*	Separazione delle parole per virgola e scomposizione dei bigrammi in unigrammi; rimozione dei duplicati per topic	*topics_cleaned*
Tokenizzazione del corpus testuale	*testi*	Tokenizzazione delle recensioni in liste di parole tramite *tokenization*	Corpus tokenizzato a livello di documento (*docs*)
Creazione del dizionario	*docs*	Costruzione del dizionario che associa ogni termine a un identificativo numerico	*dictionary*
Costruzione del corpus bag-of-words	*docs, dictionary*	Conversione di ogni documento nella rappresentazione bag-of-words	*corpus*
Calcolo della topic coherence (U_mass)	*topics_cleaned, dictionary, corpus*	Calcolo della coerenza U_mass tramite *CoherenceModel* di *gensim* e aggregazione del valore medio sui topic	Punteggio di topic coherence (U_mass)
Preparazione token per topic diversity	*topics_raw*	Separazione delle parole per virgola e scomposizione dei bigrammi in unigrammi; eliminazione dei token di lunghezza pari a un carattere	Vettore dei token associati ai topic (*all_words_clean*)
Calcolo della Topic Diversity	*all_words_clean*	Calcolo del rapporto tra numero di parole uniche e numero totale di parole considerate	Punteggio di topic diversity

La Figura 10 rappresenta il diagramma di flusso che mostra i principali step della validazione del modello statico.

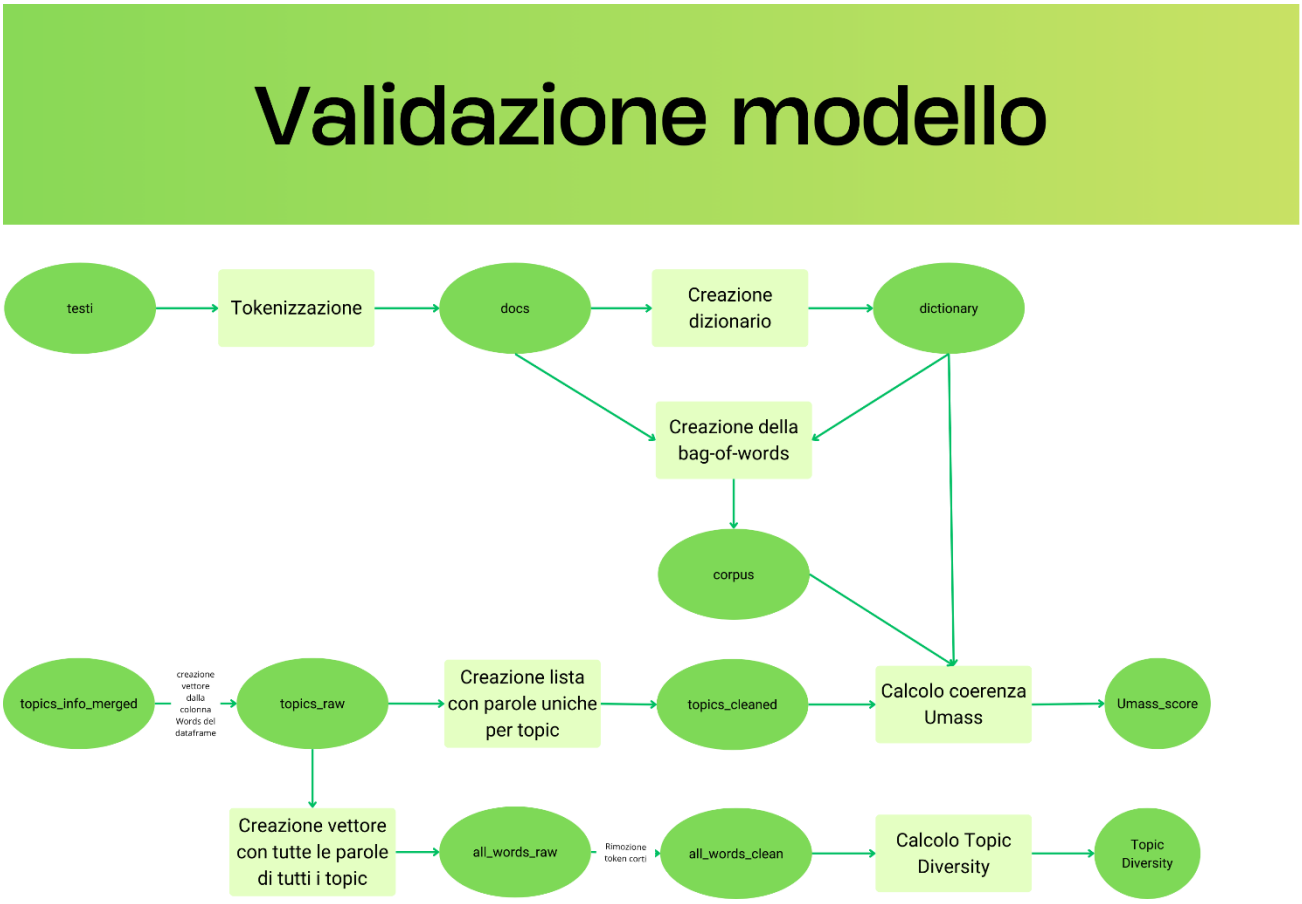

Figura 10 - Diagramma di flusso della validazione del modello

4.4.2. Validazione supervisionata del modello statico

Accanto alla validazione quantitativa non supervisionata, condotta sul modello statico, si consiglia di effettuare anche una validazione qualitativa supervisionata dei risultati, secondo un approccio human-in-the-loop. Tale fase ha l'obiettivo di verificare l'interpretabilità semantica dei topic generati e la coerenza del loro contenuto rispetto alle recensioni effettivamente associate, integrando il giudizio umano all'interno del processo di valutazione del modello.

In una prima fase, si effettua un'ispezione manuale delle parole rappresentative di ciascun topic, ovvero dei termini con maggiore rilevanza semantica individuati dal modello. Questa analisi permette di verificare la presenza di un nucleo concettuale chiaro e riconoscibile per ciascun cluster, nonché di individuare eventuali termini semanticamente ambigui, generici o incoerenti con il tema principale. I topic caratterizzati da un insieme di parole semanticamente omogeneo e facilmente interpretabile sono considerati qualitativamente validi, mentre eventuali criticità devono essere utilizzate come supporto al processo di merging.

Successivamente, la validazione qualitativa è estesa al livello documentale. Per ciascun topic, si analizzano le recensioni con maggiore similarità semantica rispetto al topic embedding. L'analisi di questi documenti consente di verificare la coerenza tra il contenuto testuale espresso nei documenti, le parole chiave del topic e l'etichetta semantica assegnata in fase di interpretazione.

Se questa verifica mostra una buona corrispondenza tra i temi espressi nelle recensioni e i topic individuati dal modello, viene confermato che l'assegnazione dei documenti ai cluster non è guidata esclusivamente da similarità superficiali, ma riflette pattern semantici interpretabili e coerenti dal punto di vista umano.

4.5. Analisi dinamica dei topic

Una volta completata la fase di identificazione e raffinamento dei topic tramite BERTopic e la validazione del modello, l'analisi può essere estesa alla dimensione temporale con l'obiettivo di studiare come la rilevanza e il contenuto lessicale dei topic statici si distribuiscano ed evolvano nel tempo.

A differenza di un approccio di Topic Modeling dinamico propriamente detto, in cui i topic vengono stimati separatamente per ciascun intervallo temporale, l'approccio adottato prevede la stima dei cluster tematici una sola volta sull'intero corpus. L'analisi temporale viene quindi condotta *ex-post*, proiettando i topic statici lungo la dimensione temporale mediante l'aggregazione dei documenti associati a ciascun topic nei diversi periodi di osservazione.

Per rendere confrontabili le osservazioni nel tempo, a ciascuna recensione bisogna anzitutto associare un periodo temporale, costruito a partire dalla data di pubblicazione. A tal fine, la variabile temporale può essere opportunamente trasformata in intervalli discreti (ad esempio anni, trimestri o mesi). L'adozione di un livello di aggregazione adeguato consente di ridurre la frammentazione temporale e di garantire una numerosità sufficiente di osservazioni per ciascun intervallo, migliorando la stabilità delle analisi. Si sottolinea che il livello di aggregazione temporale non è intrinseco al modello, ma dipende esclusivamente dalla trasformazione applicata alla variabile temporale. Di conseguenza, sebbene nello script riportato in appendice la variabile temporale sia esemplificata a livello annuale, l'analisi risulta facilmente replicabile utilizzando una diversa granularità temporale, qualora il contesto applicativo o la distribuzione dei dati lo rendano opportuno.

Sulla base dell'assegnazione di ciascuna recensione a un topic e dell'informazione temporale associata, si costruisce un dataset di base contenente, per ogni documento, il cluster tematico di appartenenza e l'anno di pubblicazione, mantenendo una corrispondenza uno a uno tra documenti, cluster e anno di

pubblicazione. Si consiglia di escludere dall'analisi le recensioni classificate come rumore.

A partire da tale struttura, si calcola la distribuzione temporale delle recensioni per topic, conteggiando per ciascun cluster il numero di recensioni associate a ogni periodo. Per ciascun topic, tali conteggi vengono successivamente normalizzati rispetto al totale delle recensioni del cluster, ottenendo una misura percentuale della distribuzione temporale interna a ciascun topic. Questa rappresentazione consente di analizzare l'evoluzione della rilevanza relativa dei topic nel tempo, evidenziando eventuali fasi di crescita, stabilità o declino, indipendentemente dalla diversa dimensione assoluta dei cluster.

Oltre alla dimensione quantitativa, l'analisi temporale viene estesa anche al contenuto semantico dei topic. Poiché i topic sono statici per costruzione, l'evoluzione è osservata esclusivamente a livello lessicale, analizzando come le parole maggiormente rappresentative di ciascun topic varino nei diversi periodi temporali.

A tal fine, i testi delle recensioni vengono sottoposti a una fase di pre-processing articolata. In particolare, viene effettuata la tokenizzazione dei testi in singole parole, seguita dalla rimozione dei termini troppo brevi, dei token contenenti caratteri non alfabetici e delle stopword standard della lingua target. Successivamente, i termini rimanenti vengono normalizzati tramite lemmatizzazione e stemming, al fine di ridurre la variabilità morfologica e migliorare la coerenza semantica delle rappresentazioni testuali a livello di termine. Infine, vengono rimosse alcune stopword specifiche di dominio, identificate manualmente, poiché altamente frequenti in determinati topic ma scarsamente informative dal punto di vista tematico.

Si evidenzia che la fase di pre-processing adottata per l'analisi temporale differisce in modo sostanziale da quella utilizzata nella fase precedente all'allenamento del modello BERTopic, pur condividendone alcune operazioni di pulizia di base. Nel modello statico, il pre-processing è condotto a livello di documento e mira alla costruzione di rappresentazioni testuali idonee al calcolo degli embedding e al successivo clustering delle recensioni. In questa fase, i testi vengono normalizzati e successivamente ricomposti in una stringa unica per ciascun documento, preservando la ricchezza semantica complessiva e privilegiando una normalizzazione moderata, coerente con l'obiettivo di catturare relazioni semantiche latenti tra i documenti.

Nell'analisi temporale, al contrario, il pre-processing è condotto a livello di singolo termine e risponde a un obiettivo differente, ossia l'estrazione di indicatori lessicali interpretabili e confrontabili nel tempo. I testi vengono pertanto scomposti in token elementari, che costituiscono l'unità di analisi per il calcolo delle misure di frequenza e della c-TF-IDF temporale. In questa fase si adotta un filtraggio

più restrittivo, sia in termini di lunghezza minima dei token sia attraverso la rimozione esplicita delle stopword standard e di dominio, al fine di ridurre il rumore e limitare l'influenza di termini altamente frequenti ma scarsamente informativi. Inoltre, la normalizzazione morfologica risulta più aggressiva, combinando lemmatizzazione e stemming, così da migliorare la stabilità dei confronti temporali.

Per analizzare l'evoluzione lessicale dei topic nel tempo, viene condotta un'analisi delle parole rilevanti per ciascuna combinazione topic–periodo mediante una misura di c-TF-IDF calcolata *ex-post* sui topic statici.

A partire dal dataset tokenizzato e pulito, bisogna innanzitutto conteggiare le occorrenze di ciascun termine all'interno di ogni coppia topic–periodo. Al fine di ridurre il rumore e limitare l'influenza di termini scarsamente informativi, vengono considerate esclusivamente le parole che presentano una frequenza minima nel periodo di riferimento.

Per ciascuna combinazione topic–periodo viene quindi calcolata la componente di term frequency (TF), definita come la frequenza relativa del termine rispetto al totale delle parole associate allo stesso topic nello stesso periodo. Successivamente, si stima la componente di inverse document frequency (IDF), calcolata sulla base del numero di coppie topic–periodo in cui ciascun termine compare, con l'obiettivo di penalizzare le parole diffuse trasversalmente nel corpus e valorizzare quelle maggiormente specifiche di determinati periodi e cluster tematici.

La misura finale di c-TF-IDF è ottenuta come prodotto tra TF e IDF, risultando elevata per i termini che sono contemporaneamente frequenti all'interno di uno specifico topic–periodo e rari nelle altre combinazioni topic–periodo. Per ciascuna coppia topic–periodo si selezionano le parole con il valore di c-TF-IDF più elevato, interpretate come le parole maggiormente rappresentative del contenuto semantico del topic in quello specifico periodo temporale.

I risultati sono successivamente aggregati in una struttura tabellare contenente, per ogni topic e per ogni periodo, l'elenco delle parole rilevanti per quel determinato periodo. A tali informazioni viene infine associata l'etichetta semantica del topic, ottenuta tramite un'operazione di join con i risultati della fase di labeling manuale condotta in precedenza, così da garantire la coerenza interpretativa tra l'analisi statica e quella temporale.

Si sottolinea che l'approccio descritto non mira a stimare topic dinamici in senso strettamente generativo, ma a fornire una lettura temporale dei topic statici, preservandone la struttura semantica complessiva e consentendo di osservare le variazioni lessicali interne ai cluster nel corso del tempo.

La Tabella 6 sintetizza il workflow dell'analisi temporale *ex-post* dei topic statici, illustrando in modo sistematico gli input da utilizzare, le trasformazioni da applicare e gli output di ciascuna fase.

Tabella 6 – Modello dinamico: input, operazioni e output

STEP	INPUT	OPERAZIONI	OUTPUT
Preparazione e validazione degli input temporali	*testi, recensioni_en$Date_clean, topics_merged*	Assegnazione degli oggetti di lavoro (docs, dates) e verifica della coerenza dimensionale tra documenti, date e topic finali	Input validati per l'analisi temporale
Costruzione variabile temporale (anno)	*dates*	Estrazione dell'anno dalla data di pubblicazione e associazione di ciascuna recensione all'anno di riferimento	Vettore *year_vec*
Costruzione dataframe temporale per distribuzione dei topic	*topics_merged, year_vec*	Creazione del dataframe recensione-topic–anno (senza testo recensione)	Dataframe *df_base_time*
Rimozione topic di rumore	*df_base_time*	Filtraggio delle osservazioni con Topic >= 0	Dataframe senza documenti di rumore
Calcolo distribuzione temporale	*df_base_time*	Conteggio recensioni per topic × year e calcolo della percentuale sul totale del topic	Dataframe *topic_time_distribution_merged*
Esportazione distribuzione temporale	*topic_time_distribution_merged*	Salvataggio risultati in formato RDS e XLSX	File di output distribuzione temporale
Costruzione dataframe base lessicale	*testi_clean, topics_merged, year_vec*	Creazione dataframe documenti–topic–anno (con testo recensione)	Dataframe *df_base*
Rimozione topic di rumore	*df_base*	Filtraggio Topic >= 0	Dataframe senza rumore
Tokenizzazione e pulizia testi	*df_base* senza rumore	Tokenizzazione, filtro lunghezza token, rimozione caratteri non alfabetici e stopword standard	Dataframe *df_tokens*
Normalizzazione morfologica	*df_tokens*	Lemmatizzazione e a seguito stemming dei token	Daframe *df_tokens* normalizzato
Rimozione stopword di dominio	*df_tokens* normalizzato, *domain_stopwords*	Eliminazione manuale di termini di dominio	Dataframe *df_tokens* finale
Conteggio parole per periodo	*df_tokens*	Conteggio delle occorrenze dei termini per ciascuna combinazione Topic-Year	Dataframe *df_ctfidf_time*
Filtro frequenza minima	*df_ctfidf_time*	Rimozione termini con frequenza assoluta inferiore alla soglia minima in ciascun Topic-Year	Dataframe *df_ctfidf_time* filtrato
Calcolo TF temporale	*df_ctfidf_time* filtrato	Calcolo del Term Frequency normalizzata all'interno di ciascuna combinazione Topic-Year	Dataframe *df_ctfidf_time* con colonna TF
Calcolo IDF temporale	Dataframe *df_ctfidf_time* con colonna TF	Calcolo IDF sulla diffusione dei termini tra le diverse combinazioni Topic-Year	Dataframe *df_ctfidf_time* con colonna IDF
Calcolo c-TF-IDF temporale	Dataframe *df_ctfidf_time* con valori TF e IDF	Calcolo della misura c-TF-IDF come prodotto tra TF e IDF per ciascuna combinazione Topic–Year	Dataframe con colonna c-TF-IDF temporale
Selezione parole rilevanti per periodo	Dataframe con colonna c-TF-IDF temporale	Selezione delle *n* parole con c-TF-IDF più elevato per Topic × Year	Dataframe *period_words*
Associazione etichette semantiche	*period_words, topic_labels_final*	Merge con etichette manuali dei topic	Dataframe *topic_words_over_time* etichettato
Esportazione risultati finali	*topic_words_over_time*	Scrittura file CSV	*topic_words_over_time.csv*

In conclusione, la figura 11 rappresenta il diagramma di flusso dell'analisi dinamica.

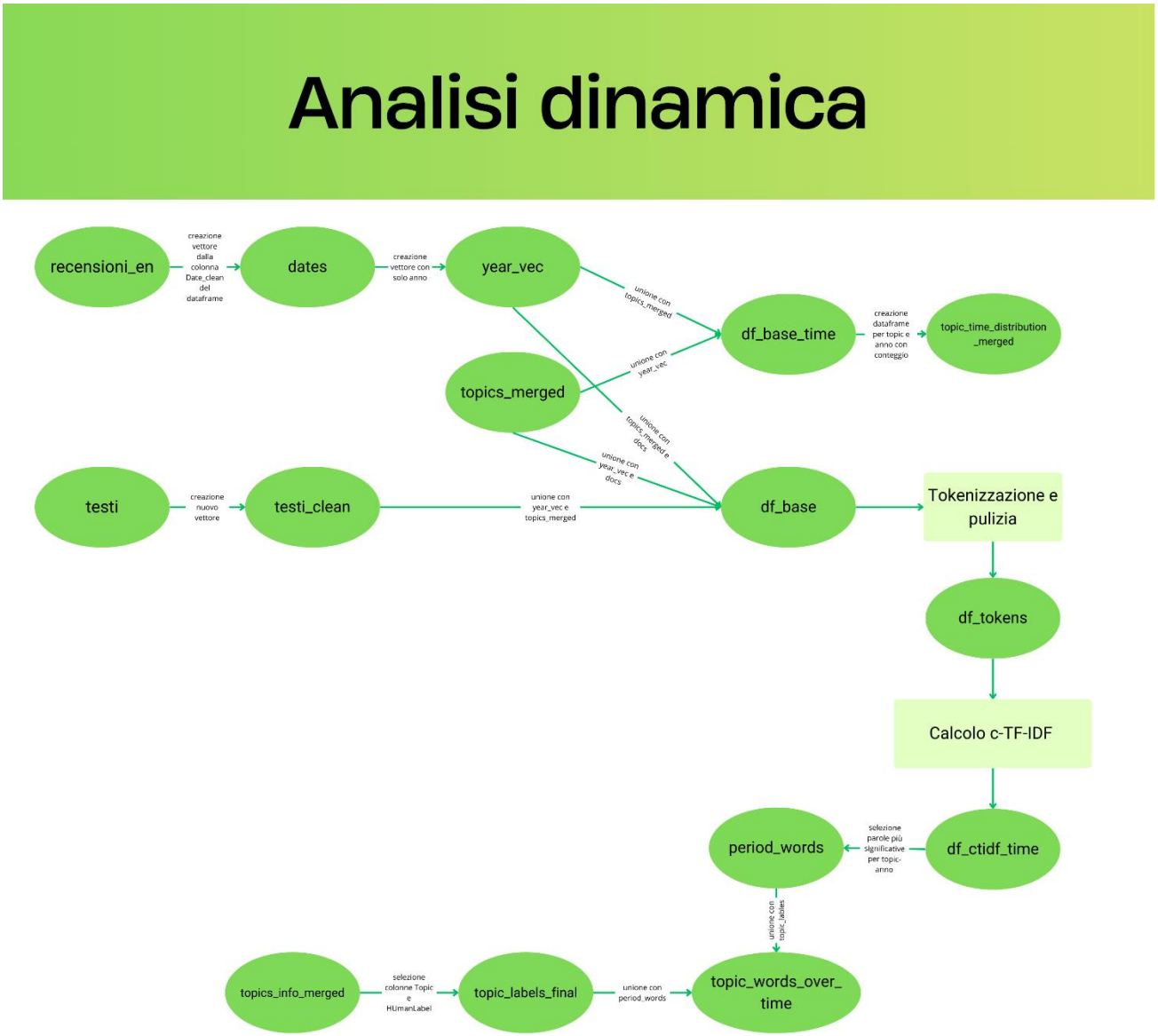

Figura 11 - Diagramma di flusso dell'analisi dinamica ex-post

4.5.1. Validazione supervisionata dell'analisi dinamica

La validazione qualitativa viene estesa alla dimensione temporale dell'analisi. In particolare, per ciascun topic e per ciascun periodo temporale considerato, bisogna esaminare le parole dinamiche emergenti, ovvero i termini che aumentano di rilevanza in specifici intervalli temporali. Tale analisi ha l'obiettivo di verificare la plausibilità semantica dell'evoluzione lessicale dei topic nel tempo, valutando se l'emergere o il declino di specifiche parole è coerente con cambiamenti funzionali o eventi contestuali rilevanti.

Se l'analisi evidenzia che le parole emergenti per ciascun periodo risultano semanticamente coerenti con il topic di riferimento e con il contesto temporale in cui compaiono, tale evidenza può essere interpretata come una forma di validazione qualitativa dell'analisi temporale, suggerendo che la dinamica osservata non sia frutto di rumore, ma rifletta un'evoluzione significativa delle tematiche discusse dagli utenti.

Nel complesso, la validazione qualitativa supervisionata mira a confermare l'interpretabilità semantica dei topic statici, la coerenza delle recensioni associate ai cluster, la plausibilità dell'evoluzione temporale del contenuto lessicale. In caso contrario, è necessario iterare il processo di modellazione, ricalibrando le principali scelte metodologiche (preprocessing, numerosità dei topic, criteri di selezione delle parole), fino al raggiungimento di una soluzione interpretabile e coerente.

L'eventuale riscontro di evidenze quali la stabilità delle parole rappresentative nel tempo, la coerenza semantica delle parole emergenti e la coerenza tematica delle recensioni associate ai topic fornisce un ulteriore supporto ai risultati della validazione quantitativa, rafforzando la fiducia nell'affidabilità complessiva del modello e delle analisi derivate.

La Figura 12 sintetizza l'intera procedura di ricerca, offrendo una visione d'insieme del flusso di analisi. Rispetto ai dettagliati flowchart presentati nelle sezioni precedenti per ogni singola fase, questa rappresentazione fornisce un quadro logico immediato.

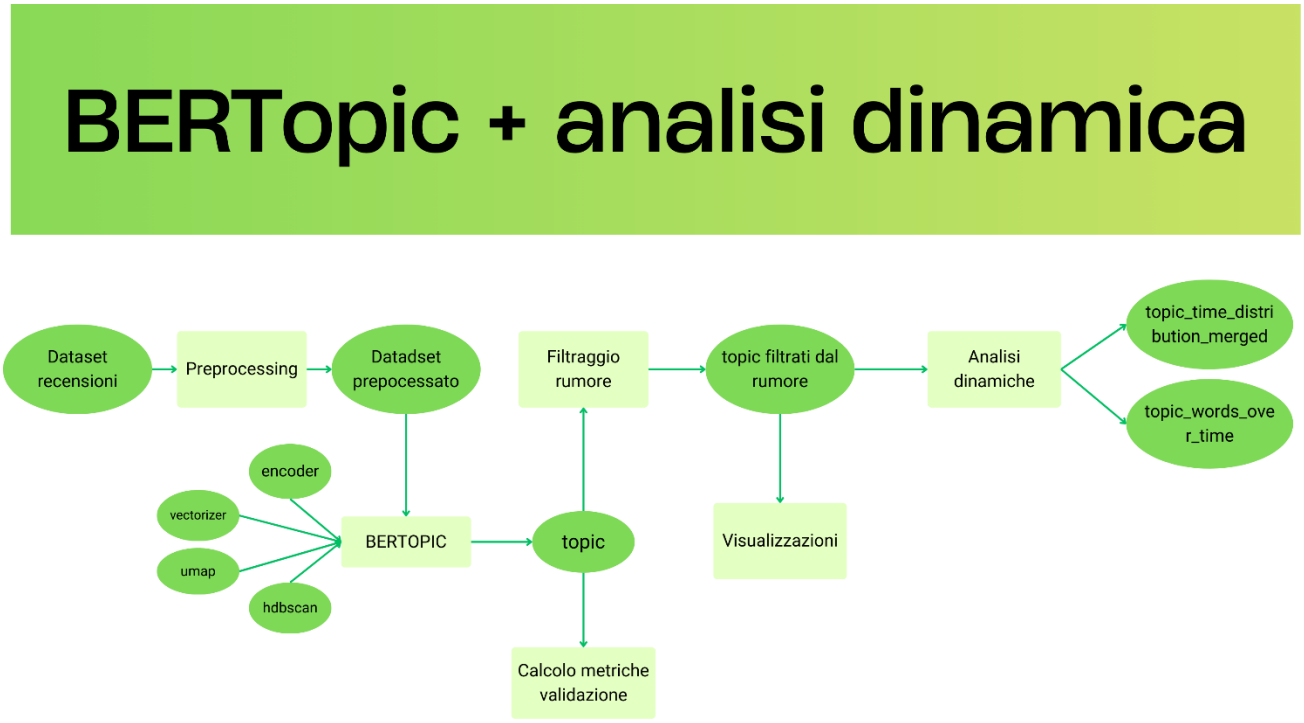

Figura 12 - Flowchart generale degli step del processo

5. Applicazione del modello alle recensioni della piattaforma Spotify

Dopo aver descritto nel capitolo precedente la metodologia di ricerca e lo script implementato per l'analisi, tale procedura viene ora applicata alle recensioni online della piattaforma Spotify, oggetto del caso di studio analizzato in questa tesi. I dati sono stati tratti dal dataset pubblico 'Spotify Reviews'[2], disponibile sulla piattaforma Kaggle.

5.1. Descrizione del dataset

Il dataset utilizzato per l'analisi è stato reperito dalla piattaforma di open data Kaggle. In particolare, è stato selezionato un dataset contenente recensioni online degli utenti relative all'applicazione Spotify, corredate da informazioni temporali necessarie per l'analisi dell'evoluzione dei topic nel tempo. I dati sono forniti in formato CSV, facilitando le operazioni di importazione, gestione e pre-processing.

Al fine di fornire una caratterizzazione preliminare del dataset dal punto di vista temporale, è stata condotta un'analisi del volume delle recensioni nel tempo sul campione iniziale, prima dell'applicazione di qualsiasi filtro o procedura di pre-processing. In particolare, il numero di recensioni per ciascun anno è stato calcolato aggregando le osservazioni in base all'anno di pubblicazione della recensione. La Figura 13 mostra la distribuzione annuale delle recensioni relative alla piattaforma Spotify, dalla quale emerge un incremento complessivo del numero di recensioni nel corso degli anni, con una crescita più marcata nel periodo recente.

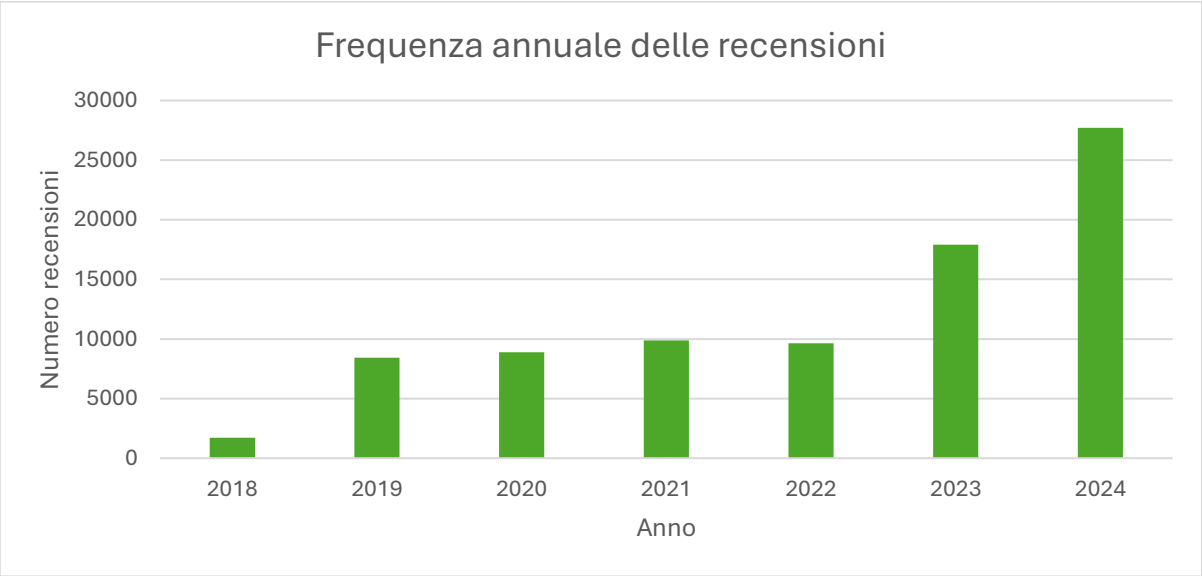

Figura 13 - Frequenza delle recensioni per anno nel dataset di Spotify

[2] https://www.kaggle.com/datasets/ashishkumarak/spotify-reviews-playstore-daily-update

Un altro elemento interessante riguarda la lunghezza delle recensioni, illustrata in Figura 14. Il grafico mostra la distribuzione cumulata delle recensioni in funzione della loro lunghezza, misurata in termini di numero di parole. Sull'asse delle ascisse è riportata la lunghezza delle recensioni, mentre sull'asse delle ordinate principale è indicato il numero di recensioni per ciascun intervallo. Sull'asse delle ordinate secondario è riportata la percentuale cumulata di recensioni. Tale rappresentazione consente di osservare la quota di recensioni al crescere della lunghezza testuale. Dall'analisi emerge che la maggior parte delle recensioni sia concentrata tra le 40 e le 80 parole, mentre il contributo delle recensioni più lunghe risulta progressivamente marginale, come quello delle recensioni particolarmente brevi.

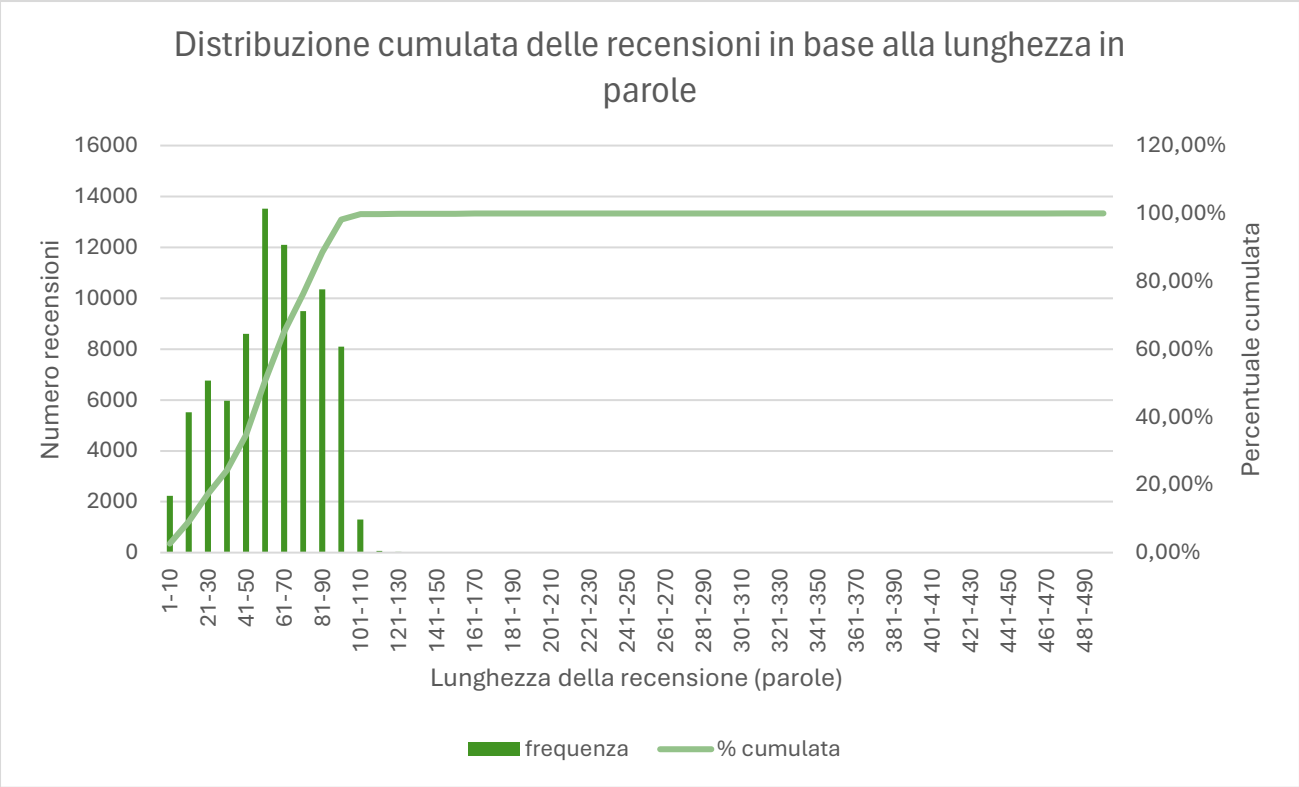

Figura 14 - Distribuzione cumulata delle recensioni in base alla lunghezza conteggiata in parole nel dataset di Spotify

Oltre al testo della recensione, il dataset include diversi metadati informativi, tra cui la valutazione numerica assegnata dall'utente, la data di pubblicazione e la versione dell'app al momento della scrittura della recensione. In particolare, la dimensione temporale risulta rilevante per contestualizzare l'emergere di specifici temi e potenziali criticità legate ad aggiornamenti o modifiche funzionali. Le informazioni relative alla versione dell'applicazione, invece, sono state escluse dall'analisi, in quanto non è stato possibile reperire uno storico completo delle versioni con le funzionalità introdotte in ciascuna release. Altri metadati, quali l'identificativo della recensione e il nome utente, risultano privi di contenuto informativo rilevante ai fini dello studio.

5.2. Applicazione del modello al dataset

A partire dal dataset iniziale così descritto, è stata quindi avviata la fase di costruzione del corpus di documenti analizzato. Il campione iniziale è costituito da 84.163 recensioni. In una prima fase, è stato applicato un filtro sulla lingua, selezionando esclusivamente le recensioni in lingua inglese, che ha ridotto il campione a 83.896 osservazioni.

Successivamente, al fine di garantire un contenuto testuale sufficientemente informativo per l'analisi, sono state eliminate le recensioni composte da meno di cinque parole, ottenendo un campione pari a 83.319 recensioni. La soglia di cinque parole è stata adottata in quanto rappresenta un compromesso tra l'eliminazione di testi eccessivamente brevi, tipicamente limitati a valutazioni generiche, e la conservazione di recensioni che, seppur brevi, presentano un minimo di contesto descrittivo utile all'identificazione dei temi.

Il testo è stato quindi sottoposto a una fase di pulizia e normalizzazione tramite la funzione *clean_text*. Durante la fase di tokenizzazione, sono stati eliminati i token di lunghezza pari o inferiore a due caratteri, in quanto generalmente costituiti da interiezioni o residui non informativi, che non contribuiscono in modo significativo alla rappresentazione semantica del testo. A seguito di tale operazione, è stato applicato un ulteriore filtro sulle recensioni con lunghezza inferiore a cinque parole. Questo passaggio si è reso necessario poiché le operazioni di pulizia e normalizzazione possono ridurre la lunghezza effettiva del testo, trasformando alcune recensioni inizialmente idonee in documenti troppo brevi per essere considerati informativi ai fini dell'analisi. Il dataset finale risulta pertanto composto da 83.202 recensioni.

Prima dell'applicazione del modello BERTopic al corpus di recensioni, sono stati definiti i principali parametri relativi alle fasi di embedding, riduzione dimensionale, clustering e rappresentazione dei topic, al fine di adattare il modello alle caratteristiche del dataset analizzato.

Per la rappresentazione semantica dei testi viene impiegato il framework *Sentence-Transformers*, selezionando un modello di embedding pre-addestrato appartenente alla famiglia Sentence-BERT (*all-MiniLM-L6-v2*). L'utilizzo di Sentence-BERT è consolidato in letteratura per la rappresentazione semantica di unità testuali di dimensione ridotta, quali frasi o brevi segmenti di testo. In particolare, il modello Sentence-BERT è stato progettato per generare embedding semanticamente informativi a livello di frase ed è stato validato su testi brevi (Reimers and Gurevych, 2019). Il modello *all-MiniLM-L6-v2* è scelto per il buon compromesso tra qualità delle rappresentazioni semantiche e complessità

computazionale, nonché per la sua ampia diffusione in applicazioni di Topic Modeling basate su embedding (Ajinaja et al., 2025; Grootendorst, 2022).

Infine, per questa fase di generazione degli embeddings si è scelta una dimensione del batch pari a 128. Tale parametro definisce il numero di documenti elaborati simultaneamente dal modello Transformer durante ogni passo dell'inferenza. La scelta di questo valore rappresenta un compromesso efficiente tra prestazioni computazionali e stabilità nella generazione delle rappresentazioni vettoriali.

La riduzione dimensionale tramite UMAP è stata configurata impostando *n_neighbors* = 10, *n_components* = 5, *min_dist* = 0.1 e utilizzando la metrica *cosine*. Nel contesto delle recensioni online, caratterizzate da testi brevi, la scelta di un numero contenuto di vicini consente di preservare la granularità dei temi emergenti, mantenendo al contempo una struttura sufficientemente stabile per il successivo clustering. Il numero di componenti è stato fissato a cinque al fine di ridurre la dimensionalità degli embeddings mantenendo una rappresentazione sufficientemente ricca delle relazioni semantiche tra le recensioni, evitando una compressione eccessiva dello spazio informativo. Il parametro *min_dist* è stato invece impostato a 0.1 per favorire una moderata compattezza delle rappresentazioni nello spazio ridotto, facilitando l'individuazione di cluster semanticamente coerenti senza imporre una separazione eccessiva tra i documenti.

Per la fase di clustering è stato adottato l'algoritmo HDBSCAN, configurato con *min_cluster_size* = 200 e *min_samples* = 25. Il valore selezionato per la dimensione minima del cluster consente di privilegiare la stabilità semantica dei topic, evitando la formazione di cluster troppo piccoli o scarsamente rappresentativi, mentre il parametro *min_samples* contribuisce a rafforzare la robustezza dei cluster individuati, riducendo l'inclusione di documenti semanticamente ambigui.

La rappresentazione testuale dei topic è stata ottenuta mediante il CountVectorizer. Il parametro *min_df* = 0,005 è stato introdotto per escludere termini estremamente rari, presenti in una quota trascurabile dei documenti, lo 0,5%, che tendono a introdurre rumore e non contribuiscono in modo significativo alla definizione dei topic. Tale soglia consente di mantenere termini sufficientemente informativi, preservando al contempo la varietà lessicale del corpus. Il parametro *max_df* = 0,85 è stato utilizzato per filtrare termini eccessivamente frequenti, che compaiono nell'85% dei documenti e risultano quindi poco discriminanti dal punto di vista tematico. L'esclusione di queste parole permette di focalizzare l'analisi su termini maggiormente caratterizzanti i singoli topic. Sono stati considerati sia unigrammi sia bigrammi, al fine di catturare non solo concetti espressi da singole parole, ma anche espressioni multi-parola semanticamente rilevanti, particolarmente frequenti nel contesto delle

recensioni online (ad esempio funzionalità dell'app, come lo *smart shuffle*, giudizi sintetici o riferimenti a specifiche caratteristiche del servizio). Infine, sono state eliminate le parole appartenenti a una lista di stopword in lingua inglese, opportunamente estesa con termini specifici del dominio (*Spotify*), con l'obiettivo di ridurre ulteriormente il rumore lessicale e migliorare la qualità interpretativa dei topic estratti.

Coerentemente con le scelte effettuate nella fase di clustering, il parametro *min_topic_size* del modello BERTopic è stato impostato pari a 200. Tale impostazione consente di garantire che ciascun topic finale sia supportato da un numero minimo adeguato di documenti, allineando la definizione dei topic alla dimensione minima dei cluster individuati da HDBSCAN e contribuendo a migliorare la stabilità e l'interpretabilità dei temi estratti. Infine, per ciascun topic sono state estratte le 5 parole più rappresentative (*top_n_words = 5*), al fine di fornire una sintesi concisa e facilmente interpretabile del contenuto semantico di ciascun tema.

A valle dell'applicazione di questo modello così settato, sono stati individuati complessivamente 14 topic distinti, compreso il cluster di rumore, al quale sono state assegnate il 13,63% delle recensioni.

Nella fase di filtraggio del rumore, i documenti inizialmente classificati come outlier sono stati riassegnati a un topic solo qualora la probabilità di appartenenza risultasse maggiore o uguale a 0,2, compromesso tra l'esigenza di ridurre il rumore e quella di preservare informazioni potenzialmente rilevanti. Mentre l'approccio standard spesso prevede l'assegnazione forzata di tutti i documenti (soglia pari a 0) o, al contrario, il mantenimento di un'alta quota di rumore non classificato, la scelta di una soglia di 0,2 è stata guidata dalla natura delle recensioni online. In tali contesti, una soglia di 0,2 garantisce che il documento abbia un legame semantico non casuale con il topic, senza però escludere recensioni che, pur nella loro brevità, risultano chiaramente riconducibili a un tema specifico. Soglie più elevate avrebbero infatti comportato una riassegnazione eccessivamente restrittiva, lasciando nel rumore documenti semanticamente coerenti con i topic, mentre soglie più basse avrebbero aumentato il rischio di assegnazioni deboli e poco interpretabili. La scelta della soglia è stata inoltre supportata da un confronto empirico tra la percentuale di documenti classificati come rumore prima e dopo il filtraggio (11,49%), che ha permesso di verificare il recupero di documenti borderline mantenendo un buon livello di coerenza semantica dei topic.

5.3. Labeling e merging dei topic

Per effettuare il labeling dei topic, sono state ispezionate le recensioni maggiormente rappresentative di ciascun topic, individuate sulla base del più elevato grado di similarità rispetto alla tematica del cluster. L'analisi congiunta delle parole chiave e dei contenuti testuali più affini ha consentito di attribuire a ciascun topic un'etichetta descrittiva, sintetizzandone il significato semantico prevalente, nonché di identificare ed unificare topic caratterizzati da contenuti semanticamente sovrapposti.

Il topic a cui afferiscono il maggior numero di recensioni è quello denominato *"*Limitazioni della versione free*"*. Le opinioni riconducibili a questo parlano delle restrizioni imposte alla versione gratuita dell'applicazione. In particolare, emerge l'impossibilità di selezionare liberamente i brani, l'obbligo di ascolto in modalità *shuffle*, e successivamente in modalità *smart shuffle*, e il numero limitato di skip consentiti. Un elemento ricorrente riguarda inoltre l'introduzione progressiva di funzionalità considerate "di base" dietro paywall.

Il secondo topic per numerosità di recensioni è quello denominato "Affidabilità e la stabilità dell'app" che include segnalazioni di malfunzionamenti e bug. Le recensioni riconducibili a questo tema evidenziano la presenza di crash improvvisi, rallentamenti dell'interfaccia, blocchi nella riproduzione dei brani e anomalie nella coda di riproduzione, soprattutto a seguito di aggiornamenti dell'applicazione.

Un ulteriore topic rilevante riguarda la "Riproduzione di podcast", che rappresenta uno dei servizi della piattaforma. Le recensioni riconducibili a questo tema si concentrano principalmente sull'esperienza di ascolto dei contenuti audio, con particolare riferimento alla gestione degli episodi dei podcast, alla continuità della riproduzione e alla facilità di accesso ai contenuti salvati.

Durante l'analisi, è emerso anche un topic a cui è stata assegnata l'etichetta "Motivazione del rating", all'interno del quale gli utenti non si limitano a descrivere situazioni e funzionalità, ma spiegano direttamente le ragioni che li hanno portati ad assegnare un determinato punteggio all'applicazione. In molte recensioni, il numero di stelle diventa esso stesso oggetto del discorso. Questo topic assume una natura trasversale, poiché integra elementi riconducibili ad altri temi, come le limitazioni della versione gratuita, la presenza di annunci pubblicitari, l'accesso ai lyrics o problemi di stabilità, riletti però in chiave valutativa. Nel complesso, il topic evidenzia come il rating non rappresenti soltanto una misura sintetica di soddisfazione, ma anche un mezzo attraverso cui gli utenti articolano un giudizio argomentato sul valore percepito del servizio. Questo aspetto risulta particolarmente rilevante

nell'ambito della Digital Voice of Customer, poiché consente di comprendere non solo cosa viene criticato o apprezzato, ma come tali elementi influenzino la valutazione finale dell'esperienza.

Un topic semanticamente vicino al precedente è rappresentato da "Aggiornamenti dell'app". Anche in questo caso, le recensioni contengono frequenti riferimenti espliciti al numero di rating assegnato, come si vede da alcune parole chiave associate al topic ("zero star" e "star"). Tuttavia, l'attenzione degli utenti non è rivolta primariamente al giudizio complessivo, bensì alle modifiche introdotte nelle versioni più recenti dell'applicazione. Le recensioni associate a questo topic sono caratterizzate da un forte confronto temporale tra la versione attuale dell'app e quelle precedenti. In questo caso, il rating emerge come una conseguenza diretta delle modifiche percepite, più che come oggetto centrale della narrazione. Per tale motivo, nonostante la parziale sovrapposizione lessicale con il topic "Motivazione del rating", i due temi sono mantenuti distinti: il primo cattura una dimensione prevalentemente valutativa e trasversale, mentre il topic sugli aggiornamenti dell'app riflette specificamente sull'evoluzione del prodotto nel tempo.

Invece, due temi inizialmente distinti che poi sono stati sovrapposti sono riconducibili alla funzione di visualizzazione dei testi delle canzoni, i *lyrics*. In particolare, i topic identificati inizialmente come separati presentavano un'elevata affinità sia in termini di parole chiave sia, soprattutto, di contenuto testuale delle recensioni maggiormente rappresentative. Le opinioni associate a entrambi i topic fanno riferimento alla progressiva limitazione dell'accesso ai *lyrics*, evidenziando questioni quali la mancata visualizzazione dei testi, l'introduzione di un limite mensile di consultazione e la necessità di sottoscrivere un abbonamento premium per usufruire della funzione. Tali elementi risultano frequentemente intrecciati a segnalazioni di bug e malfunzionamenti tecnici dei *lyrics*. Alla luce di questa forte interdipendenza semantica, i due topic sono stati unificati in un unico tema denominato *"*Accesso ai lyrics*"*.

Accanto ai topic principali descritti, l'analisi ha individuato ulteriori temi caratterizzati da una numerosità inferiore di recensioni, ma comunque rilevanti per delineare un quadro complessivo dell'esperienza utente. Tra questi rientrano, ad esempio, le opinioni relative alla presenza di annunci, alle modifiche dell'interfaccia utente e alla disponibilità di specifiche funzionalità accessorie, come il widget o gli audiolibri. Sono inoltre emersi topic legati a specifici contesti d'uso o segmenti geografici, come la disponibilità del catalogo musicale in determinati paesi o lingue. Sebbene tali temi coinvolgano una quota più ridotta di utenti, la loro individuazione contribuisce a evidenziare la varietà delle dimensioni attraverso cui si articola la Digital Voice of Customer della piattaforma Spotify.

5.4. Visualizzazioni del modello statico

Al fine di analizzare e interpretare i risultati del modello BERTopic, sono state adottate diverse visualizzazioni esplorative, volte a supportare la comprensione della struttura tematica del corpus, della distribuzione delle recensioni tra i topic e delle relazioni semantiche tra i temi individuati.

In primo luogo, è stata analizzata la distribuzione delle recensioni tra i topic finali, a seguito del merging, visibile in Figura 15. Questa visualizzazione ha il fine di valutare il peso relativo di ciascun tema all'interno del corpus e di identificare i topic maggiormente rappresentativi.

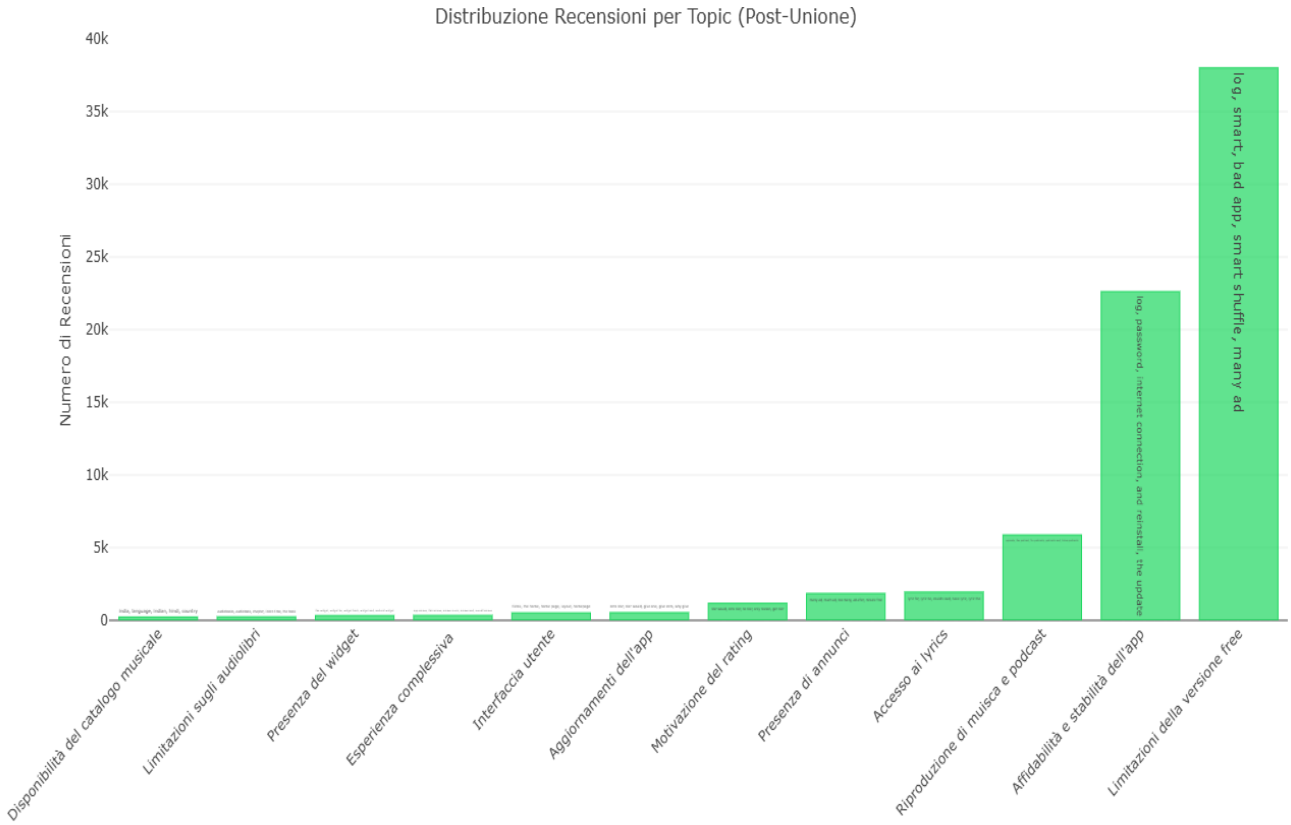

Figura 15 - Distribuzione delle recensioni per topic post-merge nel dataset di Spotify

Successivamente, è stata utilizzata una rappresentazione spaziale dei topic per analizzare la struttura semantica complessiva del modello. La mappa dei topic, riportata in Figura 16, consente di visualizzare le relazioni di prossimità tra i temi, evidenziando cluster semanticamente affini e aree di maggiore separazione tematica.

Da tale rappresentazione emerge la presenza di gruppi di topic relativamente vicini tra loro nello spazio semantico, suggerendo l'esistenza di macroaree tematiche condivise. Tuttavia, la prossimità spaziale non è stata considerata come criterio sufficiente per procedere automaticamente al merging dei topic.

L’analisi delle recensioni maggiormente rappresentative ha infatti mostrato che, nonostante la vicinanza semantica, molti di questi temi mantengono specificità contenutistiche distinte e riflettono dimensioni diverse dell’esperienza utente.

Per questo motivo, il merging è stato applicato in modo selettivo e conservativo, limitandosi ai soli casi in cui l’elevata affinità semantica risultava confermata anche a livello qualitativo. In particolare, solo due topic sono stati effettivamente unificati, come visto nel paragrafo precedente, mentre gli altri cluster visivamente individuabili nella mappa sono stati mantenuti separati al fine di preservare la capacità del modello di cogliere sfumature tematiche rilevanti.

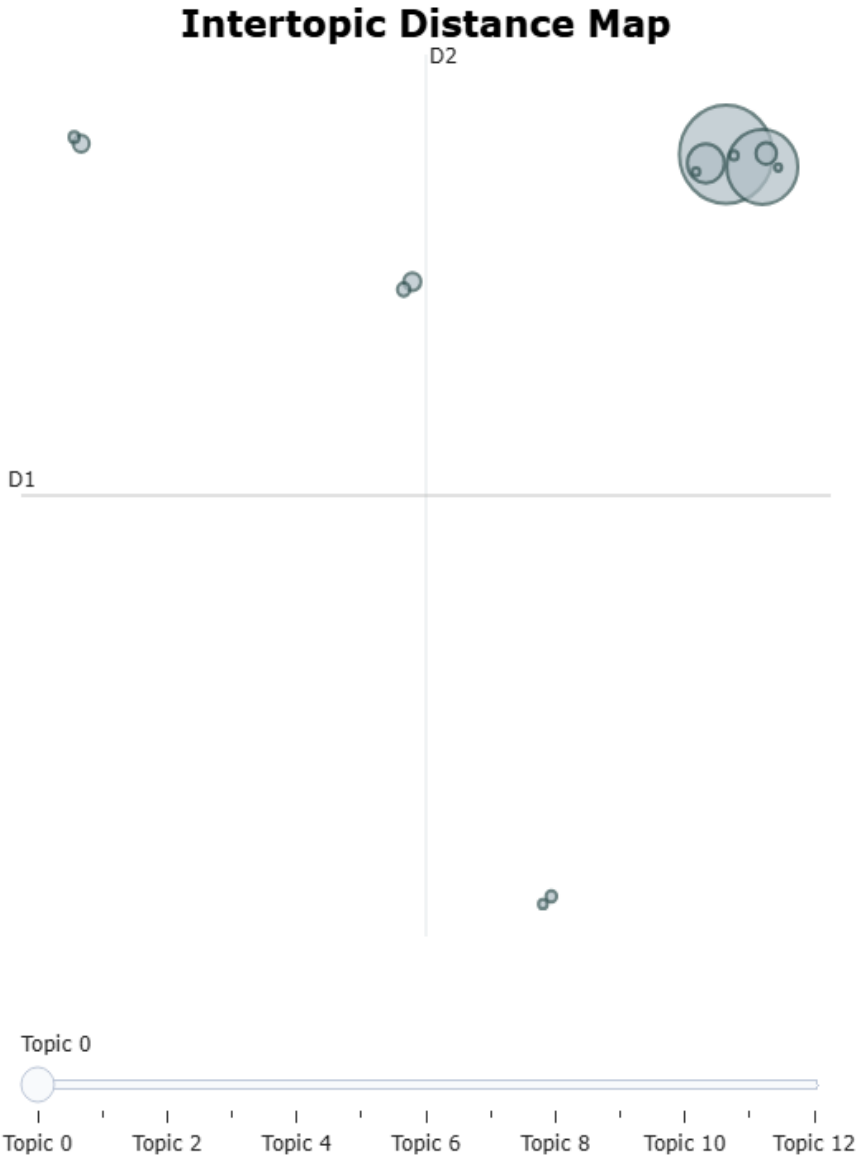

Figura 16 - Intertopic Distance Map del dataset di Spotify

A completamento dell'analisi, è stata esaminata una matrice di similarità tra i topic, in Figura 17, che consente di rappresentare le relazioni semantiche tra i topic individuati in forma quantitativa. Tale visualizzazione fornisce una lettura complementare rispetto alla mappa delle distanze inter-topic, permettendo di quantificare il grado di affinità semantica tra le coppie di topic.

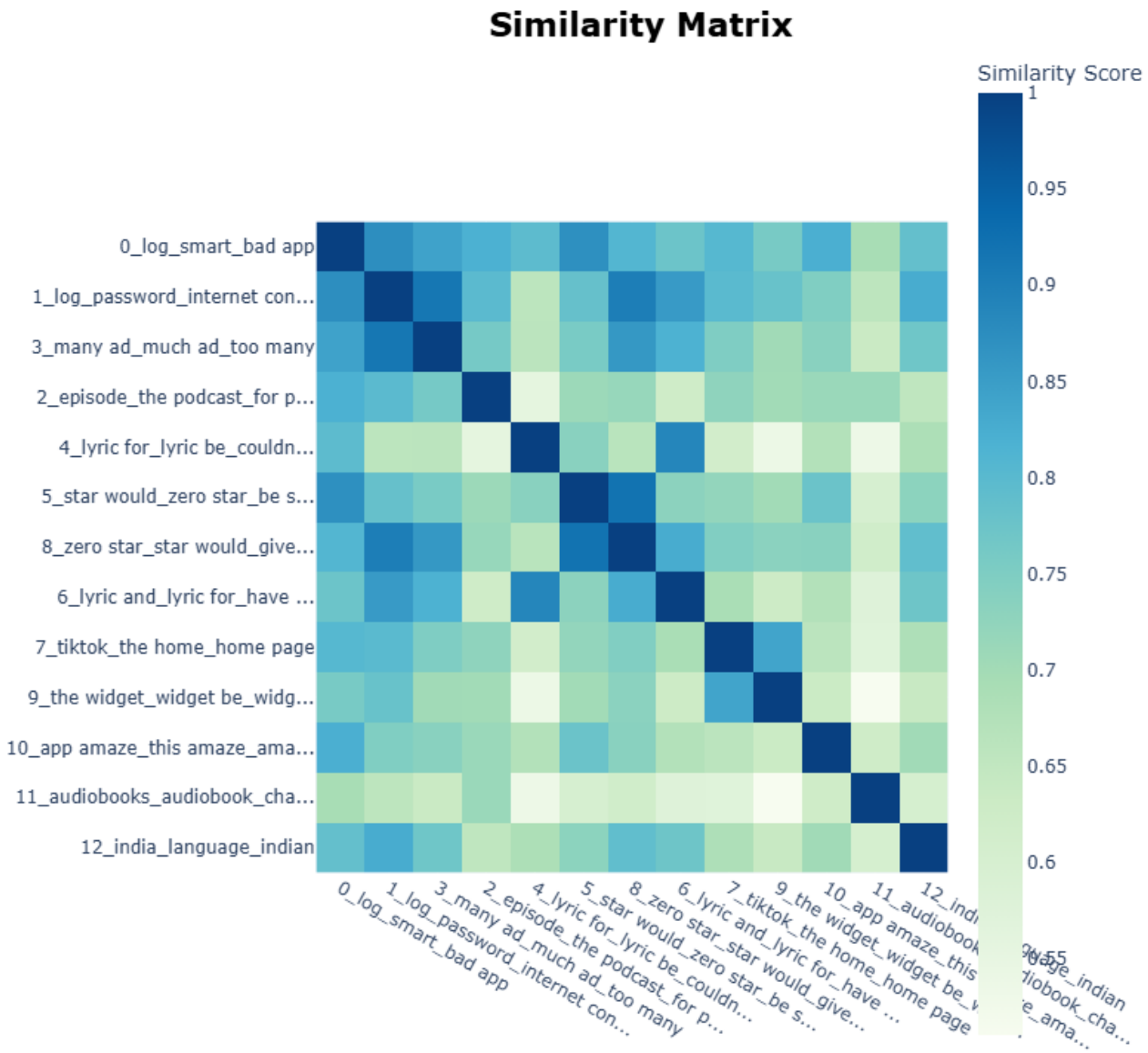

Figura 17 - Similarity Matrix del dataset di Spotify

Infine, per analizzare la composizione interna e il contenuto lessicale dei singoli temi, è stata generata la visualizzazione riportata in Figura 18. Questi grafici a barre orizzontali illustrano per ogni topic le cinque parole chiave con il punteggio c-TF-IDF più elevato.

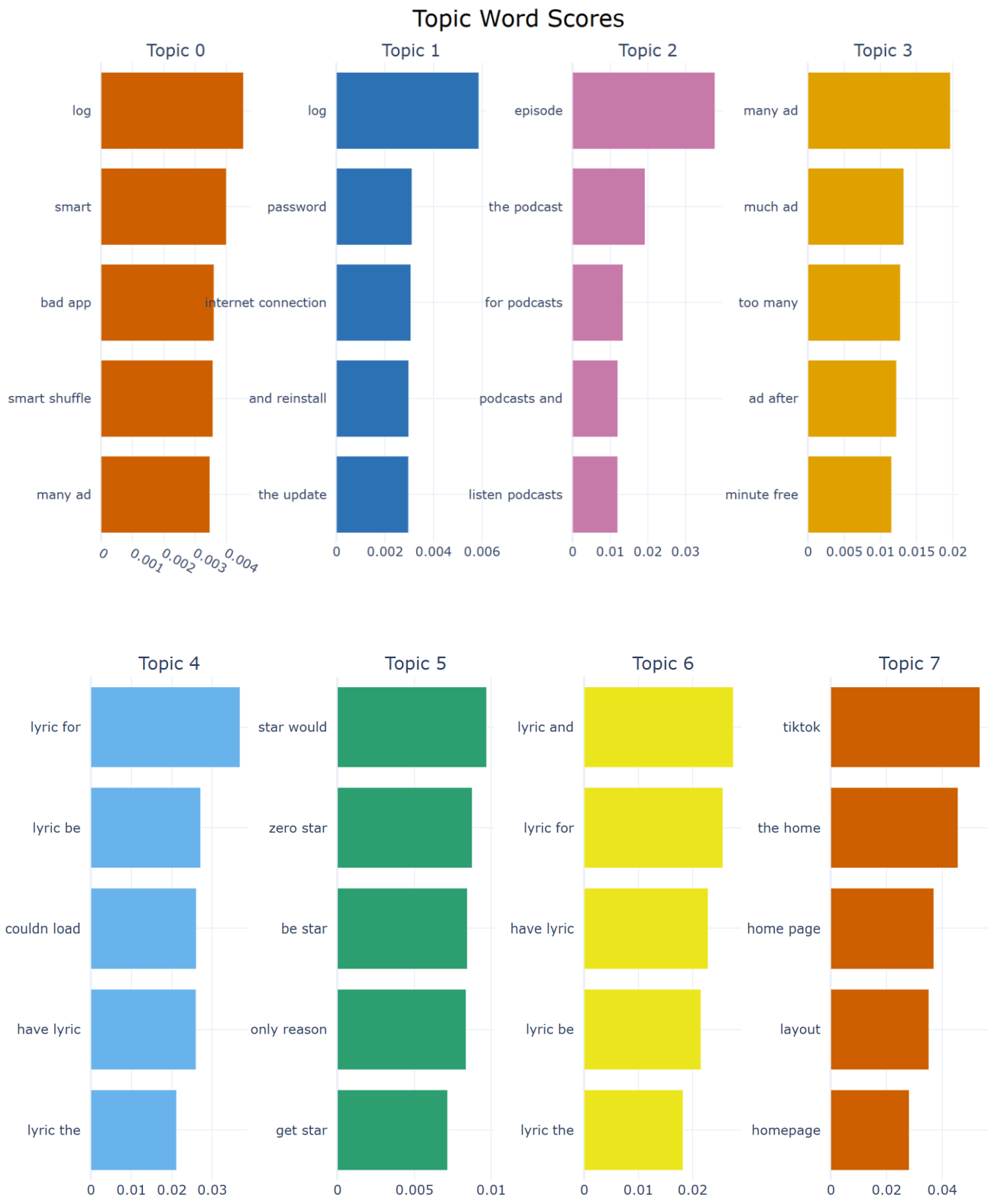
Topic Word Scores
Topic 0
log
smart
bad app
smart shuffle
many ad
0
0.001
0.002
0.003
0.004
Topic 1
log
password
internet connection
and reinstall
the update
0
0.002
0.004
0.006
Topic 2
episode
the podcast
for podcasts
podcasts and
listen podcasts
0
0.01
0.02
0.03
Topic 3
many ad
much ad
too many
ad after
minute free
0
0.005
0.01
0.015
0.02
Topic 4
lyric for
lyric be
couldn load
have lyric
lyric the
0
0.01
0.02
0.03
Topic 5
star would
zero star
be star
only reason
get star
0
0.005
0.01
Topic 6
lyric and
lyric for
have lyric
lyric be
lyric the
0
0.01
0.02
Topic 7
tiktok
the home
home page
layout
homepage
0
0.02
0.04

Figura 18 - Punteggi c-TF-IDF per le parole chiave dei topic

5.5. Metriche di validazione

Nella fase di validazione del modello per il dataset di Spotify, il valore di topic coherence complessivo ottenuto è stato pari a *U_mass* = – 2.01, indicando una buona coerenza semantica globale dei topic individuati. Tale risultato è da considerarsi pienamente soddisfacente alla luce della natura del dataset analizzato, costituito da recensioni brevi, non strutturate e intrinsecamente rumorose. Sebbene la metrica *U_mass* non ammetta soglie assolute, il valore ottenuto risulta coerente con quanto osservato in letteratura: modelli ritenuti validi presentano valori di *U_mass* anche inferiori, suggerendo che il risultato ottenuto è compatibile con una buona qualità complessiva del modello (Rahimi et al., 2024).

Il risultato quantitativo è coerente con le evidenze emerse dalla precedente analisi qualitativa: l'ispezione delle recensioni ad alta probabilità di appartenenza ai topic e il processo di labeling manuale hanno confermato la presenza di nuclei semantici ben definiti e interpretabili, rafforzando l'affidabilità del punteggio di coerenza ottenuto.

Il valore di topic diversity ottenuto è pari a 0.577. Ciò significa che oltre il 57% del lessico rappresentativo è specifico dei singoli topic. Il valore di topic diversity ottenuto può essere considerato soddisfacente a fronte di un vocabolario parzialmente condiviso tipico del dominio applicativo e se confrontato con quanto riportato in letteratura, dove vengono ritenuti accettabili anche valori inferiori (Dieng et al., 2020). Valori più elevati di topic diversity sono tipicamente osservati in dataset composti da documenti semanticamente più distinti tra loro, come nel caso di analisi di domini tecnologici (Jin et al., 2025). Nel presente studio, il corpus analizzato riguarda un'unica applicazione e un contesto tematico relativamente omogeneo, condizione che tende naturalmente a ridurre la diversità lessicale tra i topic e rende quindi non realistico il confronto diretto con tali benchmark più elevati.

Inoltre, questo risultato è coerente anche con le visualizzazioni esplorative del modello: la parziale sovrapposizione osservata nella *Intertopic Distance Map* non compromette la validità della segmentazione tematica, ma riflette piuttosto la comune appartenenza dei topic a un macro-tema. Al contempo, l'elevato livello di diversità terminologica garantisce che ciascun topic catturi sfumature lessicali e questioni specifiche.

Nel complesso, i risultati ottenuti dalla validazione non supervisionata indicano che il modello di Topic Modeling adottato presenta un buon equilibrio tra coerenza semantica e capacità discriminante, fornendo una base solida per le analisi temporali e interpretative sviluppate nel proseguo del lavoro.

5.6. Applicazione dell'analisi dinamica ex-post

L'analisi dinamica *ex-post* è stata applicata ai topic statici ottenuti nella fase precedente con l'obiettivo di ricostruirne l'evoluzione temporale sia in termini quantitativi, sia semantici.

Una prima scelta metodologica ha riguardato il livello di aggregazione temporale: le date sono state trasformate in anno, in quanto tale granularità è risultata la più adeguata rispetto alle caratteristiche del dataset. Prove preliminari condotte su finestre più ristrette (trimestri e mesi) hanno evidenziato una forte instabilità delle distribuzioni, soprattutto per i topic meno frequenti, nei quali il numero di recensioni per periodo diventava troppo esiguo per consentire interpretazioni affidabili. Inoltre, un'aggregazione eccessivamente fine tendeva a frammentare artificialmente fenomeni di medio periodo, rendendo poco visibili sia i trend strutturali sia gli shock. L'aggregazione annuale ha quindi rappresentato un compromesso tra capacità di cogliere variazioni significative e necessità di garantire una base dati sufficientemente robusta per il calcolo delle metriche e delle parole chiave dinamiche.

Per quanto riguarda l'analisi semantica dinamica, è stata effettuata una tokenizzazione *ex-post* dei testi, indipendente dal pre-processing utilizzato in BERTopic, al fine di identificare le parole più rappresentative di ogni topic in ciascun periodo. In questa fase sono stati introdotti alcuni parametri chiave, la cui scelta è stata guidata dall'obiettivo di ridurre il rumore linguistico senza compromettere la capacità descrittiva del vocabolario.

In particolare, è stata adottata una soglia minima di lunghezza dei token pari a quattro caratteri, per escludere termini molto brevi che, nelle recensioni online, risultano spesso poco informativi o ambigui, riducendo la presenza di parole funzionali non rilevanti per la caratterizzazione semantica. Sono state rimosse sia le stopword generiche sia un insieme di domain stopwords specifiche del contesto, come *spotify*, *playlist*, *song*, *listen*, *pause* e *log,* poiché termini eccessivamente frequenti e trasversali a tutti i topic, che avrebbero ridotto il potere discriminante dell'analisi c-TF-IDF. Infine, è stata introdotta una soglia minima di frequenza ($n \geq 3$) per considerare una parola come rilevante in un dato periodo, scelta per evitare che termini occasionali o errori ortografici influenzassero le parole chiave dinamiche.

Tali parametri non rappresentano valori assoluti, ma configurazioni ragionevoli rispetto alla dimensione e alla natura del dataset analizzato: in contesti con volumi maggiori o linguaggi più specialistici, soglie diverse potrebbero risultare più appropriate.

Le parole caratterizzanti sono state individuate tramite una variante del c-TF-IDF calcolata su base topic–anno, in modo da privilegiare i termini distintivi di uno specifico periodo rispetto all'intero

corpus. Il numero di parole estratte per ciascuna coppia temporale è stato fissato a sei, ma rappresenta un ulteriore parametro modificabile in funzione del livello di dettaglio desiderato.

5.7. Risultati dell'analisi dinamica

Per approfondire la comprensione della Digital Voice of Customer, è stata condotta l'analisi dinamica dei topic, volta a esaminare come la numerosità delle recensioni e il vocabolario associato a ciascun tema si siano evoluti nel tempo. I risultati mostrano che alcuni topic evidenziano pattern stabili nel tempo, altri manifestano picchi improvvisi o cambiamenti nella composizione del vocabolario. Questo riflette sia l'introduzione di nuove funzionalità sia modifiche di policy o di interfaccia. Di seguito vengono illustrati i principali topic per numerosità e quelli più interessanti in termini di cambiamenti dinamici, con un'analisi congiunta della loro evoluzione quantitativa e semantica.

Nel corso del tempo il topic 0, associato alle "Limitazioni della versione free", rimane abbastanza stabile, come mostrato in Figura 19. Il topic mostra un'evoluzione semantica coerente con i principali cambiamenti funzionali del servizio. Nei primi anni, dal 2018 al 2021, le parole chiave più frequenti sono legate a restrizioni operative della versione gratuita, come *skip*, *random* e *minute*, suggerendo una percezione stabile e costante delle limitazioni imposte agli utenti non premium.

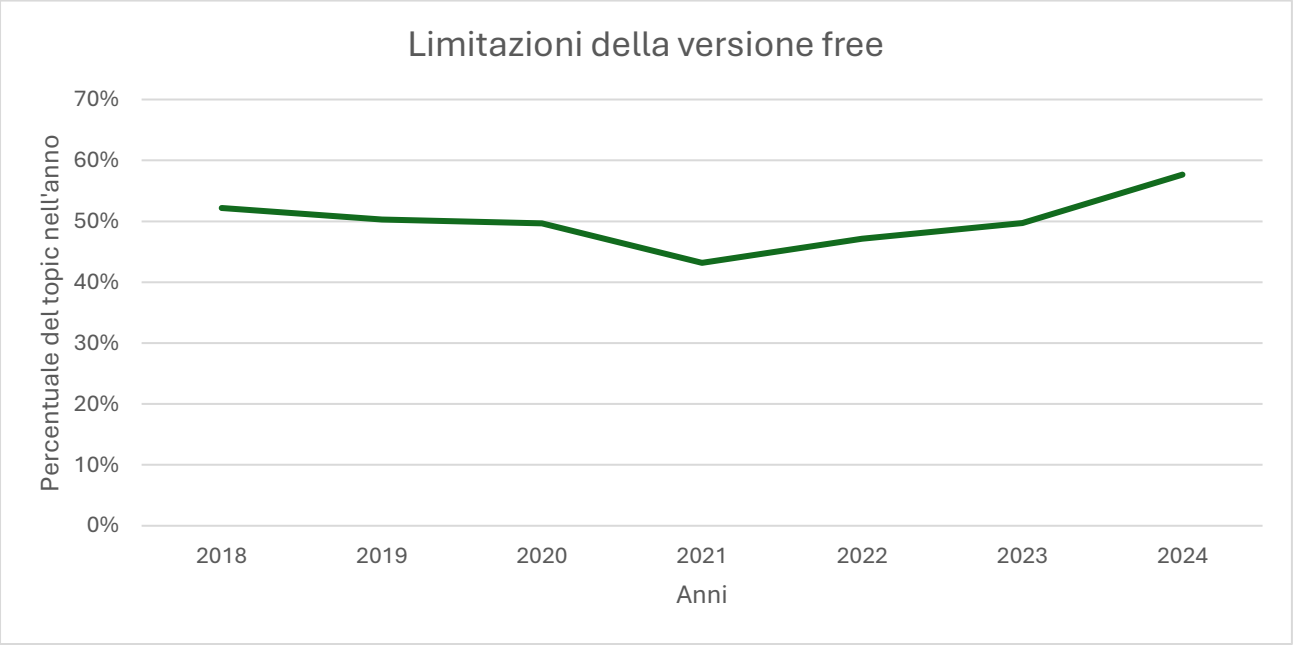

Figura 19 - Percentuale di recensioni appartenenti al topic "Limitazioni della versione free" negli anni

Nel 2022, in concomitanza con la ripresa dell'importanza del topic, dopo la leggera flessione registrata nel 2021, emerge un cambiamento nel vocabolario del topic: compare il termine *shuffle,* come evidenziato in Figura 20. Nel 2023 entra nel vocabolario del topic anche la parola *smart.* Questi risultati

sono coerenti con l'introduzione, prima, dell'opzione *shuffle* obbligatoria per utenti free nel 2022 e, successivamente, della funzione *smart shuffle* sempre obbligatoria per utenti free nel marzo 2023. Tali modifiche hanno influenzato il comportamento della riproduzione per gli utenti free, generando un aumento delle recensioni legate alla percezione di controllo limitato. Nel 2024 il tema rimane centrale con il 58% delle recensioni associate a questo topic e il vocabolario che si stabilizza attorno a *shuffle*, *premium*, e *smart*, confermando che le limitazioni della versione gratuita continuano a rappresentare una delle principali fonti di discussione per gli utenti.

Topic	Year	period_words	HumanLabel
0	2018	offlin, skip, random, crash, connect, phone	Limitazioni della versione free
0	2019	save, phone, connect, updat, librari, download	Limitazioni della versione free
0	2020	skip, minut, random, phone, download, connect	Limitazioni della versione free
0	2021	skip, minut, connect, random, load, phone	Limitazioni della versione free
0	2022	skip, random, shuffl, paus, play, phone	Limitazioni della versione free
0	2023	shuffl, premium, skip, smart, updat, random	Limitazioni della versione free
0	2024	shuffl, premium, skip, smart, offlin, random	Limitazioni della versione free

Figura 20 - Parole rappresentative del topic "Limitazioni della versione free" per anno

Il topic 1, etichettato come "Affidabilità e stabilità dell'app", rappresenta una quota rilevante delle recensioni annuali, sebbene la sua incidenza percentuale sul totale diminuisca progressivamente nel tempo, come mostrato in Figura 21: dal 39% nel 2018 al 24% nel 2024. Ciò suggerisce che, pur aumentando in valore assoluto, il tema perde centralità relativa rispetto ad altri argomenti emergenti.

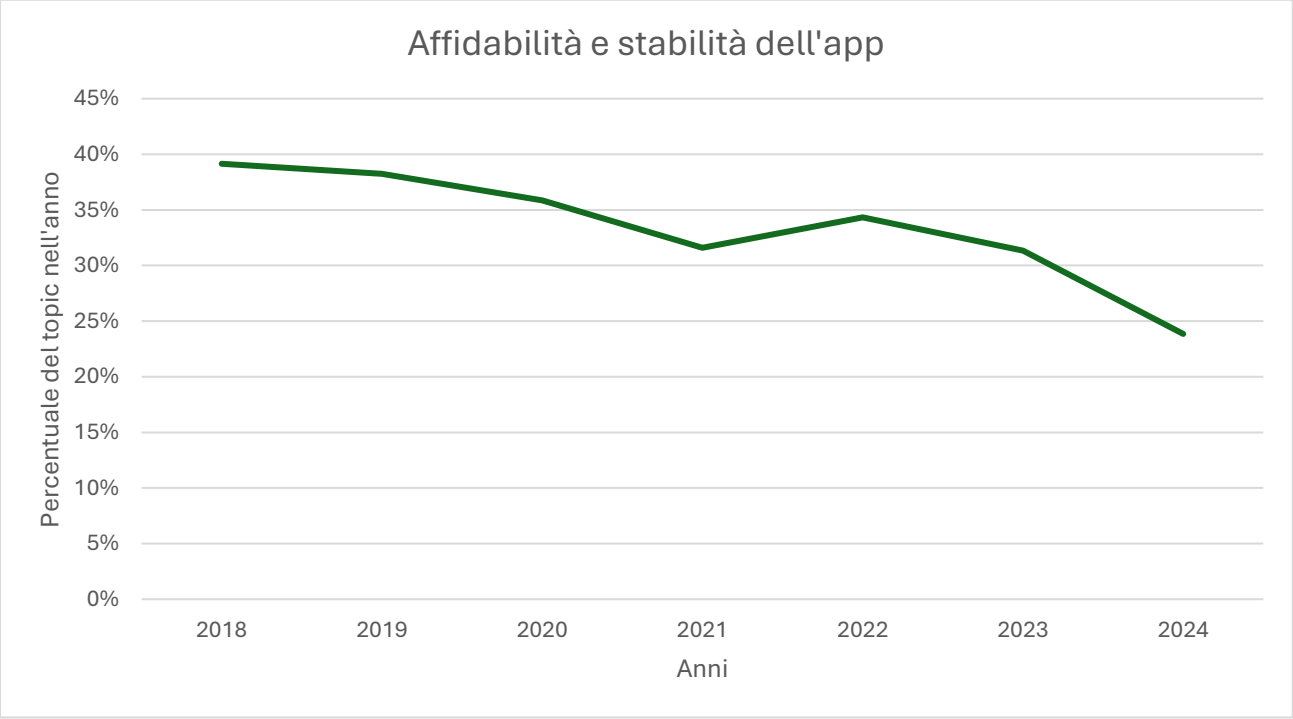

Figura 21 - Percentuale di recensioni appartenenti al topic "Affidabilità e stabilità dell'app" negli anni

L’analisi dell’evoluzione delle parole chiave conferma la natura tecnica del topic e ne rafforza l’interpretazione. A livello semantico, il topic rimane abbastanza stabile nel tempo. Nei primi anni il vocabolario è dominato da termini come *crash*, *connect*, *phone* e *offline*, ai quali, nel 2020–2021, si aggiungono parole come *freeze*, *reinstall* e *load*. A partire dal 2023 il topic mostra un cambiamento semantico parzialmente sovrapposto a quello del topic sulle limitazioni della versione free, con l’ingresso di termini quali *smart*, *shuffle* e *skip*, insieme a *uninstall* o *download*. Questo risultato è coerente con aggiornamenti introdotti nel periodo, come quello che ha implementato la funzione *smart shuffle*, che, oltre a incidere sull’esperienza di ascolto, sembra aver generato problemi di stabilità e compatibilità su alcuni dispositivi. In Figura 22 è possibile visualizzare le parole chiave del topic per ogni periodo.

Topic	Year	period_words	HumanLabel
1	2018	offlin, connect, crash, phone, premium, shuffl	Affidabilità e stabilità dell'app
1	2019	phone, save, connect, updat, download, librari	Affidabilità e stabilità dell'app
1	2020	phone, connect, reinstal, account, freez, download	Affidabilità e stabilità dell'app
1	2021	connect, phone, internet, load, download, freez	Affidabilità e stabilità dell'app
1	2022	phone, connect, skip, random, updat, download	Affidabilità e stabilità dell'app
1	2023	shuffl, premium, updat, skip, smart, uninstal	Affidabilità e stabilità dell'app
1	2024	premium, shuffl, offlin, skip, smart, download	Affidabilità e stabilità dell'app

Figura 22 - Parole rappresentative del topic “Affidabilità e stabilità dell’app” per anno

Il topic 2 “Riproduzione di podcast” registra un incremento particolarmente rilevante nel 2021, quando raggiunge il valore massimo di 1.544 recensioni e il 18% delle recensioni annuali totali. Questo picco, ben visibile in Figura 23, è coerente sia con il crescente interesse del pubblico nel format dei podcast sia con l’acquisizione dell’esclusiva del podcast *The Joe Rogan Experience* nel settembre 2020.

Figura 23 - Percentuale di recensioni appartenenti al topic "Riproduzione di podcast" negli anni

L'analisi delle parole chiave nei diversi periodi, in Figura 24, rafforza questa interpretazione: nei primi anni il topic è dominato da termini legati alla fruizione dei podcast, come *episode*, *download*, *pause* e *offline* indicando un uso ancora emergente del formato. Nel 2020 compaiono parole come *exclusive* che riflettono l'introduzione di contenuti esclusivi, come il podcast *The Joe Rogan Experience* e nel 2021 il termine *rogan* entra nel vocabolario del topic. Nel 2022 si osserva una contrazione significativa del numero di recensioni associate al topic (–38%), accompagnata da un vocabolario più neutro (*pause*, *stop*, *restart*), indicativo di una fase di assestamento dell'offerta podcast. Negli anni successivi il topic continua ad assistere alla riduzione progressiva del suo peso relativo sul totale delle recensioni, segnalando una normalizzazione del formato podcast rispetto ad altri temi emergenti. Parallelamente, nel 2024 emergono nuove parole chiave come *audiobook*, *audio* e *video*, che indicano un'ulteriore evoluzione dell'esperienza di ascolto verso contenuti multiformato e confermano il ruolo del topic come indicatore sensibile dei cambiamenti strategici della piattaforma.

Topic	Year	period_words	HumanLabel
2	2018	podcast, episod, download, paus, librari, offlin	Riproduzione di podcast
2	2019	podcast, episod, load, paus, download, phone	Riproduzione di podcast
2	2020	podcast, episod, download, exclus, crash, restart	Riproduzione di podcast
2	2021	podcast, episod, rogan, download, paus, freez	Riproduzione di podcast
2	2022	podcast, episod, paus, stop, restart, download	Riproduzione di podcast
2	2023	podcast, episod, random, content, rogan, start	Riproduzione di podcast
2	2024	podcast, episod, audiobook, audio, rogan, video	Riproduzione di podcast

Figura 24 - Parole rappresentative del topic "Riproduzione di podcast" per anno

Il topic 4, relativo all'"Accesso ai lyrics", mostra un andamento crescente nel periodo analizzato, passando da un numero marginale di recensioni a 1.243 recensioni nel 2024. Dopo una prima fase di crescita graduale tra il 2019 e il 2022, in cui il topic rimane residuale (circa l'1% delle recensioni annuali), si osserva una brusca accelerazione nel 2023, quando il numero di recensioni aumenta di quasi sei volte rispetto all'anno precedente, seguita da un ulteriore incremento nel 2024. Questo cambiamento, visibile in Figura 25, rappresenta uno dei salti più marcati tra tutti i topic analizzati e suggerisce l'effetto di un evento rilevante.

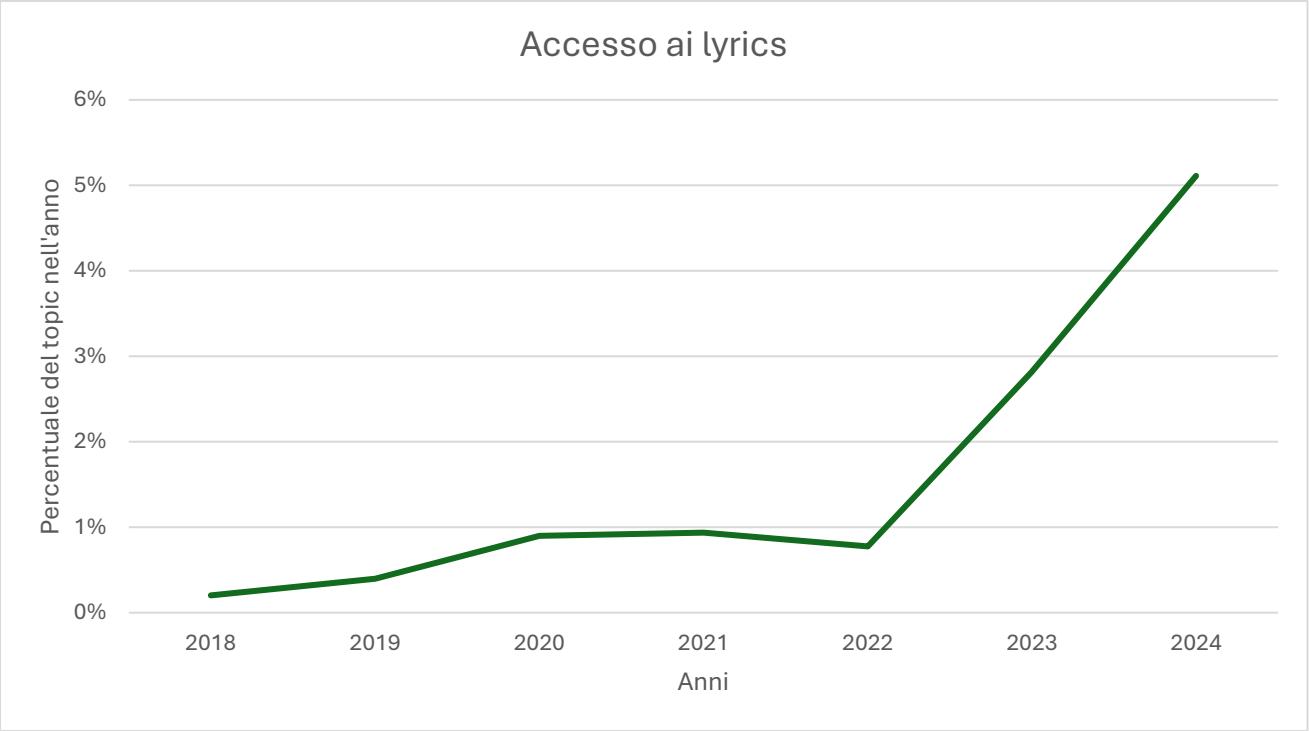

Figura 25 - Percentuale di recensioni appartenenti al topic "Accesso ai lyrics" negli anni

L'analisi delle parole chiave, in Figura 26, spiega il motivo di questo forte incremento. Nei primi anni il vocabolario del topic è legato alla semplice fruizione dei testi (*lyric*, *read*, *sing*), indicando un interesse funzionale ma limitato. In questi anni compaiono tra le parole Genius e Musixmatch, che riflettono le recensioni in cui gli utenti fanno riferimento alle piattaforme utilizzate da Spotify per fornire i testi: Genius per la funzione *Behind the Lyrics*, introdotta tra il 2019 e il 2020, e Musixmatch per i testi sincronizzati completi, reintrodotti a partire dal 2022. Nel 2023, in concomitanza con l'impennata del numero di recensioni associate al topic, il vocabolario si arricchisce di termini come *premium* e *update*. Questo cambiamento coincide con l'implementazione dell'aggiornamento che introduce il limite di accesso ai lyrics per gli utenti free, che ha trasformato una funzionalità precedentemente percepita come standard in un elemento di restrizione, generando un forte aumento delle recensioni. Nel 2024 il topic si consolida ulteriormente, con parole chiave come *limit* e *month*, a

indicare una stabilizzazione della questione relativa al numero limitato di lyrics disponibili mensilmente per gli utenti free.

Topic	Year	period_words	HumanLabel
4	2018	lyric, music	Accesso ai lyrics
4	2019	lyric, genius, stori, sing, request, read	Accesso ai lyrics
4	2020	lyric, genius, featur, incorpor, hope, sing	Accesso ai lyrics
4	2021	lyric, featur, read, account, appl, download	Accesso ai lyrics
4	2022	lyric, translat, english, musixmatch, dialog, sync	Accesso ai lyrics
4	2023	lyric, load, couldn, updat, sing, premium	Accesso ai lyrics
4	2024	lyric, limit, load, premium, month, couldn	Accesso ai lyrics

Figura 26 - Parole rappresentative del topic "Accesso ai lyrics" per anno

Il topic relativo all'*Interfaccia utente,* dopo una presenza marginale tra il 2018 e il 2022, con un numero di recensioni sempre molto limitato, registra un aumento eccezionale nel 2023, quando le recensioni aumentano del 911%, per poi ridursi drasticamente nel 2024, come si vede in Figura 27. Questo pattern suggerisce una reazione fortemente concentrata nel tempo da parte degli utenti, tipica dei cambiamenti visivi e di interazione che impattano immediatamente sull'esperienza d'uso.

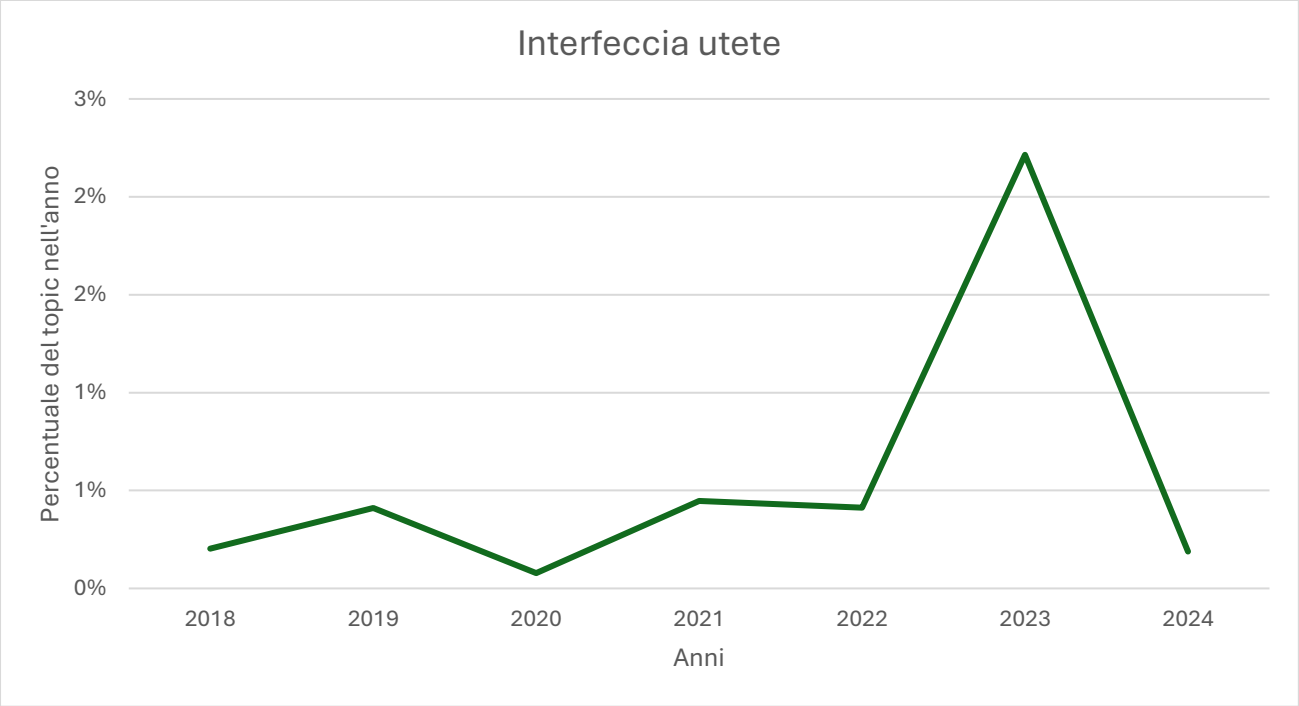

Figura 27 - Percentuale di recensioni appartenenti al topic "Interfaccia utente" negli anni

L'analisi delle parole chiave nei diversi periodi conferma questa interpretazione: nei primi anni il vocabolario del topic è dominato da termini come *home*, *layout*, *page* e *screen*. La discontinuità semantica emerge nel 2023, con la comparsa del termine *TikTok* che segnala chiaramente l'introduzione di un'interfaccia ispirata allo stile "feed verticale" tipico del social media. Questo

redesign ha generato una forte reazione da parte degli utenti, come suggerito dall'esplosione del numero di recensioni e dall'aumento della significatività del topic nello stesso anno. Nel 2024 il topic torna rapidamente a livelli marginali, mentre il vocabolario mantiene alcuni riferimenti al layout di *TikTok*, come è possibile vedere in Figura 28.

Topic	Year	period_words	HumanLabel
7	2018	home, page, button, screen, chang	Interfaccia utente
7	2019	layout, home, updat, confus, strang, previous	Interfaccia utente
7	2020	layout, notif, home, terribl, recent, artist	Interfaccia utente
7	2021	home, layout, page, screen, homepag, scroll	Interfaccia utente
7	2022	home, layout, page, blend, screen, scroll	Interfaccia utente
7	2023	tiktok, home, scroll, screen, homepag, page	Interfaccia utente
7	2024	tiktok, scroll, homepag, tile, page, layout	Interfaccia utente

Figura 28 - Parole rappresentative del topic "Interfaccia utente" per anno

Il topic "Presenza del widget" evidenzia una dinamica fortemente legata a un evento specifico avvenuto nel 2019. In questo anno il topic registra un picco improvviso, come visibile in Figura 29, arrivando a rappresentare più del 3% delle recensioni annuali. Tale esplosione è direttamente riconducibile alla decisione di Spotify di rimuovere il widget per Android. Infatti, nel 2019 il vocabolario è dominato da termini come *widget*, *remove*, *android* e *bring*, che esprimono esplicitamente la richiesta di ripristino della funzionalità. Negli anni successivi il numero di recensioni legate al tema torna su livelli molto bassi e il vocabolario riflette principalmente commenti su funzionalità correlate al widget, come mostrato in Figura 30.

Figura 29 - Percentuale di recensioni appartenenti al topic "Presenza del widget" negli anni

Topic	Year	period_words	HumanLabel
9	2018	widget, buggi, screen, phone, stop, doesn	Presenza del widget
9	2019	widget, remov, android, screen, bring, home	Presenza del widget
9	2020	widget, remov, updat, random, recent, star	Presenza del widget
9	2021	widget, homescreen, notif, control, button, screen	Presenza del widget
9	2022	widget, bluetooth, remov, close, custom, connect	Presenza del widget
9	2023	widget, colour, homescreen, ugli, remov, set	Presenza del widget
9	2024	widget, break, android, updat, auto, home	Presenza del widget

Figura 30 - Parole rappresentative del topic "Presenza del widget" per anno

Il topic 11 riguardante le "Limitazioni sugli audiolibri" rappresenta un tema emergente che compare solo a partire dal 2020, in concomitanza con l'introduzione degli audiolibri all'interno della piattaforma. Nei primi anni in cui è presente il topic rimane marginale, come da Figura 31, con pochissime recensioni e un peso trascurabile sul totale annuale, segno di una fase iniziale di adozione del nuovo formato. La situazione cambia nel 2023, quando viene introdotto il limite di 15 ore mensili di ascolto per gli audiolibri. Il limite viene introdotto a fine 2023 ed è per questo che le recensioni sono ancora di più nel 2024 e la parola *limit* compare solo dal 2024, come è possibile notare in Figura 32.

Figura 31 - Percentuale di recensioni appartenenti al topic "Limitazione sugli audiolibri" negli anni

Topic	Year	period_words	HumanLabel
11	2020	chapter, audiobook, book, play, music	Limitazioni sugli audiolibri
11	2021	book, chapter, audio, stop	Limitazioni sugli audiolibri
11	2022	audiobook, book, purchas, audio, subscript, star	Limitazioni sugli audiolibri
11	2023	audiobook, book, hour, audio, chapter, audibl	Limitazioni sugli audiolibri
11	2024	audiobook, book, chapter, hour, audio, limit	Limitazioni sugli audiolibri

Figura 32 - Parole rappresentative del topic "Limitazione sugli audiolibri" per anno

Oltre ai topic analizzati in questo capitolo, l'applicazione del modello ha fatto emergere ulteriori topic caratterizzati da una bassa variazione temporale o da una scarsa significatività statistica. Tali cluster, non presentando mutamenti qualitativi rilevanti ai fini della presente ricerca, sono stati esclusi dalla discussione dettagliata per favorire la focalizzazione sui trend maggiormente impattanti.

Per una visione esaustiva delle distribuzioni di frequenza delle parole chiave dei rimanenti topic, si rimanda alle tabelle e ai grafici riportati nell'Appendice B.

6. Discussione dei risultati potenziali e implicazioni manageriali

6.1. Dal monitoraggio reattivo alla diagnostica proattiva

L'analisi dinamica dei topic emergenti da fonti testuali offre al management uno strumento prezioso per monitorare l'evoluzione della Digital Voice of Customer. Come dimostrato in letteratura (Tirunillai and Tellis, 2014), i contenuti generati dagli utenti (UGC) costituiscono una fonte di informazioni ad alta frequenza temporale capace di estrarre le dimensioni latenti della qualità e della soddisfazione dei consumatori. Lo studio sottolinea come un'analisi dinamica permetta di mappare le percezioni degli utenti nei confronti del brand nel tempo.

L'efficacia di tali analisi dipende però dalla capacità del modello di superare la semplice estrazione di parole chiave per cogliere la complessità del ragionamento. A differenza delle tecniche tradizionali basate sulla frequenza dei termini, l'approccio qui adottato comprende la semantica latente delle recensioni (Grootendorst, 2022). Grazie all'uso degli embeddings, il modello è in grado di riconoscere la correlazione tra esperienze diverse anche quando espresse con vocabolari differenti (ad esempio, associando concettualmente termini come *crash*, *freeze* o *bug*). Questo aspetto è particolarmente rilevante in contesti applicativi come quello della Digital VoC, dove gli utenti descrivono problemi analoghi con registri linguistici molto diversi.

L'analisi temporale dei topic consente di identificare picchi improvvisi nella discussione, spesso riconducibili a modifiche di prodotto, aggiornamenti di policy o eventi esterni. Le variazioni del linguaggio degli utenti sono collegate a queste specifiche decisioni aziendali. Tali variazioni rappresentano segnali deboli che, se intercettati tempestivamente, permettono all'organizzazione di attivare meccanismi di risposta proattiva prima che l'insoddisfazione si traduca in conseguenze economiche o reputazionali.

Dal punto di vista manageriale, tale approccio assume valore soprattutto come sistema diagnostico capace di integrare i limiti di indicatori tradizionali come il Net Promoter Score (NPS) (Reichheld, 2003). Quest'ultimo, infatti, agisce come segnale ritardato che comunica che un problema esiste senza spiegarne l'origine. L'analisi dinamica dei topic, individuati tramite BERTopic, fornisce la dimensione qualitativa necessaria per comprendere il perché del cambiamento. Se un calo dell'NPS indica una diminuzione della fedeltà, lo studio dei topic permette di risalire alla causa specifica, ad esempio, un malfunzionamento tecnico o una policy di prezzo sgradita, trasformando un numero astratto in un insight azionabile.

6.2 Esempi di discontinuità e trend

Si precisa che l'analisi del dataset di Spotify non costituisce l'obiettivo finale della tesi, ma rappresenta un caso dimostrativo volto a validare la metodologia proposta. I risultati qui presentati hanno quindi la funzione di mostrare come lo script sviluppato consenta di estrarre, monitorare e interpretare la Digital Voice of Customer in modo replicabile, indipendentemente dal dominio applicativo e dalla piattaforma considerata (Tirunillai and Tellis, 2014).

Un esempio emblematico emerso dall'analisi riguarda il topic relativo alla "Riproduzione di podcast", il cui andamento mostra come l'introduzione di contenuti esclusivi generi variazioni sia quantitative sia semantiche nella VoC. Analogamente, il topic "Accesso ai lyrics" evidenzia come modifiche nelle policy possano produrre reazioni rapide e consistenti, confermando che l'analisi tematica è in grado di cogliere discontinuità legate a decisioni strategiche della piattaforma.

L'analisi congiunta dei trend quantitativi e dell'evoluzione delle parole chiave consente inoltre di interpretare gli effetti delle dinamiche osservate nel medio periodo. Emblematico è il caso del topic "Affidabilità dell'app", che evidenzia una progressiva diminuzione del volume di discussione. Tale andamento può essere letto come un segnale di miglioramento della stabilità percepita del servizio e di una conseguente riduzione delle segnalazioni da parte degli utenti.

In modo analogo, il topic "Interfaccia utente" mostra come interventi di design o aggiornamenti funzionali possano generare reazioni intense. La riduzione della quota di recensioni associate a questo tema nell'anno successivo suggerisce che le modifiche introdotte siano state successivamente ricalibrate, verosimilmente in risposta al feedback espresso dagli utenti, evidenziando il ruolo delle recensioni come meccanismo informativo a supporto dei processi decisionali aziendali.

L'introduzione di nuovi format o funzionalità, come nel caso delle "Limitazioni sugli audiolibri", ha fatto emergere un topic precedentemente assente, fenomeno coerente con l'idea che i dati testuali possano rivelare bisogni latenti e nuove dimensioni dell'esperienza utente. Ciò conferma che l'analisi tematica non si limita a descrivere l'esistente, ma costituisce uno strumento di esplorazione per orientare l'evoluzione dell'offerta (Rahimi et al., 2024).

6.3 Visualizzazioni per il decision-making

Un elemento centrale del contributo applicativo di questo lavoro riguarda l'usabilità dello script sviluppato. Le visualizzazioni prodotte non hanno solo finalità descrittive, ma costituiscono strumenti operativi per le aziende. In particolare, l'output *topic_time_distribution_merged* consente di individuare picchi anomali e periodi di discontinuità nel volume delle recensioni associate al topic,

facilitando il collegamento tra queste variazioni ed eventi quali rilasci di nuove versioni, cambi di policy o campagne promozionali. L'output *topic_words_over_time*, invece, permette di analizzare l'evoluzione semantica interna ai topic, evidenziando come il significato associato a un topic cambi nel tempo e supportando l'interpretazione qualitativa. Strumenti di questo tipo rispondono all'esigenza di rendere i risultati del text mining comprensibili e azionabili per figure manageriali non specialistiche (Dhar, 2013). Questo si traduce nella capacità di orientare i reparti aziendali. L'identificazione di un picco di lamentele tecniche, ad esempio, potrebbe permettere al reparto IT di isolare bug specifici, mentre lo studio dei topic legati alle nuove funzionalità orienta il marketing nella calibrazione della comunicazione (Reinartz et al., 2019).

L'applicazione dell'analisi della Digital VoC tramite un framework di questo tipo consente quindi di costruire un'infrastruttura di ascolto continuo capace di dialogare con i processi aziendali esistenti. In generale, l'approccio proposto può offrire al management aziendale:

- capacità predittiva e interventi preventivi grazie all'individuazione tempestiva di picchi nel numero di recensioni o cambiamenti semantici (Dhar, 2013);
- supporto alla customer experience e alla fidelizzazione attraverso la comprensione delle reazioni degli utenti a caratteristiche e innovazioni del prodotto o del servizio (Verhoef et al., 2009);
- raccomandazioni strategiche per il monitoraggio della qualità trasformando i topic in indicatori continui delle aree di forza e debolezza.

Sotto il profilo della Gestione della Qualità, l'integrazione di questo modello nei processi aziendali può aiutare il ciclo PDCA (Plan-Do-Check-Act), in particolare nella fase di Check. L'analisi dei topic permette una misurazione della qualità percepita molto più granulare rispetto ai controlli statistici tradizionali, poiché rileva non solo la presenza di una non-conformità, ma ne esplicita la causa attraverso il linguaggio dell'utente (Chan and Wu, 2002). Poter monitorare se un problema scompare dal corpus di recensioni degli utenti dopo un aggiornamento permette di verificare l'efficacia delle soluzioni adottate. In questo modo, le opinioni dei clienti diventano una guida strategica per eliminare i difetti e aumentare il valore percepito del servizio. In definitiva, l'approccio dinamico può supportare il Total Quality Management (TQM) ponendo l'ascolto del cliente al centro del miglioramento continuo (Korkmaz and Barstuğan, 2020).

In sintesi, l'analisi dinamica e semantica dei topic rappresenta un ponte tra dati testuali non strutturati e processi decisionali, consentendo di trasformare qualunque corpus documentale in una fonte di

insight operativi. La scelta di analizzare l'evoluzione temporale in modalità *ex-post* risulta coerente con tali finalità: l'attenzione non è rivolta alla modellazione formale del processo generativo, bensì alla produzione di indicatori trasparenti, interpretabili e immediatamente utilizzabili dalle aziende. Lo script sviluppato, grazie alla flessibilità rispetto alle fonti e alle visualizzazioni, si configura come uno strumento integrabile nei sistemi di gestione della qualità e della customer experience.

Tuttavia, le implicazioni manageriali discusse nel presente lavoro devono essere interpretate alla luce di alcuni limiti strutturali dei dati di Digital Voice of Customer. Questi ultimi rappresentano un campione spontaneo e non probabilistico. Come detto in precedenza, le recensioni online tendono a essere prodotte prevalentemente da utenti con esperienze particolarmente positive o negative e possono non riflettere in modo fedele la percezione della totalità dei clienti (Hu et al., 2007). Inoltre, gli insight derivati descrivono il segmento degli utenti digitalmente attivi e non l'universo complessivo dei consumatori (Schivinski and Dabrowski, 2016). Un ulteriore limite riguarda il fatto che le relazioni tra topic e performance aziendale non possono essere interpretate in termini causali, ma come segnali utili per orientare indagini più approfondite (Dhar, 2013).

Nonostante tali vincoli, la metodologia proposta consente di monitorare in modo sistematico la Digital VoC, individuare precocemente aree critiche e valutare l'impatto di cambiamenti di prodotto o servizio.

7. Conclusioni

Il lavoro si colloca nell'ambito della letteratura sulla Digital Voice of Customer che, attraverso l'impiego di tecniche di Topic Modeling neurale e di analisi dinamica dei topic, evidenzia il potenziale di tali approcci nel convertire grandi volumi di recensioni online in elementi di supporto alle aziende. Il contributo principale di questo lavoro non è tanto l'analisi puntuale delle recensioni della piattaforma Spotify, quanto la definizione e la validazione di un metodo replicabile per l'estrazione, il monitoraggio e l'interpretazione della Digital Voice of Customer. La procedura è stata progettata per essere riutilizzabile su differenti corpus e contesti aziendali. L'applicazione al dataset di Spotify funge da caso di studio per mostrare come la procedura operi nella pratica e quali output interpretabili può generare.

È stata implementata e documentata una metodologia che produce topic stabili e interpretabili per ogni periodo temporale. Le metriche di qualità e i controlli manuali effettuati hanno dimostrato la coerenza semantica dei cluster e la capacità del modello di evidenziare sia trend sia shock legati ad eventi. L'analisi del vocabolario associato ai topic su finestre temporali ha mostrato come la comparsa di nuove parole chiave renda interpretabili i cambiamenti di frequenza dei topic. Tutti i passaggi sono stati accompagnati da suggerimenti operativi per facilitare la riproducibilità in contesti diversi.

La decisione di adottare BERTopic per clusterizzare l'intero corpus e successivamente analizzare le variazioni temporali *ex-post* è stata guidata da criteri di interpretabilità, robustezza e coerenza con gli obiettivi applicativi. I modelli dinamici tradizionali, infatti, descrivono l'evoluzione dei topic attraverso meccanismi di transizione che favoriscono una continuità semantica tra epoche. Tale impostazione è efficace quando i temi evolvono in modo progressivo, ma può risultare meno adatta in contesti caratterizzati da discontinuità improvvise, tipiche della Digital Voice of Customer. In quest'ultima, infatti, sono presenti picchi repentini e la nascita di nuove parole chiave, dovuti a rilasci di prodotto, cambi di policy, campagne di marketing o bug. L'approccio *ex-post* permette di evitare i vincoli di dipendenza temporale (smoothness) che potrebbero mascherare questi shock e garantisce che i topic rimangano confrontabili tra diverse epoche poiché estratti da uno spazio semantico comune e stabile. Inoltre, questa soluzione garantisce maggiore flessibilità operativa. Può essere applicata a corpus con granularità temporale eterogenea, periodi sbilanciati o sequenze irregolari di osservazioni, condizioni frequenti nei dataset aziendali.

Per quanto riguarda le limitazioni del lavoro è doveroso citare l'assegnazione del documento a un solo topic, aspetto tipico di BERTopic che potrebbe semplificare eccessivamente recensioni molto lunghe

o articolate che trattano più argomenti contemporaneamente. Un secondo limite riguarda l'analisi dell'evoluzione dei topic: affinché i trend temporali siano affidabili è necessario disporre di finestre con un numero sufficiente di osservazioni. In presenza di periodi molto sbilanciati la capacità di cogliere piccoli shock può ridursi o risultare instabile. Inoltre, nonostante il c-TF-IDF, la validazione semantica finale rimane parzialmente manuale e la validazione quantitativa dell'evoluzione temporale è meno sviluppata rispetto alla valutazione del modello statico.

Il lavoro svolto apre diverse direzioni di sviluppo. Un primo ambito riguarda l'integrazione di modelli linguistici per il labeling automatico dei topic. L'utilizzo di LLM per generare etichette e spiegazioni potrebbe ridurre in modo significativo l'intervento manuale, migliorando l'uniformità semantica.

Un secondo sviluppo rilevante consiste nella combinazione tra Topic Modeling e analisi del sentiment. Affiancare all'evoluzione dei topic una misura del sentiment consentirebbe di comprendere come cambia la percezione degli utenti rispetto a ciascun tema. L'integrazione del sentiment potrebbe inoltre essere estesa al livello lessicale: l'interpretazione delle parole chiave associate ai topic risulterebbe più ricca distinguendo se i termini compaiono in contesti prevalentemente positivi o negativi nel corso del tempo. Ciò permetterebbe di cogliere variazioni semantiche sottili e di supportare interventi mirati su funzionalità, processi o comunicazione.

Ulteriore direzione riguarda l'estensione multi-corpus. L'applicazione congiunta a fonti eterogenee, come ticket di assistenza, social media, forum e survey, permetterebbe di valutare la robustezza dell'approccio e di costruire indicatori più completi, capaci di integrare segnali provenienti da touchpoint differenti della customer journey.

Sarebbe infine utile effettuare un confronto sperimentale controllato con modelli dinamici di Topic Modeling, al fine di quantificare il trade-off tra continuità semantica e sensibilità agli shock informativi.

In conclusione, questa tesi mette a disposizione non solo un'analisi applicata al caso studio di Spotify, ma soprattutto uno script e una procedura riproducibile, pensata per essere adottata e adattata in diversi contesti aziendali.

8. Bibliografia

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A., 2023. Topic modeling algorithms and applications: A survey. Information Systems 112, 102131. https://doi.org/10.1016/j.is.2022.102131

Aguwa, C.C., Monplaisir, L., Turgut, O., 2012. Voice of the customer: Customer satisfaction ratio based analysis. Expert Systems with Applications 39, 10112–10119. https://doi.org/10.1016/j.eswa.2012.02.071

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures q.

Ajinaja, M.O., Fakoya, J.T., Ogunwale, Y.E., Omoniyi, J.K., Ibiyomi, M.A., Abiona, A.A., Akinola, D., 2025. A Comparative Evaluation of Probabilistic and Transformer-Based Topic Models Across Diverse and Multilingual Text Corpora. Neural Process Lett 58, 9. https://doi.org/10.1007/s11063-025-11820-3

Alamsyah, A., Girawan, N.D., 2023. Improving Clothing Product Quality and Reducing Waste Based on Consumer Review Using RoBERTa and BERTopic Language Model. BDCC 7, 168. https://doi.org/10.3390/bdcc7040168

Ali, M., van Berkel, N., Tag, B., Paananen, V., Oppenlaender, J., Yatani, K., Hosio, S., 2025. Investigating mental wellbeing self-care in higher education using BERTopic modeling. Discov Ment Health 5, 204. https://doi.org/10.1007/s44192-025-00323-1

Baier, D., Decker, R., Asenova, Y., 2025. Collecting and Analyzing User-Generated Content for Decision Support in Marketing Management: An Overview of Methods and Use Cases. Schmalenbach J Bus Res 77, 419–455. https://doi.org/10.1007/s41471-025-00208-7

Barravecchia, F., Franceschini, F., Mastrogiacomo, L., Zaki, M., 2021. Research on product-service systems: topic landscape and future trends. JMTM 32, 208–238. https://doi.org/10.1108/JMTM-04-2020-0164

Barravecchia, F., Mastrogiacomo, L., Casadesús Fa, M., Franceschini, F., 2024. MOBI-Qual: a common framework to manage the product-service system quality of shared mobility. Flex Serv Manuf J 36, 1359–1398. https://doi.org/10.1007/s10696-023-09520-y

Barravecchia, F., Mastrogiacomo, L., Franceschini, F., 2023. Product quality tracking based on digital Voice-of-Customers. Total Quality Management & Business Excellence 34, 1386–1409. https://doi.org/10.1080/14783363.2023.2177147

Barravecchia, F., Mastrogiacomo, L., Franceschini, F., 2022. Digital voice-of-customer processing by topic modelling algorithms: insights to validate empirical results. IJQRM 39, 1453–1470. https://doi.org/10.1108/IJQRM-07-2021-0217

Barravecchia, F., Mastrogiacomo, L., Franceschini, F., 2020. Categorizing Quality Determinants in Mining User-Generated Contents. Sustainability 12, 9944. https://doi.org/10.3390/su12239944

Bianchi, F., Terragni, S., Hovy, D., 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, in: Zong, C., Xia, F., Li, W., Navigli, R. (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Presented at the ACL-IJCNLP 2021, Association for Computational Linguistics, Online, pp. 759–766. https://doi.org/10.18653/v1/2021.acl-short.96

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55, 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association 112, 859–877. https://doi.org/10.1080/01621459.2017.1285773

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning - ICML '06. Presented at the the 23rd international conference, ACM Press, Pittsburgh, Pennsylvania, pp. 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D.M., Ng, A.Y., Jordan, M.I. l, 2003. Latent Dirichlet Allocation [WWW Document]. URL https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed 6.7.25).

Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J., 2015. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Trans. Knowl. Discov. Data 10, 5:1-5:51. https://doi.org/10.1145/2733381

Chan, L.-K., Wu, M.-L., 2002. Quality function deployment: A literature review.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., Blei, D., 2009. Reading Tea Leaves: How Humans Interpret Topic Models, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Dhar, V., 2013. Data science and prediction. Commun. ACM 56, 64–73. https://doi.org/10.1145/2500499

Dieng, A.B., Ruiz, F.J.R., Blei, D.M., 2020. Topic Modeling in Embedding Spaces. Transactions of the Association for Computational Linguistics 8, 439–453. https://doi.org/10.1162/tacl_a_00325

Donald, J., Banner, J., Satria, R., Tania, W., James, W., 2024. Sentiment analysis of user-generated content. 2024.

Griffin, A., Hauser, J.R., 1993. The Voice of the Customer. Marketing Science 12, 1–27. https://doi.org/10.1287/mksc.12.1.1

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arXiv.2203.05794

Hankar, M., Kasri, M., Beni-Hssane, A., 2025. A comprehensive overview of topic modeling: Techniques, applications and challenges. Neurocomputing 628, 129638. https://doi.org/10.1016/j.neucom.2025.129638

Hu, N., Pavlou, P.A., Zhang, J. (Jennifer), 2007. Why Do Online Product Reviews Have a J-Shaped Distribution? Overcoming Biases in Online Word-of-Mouth Communication. SSRN Journal. https://doi.org/10.2139/ssrn.2380298

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 78, 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Jin, X., Zhou, W., Zhu, Q., Wang, W., Xu, G., 2025. Research on the analysis and application of technological supply and demand structure based on LDA and BERTopic models. Cognitive Robotics 5, 260–275. https://doi.org/10.1016/j.cogr.2025.07.001

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., Sagiadinos, V., 2019a. Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. Expert Systems with Applications 116, 472–486. https://doi.org/10.1016/j.eswa.2018.09.037

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., Sagiadinos, V., 2019b. Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. Expert Systems with Applications 116, 472–486. https://doi.org/10.1016/j.eswa.2018.09.037

Korkmaz, M., Barstuğan, M., 2020. A Deep Learning-Based Quality Control Application. European Journal of Science and Technology. https://doi.org/10.31590/ejosat.804744

Krishnan, A., 2023. Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis. https://doi.org/10.48550/arXiv.2308.11520

Lau, J.H., Armendariz, C., Lappin, S., Purver, M., Shu, C., 2020. How Furiously Can Colorless Green Ideas Sleep? Sentence Acceptability in Context. Transactions of the Association for Computational Linguistics 8, 296–310. https://doi.org/10.1162/tacl_a_00315

Li, S., Liu, F., Zhang, Y., Zhu, B., Zhu, H., Yu, Z., 2022. Text Mining of User-Generated Content (UGC) for Business Applications in E-Commerce: A Systematic Review. Mathematics 10, 3554. https://doi.org/10.3390/math10193554

López, A.B., Pastor-Galindo, J., Ruipérez-Valiente, J.A., 2024. LLM-assisted topic modeling for hate speech characterization. https://doi.org/10.22541/au.172966882.21215291/v1

Lupșa-Tătaru, F.R., Lixăndroiu, R.C., Lupșa-Tătaru, D.A., 2023. A Sustainable Analysis Regarding the Impact of Tourism on Food Preferences in European Capitals. Sustainability 15, 14899. https://doi.org/10.3390/su152014899

Mastrogiacomo, L., Barravecchia, F., Franceschini, F., Marimon, F., 2021. Mining quality determinants of product-service systems from user-generated contents. Quality Engineering 33, 425–442. https://doi.org/10.1080/08982112.2021.1877305

McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426

Melzner, J., Bonezzi, A., Meyvis, T., 2023. Information Disclosure in the Era of Voice Technology. Journal of Marketing 87, 491–509. https://doi.org/10.1177/00222429221138286

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing Semantic Coherence in Topic Models, in: Barzilay, R., Johnson, M. (Eds.), Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2011, Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 262–272.

Murshed, B.A.H., Mallappa, S., Abawajy, J., Saif, M.A.N., Al-ariki, H.D.E., Abdulwahab, H.M., 2023. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. Artif Intell Rev 56, 5133–5260. https://doi.org/10.1007/s10462-022-10254-w

Mustak, M., Hallikainen, H., Laukkanen, T., Plé, L., Hollebeek, L.D., Aleem, M., 2024. Using machine learning to develop customer insights from user-generated content. Journal of Retailing and Consumer Services 81, 104034. https://doi.org/10.1016/j.jretconser.2024.104034

Nedungadi, P., Veena, G., Tang, K.-Y., Menon, R.R.K., Raman, R., 2025. AI Techniques and Applications for Online Social Networks and Media: Insights From BERTopic Modeling. IEEE Access 13, 37389–37407. https://doi.org/10.1109/ACCESS.2025.3543795

Opitz, J., Möller, L., Michail, A., Padó, S., Clematide, S., 2025. Interpretable Text Embeddings and Text Similarity Explanation: A Survey. https://doi.org/10.48550/arXiv.2502.14862

Palese, B., Usai, A., 2018. The relative importance of service quality dimensions in E-commerce experiences. International Journal of Information Management 40, 132–140. https://doi.org/10.1016/j.ijinfomgt.2018.02.001

Paltoglou, G., Thelwall, M., 2010. A Study of Information Retrieval Weighting Schemes for Sentiment Analysis, in: Hajič, J., Carberry, S., Clark, S., Nivre, J. (Eds.), Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Presented at the ACL 2010, Association for Computational Linguistics, Uppsala, Sweden, pp. 1386–1395.

Parasurman, A., Zeithaml, V.A., Berry, L.L., 1988. SERVQUAL-A-Multiple-Item-Scale-for-Measuring-Consumer-Perceptions-of-Service-Quality-libre [WWW Document]. URL https://d1wqtxts1xzle7.cloudfront.net/46268843/SERVQUAL-A-Multiple-Item-Scale-for-Measuring-Consumer-Perceptions-of-Service-Quality-libre.pdf?1465197022=&response-content-disposition=inline%3B+filename%3DSERVQUAL_A_Multiple_Item_Scale_for_Measu.pdf&Expires=1771525195&Signature=WU0NBXJ5XmoX497gbioDAScY8mBS46cpJBswyDCv1Cw9TQoHRnbliEdBtIGKOW0DtWOhL44czlgis7DMctGTD1UnnS8V-N0EVqx4ezOshBER0VgLnVCWPZLfGLjH38cIS0mSVvDZikFnCl5ji1OrXpZ2Wrd4jgpdXUPS51mbPCKlfbxZYUFLCK31U1S-L4X9I7m9x9bZo6q9SKoNoNApYRtK~Q0oYVKsE5sfoLBECCAniTZk4uj5RwkFIdi6o~VP3zl3G0MXNxqrrhJupjsCGybnSFDWBay5x~P7ljYdhjRHG4AVhmwUMDwiNcyrv3Xv1fdiplZ5Lv-02athJVGO0A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA (accessed 2.19.26).

Park, K., Cha, M., Rhim, E., 2018. Positivity Bias in Customer Satisfaction Ratings, in: Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. pp. 631–638. https://doi.org/10.1145/3184558.3186579

Rahimi, H., Mimno, D., Hoover, J.L., Naacke, H., Constantin, C., Amann, B., 2024. Contextualized Topic Coherence Metrics.

Reichheld, F.F., 2003. The One Number You Need to Grow.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084

Reinartz, W., Wiegand, N., Imschloss, M., 2019. The impact of digital transformation on the retailing value chain. International Journal of Research in Marketing 36, 350–366. https://doi.org/10.1016/j.ijresmar.2018.12.002

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural Topic Models for Open-Ended Survey Responses. American J Political Sci 58, 1064–1082. https://doi.org/10.1111/ajps.12103

Santos, M.L.B.D., 2022. The "so-called" UGC: an updated definition of user-generated content in the age of social media. OIR 46, 95–113. https://doi.org/10.1108/OIR-06-2020-0258

Schivinski, B., Dabrowski, D., 2016. The effect of social media communication on consumer perceptions of brands. Journal of Marketing Communications 22, 189–214. https://doi.org/10.1080/13527266.2013.871323

Sharma, N.K., Kumar, V., Singh, A., Verma, P., 2025. International Journal of Quality & Reliability Management. https://doi.org/10.1108/ijqrm

Srinivas, S., Ramachandiran, S., 2020. Discovering Airline-Specific Business Intelligence from Online Passenger Reviews: An Unsupervised Text Analytics Approach. https://doi.org/10.48550/arXiv.2012.08000

Subhashini, L.D.C.S., Li, Y., Zhang, J., Atukorale, A.S., Wu, Y., 2021. Mining and classifying customer reviews: a survey. Artif Intell Rev 54, 6343–6389. https://doi.org/10.1007/s10462-021-09955-5

Tangherlini, T.R., Chen, R., 2024. Travels with BERT: Surfacing the intertextuality in Hans Christian Andersen's travel writing and fairy tales through the network lens of large language model-based topic modeling. Orbis Litterarum 79, 519–562. https://doi.org/10.1111/oli.12458

Thapa, M., Kapoor, P., Kaushal, S., Sharma, I., 2024. A Review of Contextualized Word Embeddings and Pre-Trained Language Models, with a Focus on GPT and BERT:, in: Proceedings of the 1st International Conference on Cognitive & Cloud Computing. Presented at the International Conference on Cognitive & Cloud Computing, SCITEPRESS - Science and Technology Publications, Jaipur, India, pp. 205–214. https://doi.org/10.5220/0013305900004646

Tirunillai, S., Tellis, G.J., 2014. Extracting Dimensions of Consumer Satisfaction with Quality from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. SSRN Journal. https://doi.org/10.2139/ssrn.2408855

Tripodi, R., 2021. How Contextualized Word Embeddings Represent Word Senses, in: Fersini, E., Passarotti, M., Patti, V. (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-It 2021). Presented at the CLiC-it 2021, CEUR Workshop Proceedings, Milan, Italy, pp. 337–345.

Vayansky, I., Kumar, S.A.P., 2020. A review of topic modeling methods. Information Systems 94, 101582. https://doi.org/10.1016/j.is.2020.101582

Verhoef, P.C., Lemon, K.N., Parasuraman, A., Roggeveen, A., Tsiros, M., Schlesinger, L.A., 2009. Customer Experience Creation: Determinants, Dynamics and Management Strategies. Journal of Retailing 85, 31–41. https://doi.org/10.1016/j.jretai.2008.11.001

Verleysen, M., François, D., 2005. The Curse of Dimensionality in Data Mining and Time Series Prediction, in: Cabestany, J., Prieto, A., Sandoval, F. (Eds.), Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 758–770. https://doi.org/10.1007/11494669_93

Wang, X., McCallum, A., 2006. Topics over time: a non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06. Association for Computing Machinery, New York, NY, USA, pp. 424–433. https://doi.org/10.1145/1150402.1150450

Wani, M.A., ElAffendi, M., Shakil, K.A., 2024. AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing. Computers 13, 264. https://doi.org/10.3390/computers13100264

Yi, J., Oh, Y.K., Kim, J.-M., 2025. Unveiling the drivers of satisfaction in mobile trading: Contextual mining of retail investor experience through BERTopic and generative AI. Journal of Retailing and Consumer Services 82, 104066. https://doi.org/10.1016/j.jretconser.2024.104066

Zhang, Y., Jin, R., Zhou, Z.-H., 2010. Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. & Cyber. 1, 43–52. https://doi.org/10.1007/s13042-010-0001-0

Appendice A

```
#SETUP E CONFIGURAZIONE AMBIENTE
#le installazioni vanno eseguite solo alla prima configurazione (decommentare per eseguire
l'installazione)
#install.packages('reticulate')
#install.packages('stringr')
#install.packages('dplyr')
#install.packages("cld3")
#install.packages("tokenizers")
#install.packages("textstem")
#install.packages("SnowballC")
#install.packages("lubridate")
#install.packages("plotly")
#install.packages("openxlsx")
#install.packages("tidytext")

library(reticulate)
#eseguire virtualenv_create solo alla prima configurazione dell'ambiente
#virtualenv_create("bertopic-env",required=TRUE)
use_virtualenv("bertopic-env", required = TRUE)
#eseguire py_install solo alla prima configurazione dell'ambiente
#py_install(c("bertopic", "umap-learn", "hdbscan", "sentence-transformers"))
library(stringr)
library(dplyr)
library(cld3)
library(tokenizers)
library(textstem)
library(SnowballC)
library(lubridate)
library(plotly)
library(openxlsx)
library(tidytext)

#assicurarsi che I file di input siano nella stessa working directory
recensioni <- read.csv("Spotify.csv", stringsAsFactors = FALSE, sep=";")

#PREPROCESSING DEL TESTO
#conversione in UTF-8
recensioni$Review <- iconv(recensioni$Review, to="UTF-8", sub="")
#rimozione caratteri invisibili
recensioni$Review <- gsub("[\r\n\t]", " ", recensioni$Review)
#rimozione recensioni vuote o NA
recensioni <- recensioni[!is.na(recensioni$Review) & recensioni$Review != "", ]
#filtro lingua inglese
lang <- cld3::detect_language(recensioni$Review)
recensioni_en <- recensioni[!is.na(lang) & lang == "en", ]
cat("Dopo filtro lingua:", nrow(recensioni_en), "\n")
#filtro recensioni con più di 5 parole
recensioni_en <- recensioni_en[sapply(strsplit(recensioni_en$Review, " "), length) >= 5, ]
cat("Dopo filtro >=5 parole:", nrow(recensioni_en), "\n")

clean_text <- function(txt) {
  txt <- txt %>%
    stringr::str_to_lower() %>%
    stringr::str_replace_all("[\r\n\t]", " ") %>%
    stringr::str_replace_all("http\\S+|www\\S+", " ") %>%
    stringr::str_replace_all("[^a-z ]", " ") %>%
    stringr::str_replace_all("\\s+", " ") %>%
    stringr::str_trim()
  #tokenizzazione
  tokens <- tokenizers::tokenize_words(txt)[[1]]
  #rimozione token troppo corti
  tokens <- tokens[nchar(tokens) > 2]
  #lemmatizzazione
  tokens <- textstem::lemmatize_words(tokens)
  #ricomposizione del testo
  txt_clean <- paste(tokens, collapse = " ")
  return(txt_clean)}

recensioni_en$CleanReview <- sapply(recensioni_en$Review,clean_text,USE.NAMES = FALSE)
recensioni_en$CleanReview <- as.character(recensioni_en$CleanReview)

#filtro recensioni con almeno 5 parole dopo pulizia
recensioni_en <- recensioni_en[sapply(strsplit(recensioni_en$CleanReview, " "), length) >= 5, ]
cat("Dopo pulizia >=5 parole:", nrow(recensioni_en), "\n")
```

```
recensioni_en$Date_clean <- as.Date(
  substr(recensioni_en$Date, 1, 10),
  format = "%d/%m/%Y")
stopifnot(!all(is.na(recensioni_en$Date_clean)))
recensioni_en$Year <- format(recensioni_en$Date_clean, "%Y")

saveRDS(recensioni_en, "recensioni_pulite.rds")
testi <- recensioni_en$CleanReview

#CREAZIONE E ALLENAMENTO MODELLO STATICO
bertopic <- import("bertopic")

from_sentence_transformers <- import("sentence_transformers")$SentenceTransformer
encoder <- from_sentence_transformers("all-MiniLM-L6-v2")
embeddings <- encoder$encode(testi, batch_size = 128L)

umap <- import("umap")
umap_model <- umap$UMAP(
  n_neighbors = 10L,
  n_components = 5L,
  min_dist = 0.1,
  metric = "cosine",
  random_state = 42L)

hdbscan <- import("hdbscan")
hdbscan_model <- hdbscan$HDBSCAN(
  min_cluster_size = 200L,
  min_samples = 25L,
  metric = "euclidean",
  prediction_data = TRUE)

sklearn <- import("sklearn.feature_extraction.text")
CountVectorizer <- sklearn$CountVectorizer
vectorizer <- CountVectorizer(
  min_df = 0.005,
  max_df = 0.85,
  stop_words = c("english", "spotify"),
  ngram_range = tuple(1L, 2L))

model <- bertopic$BERTopic(
  embedding_model = encoder,
  umap_model = umap_model,
  hdbscan_model = hdbscan_model,
  vectorizer_model = vectorizer,
  representation_model = NULL,
  min_topic_size = 200L,
  top_n_words = 5L,
  language = "english",
  calculate_probabilities = TRUE,
  verbose = TRUE)

joblib <- import("joblib")
parallel_backend <- joblib$parallel_backend

with(parallel_backend("threading", n_jobs = -1L), {
  res <- model$fit_transform(testi, embeddings)})

topics <- res[[1]]
probs  <- res[[2]]
saveRDS(model, "bertopic_spotify_final.rds")
saveRDS(topics, "topics_static_final.rds")
saveRDS(probs, "probs_static_final.rds")
saveRDS(testi, "texts_static_final.rds")
saveRDS(recensioni_en$Date_clean, "dates_static_final.rds")
saveRDS(embeddings, "embeddings_static_final.rds")

#TEST
#statistiche sui topic
sum(topics == -1) / length(topics) * 100
length(unique(topics))
head(topics, 10)
table(topics)
#statistiche sulle probabilità
dim(probs)
probs_df <- as.data.frame(probs)
head(probs_df[, 1:5])
```

```
#FILTRAGGIO DEL RUMORE
topics_refined <- model$reduce_outliers(
  testi,
  topics,
  probabilities = probs,
  strategy = "probabilities",
  threshold = 0.2)
topics_refined <- as.integer(unlist(topics_refined))

c(noise_before = mean(topics == -1),
  noise_after  = mean(topics_refined == -1))

tab_before <- as.data.frame(table(topics))
tab_after  <- as.data.frame(table(topics_refined))
colnames(tab_before) <- c("Topic", "Before")
colnames(tab_after)  <- c("Topic", "After")
comparison <- merge(
  tab_before,
  tab_after,
  by = "Topic",
  all = TRUE)
comparison[is.na(comparison)] <- 0
comparison

sum(topics == -1 & topics_refined != -1)

saveRDS(comparison, "topic_allocation_before_after.rds")
saveRDS(topics_refined, "topics_refined_vector.rds")

#SIMILARITA' TOPIC RECENSIONE
cosine_sim <- function(a, b) {sum(a * b) / (sqrt(sum(a^2)) * sqrt(sum(b^2)))}

topic_embeddings<- model$topic_embeddings_
doc_embeddings<- embeddings

valid_topics<- sort(unique(topics_refined[topics_refined >= 0]))

similarity_matrix <- sapply(
  valid_topics,
  function(t) {
    topic_emb <- topic_embeddings[t + 1, ]  # +1 per indexing R
    apply(doc_embeddings, 1, function(d)
      cosine_sim(d, topic_emb))})
colnames(similarity_matrix) <- paste0("Topic_", valid_topics)
reviews_original <- recensioni_en$Review
topic_examples <- lapply(
  seq_along(valid_topics),
  function(i) {
    t <- valid_topics[i]
    sim_values <- similarity_matrix[, i]
    idx <- which(topics_refined == t)
    idx <- idx[order(sim_values[idx], decreasing = TRUE)]
    data.frame(
      Topic        = t,
      Similarity   = round(sim_values[idx], 3),
      Review       = reviews_original[idx],
      stringsAsFactors = FALSE)})
topic_examples_df <- dplyr::bind_rows(topic_examples)

write.csv2(
  topic_examples_df,
  "all_topics_all_reviews_ordered_similarity.csv",
  row.names = FALSE)

#LABELING
topic_info <- as.data.frame(model$get_topic_info())

topic_info <- topic_info[topic_info$Topic != -1, ]

topic_info$Words <- sapply(
  topic_info$Representation,
  function(x) paste(x, collapse = ", "))

topic_info$AutoLabel <- sapply(
  topic_info$Representation,
  function(x) paste(x[1:3], collapse = "_"))
```

```
topic_info_for_labeling <- topic_info[, c(
  "Topic",
  "Count",
  "Words",
  "AutoLabel")]

saveRDS(topic_info_for_labeling, "topic_info_for_labeling.rds")
write.xlsx(
  topic_info_for_labeling,
  "topic_info_for_labeling.xlsx",
  overwrite = TRUE)

#RE-IMPORT DOPO LABELING
topic_labels <- read.xlsx("topic_info_for_labeling.xlsx")

stopifnot("HumanLabel" %in% colnames(topic_labels))
stopifnot(!any(is.na(topic_labels$HumanLabel)))

topic_info_labeled <- merge(
  topic_info,
  topic_labels[, c("Topic", "HumanLabel")],
  by = "Topic",
  all.x = TRUE)

stopifnot(!any(is.na(topic_info_labeled$HumanLabel)))

saveRDS(topic_info_labeled, "topic_info_labeled_final.rds")

#MERGE TOPIC SIMILI
topics_merged <- topics_refined

topics_merged[topics_merged == 6]  <- 4

table(topics_merged)

topic_info_merged <- topic_info_labeled %>%
  filter(Topic != 6)

topic_info_merged$HumanLabel[topic_info_merged$Topic == 4] <- "Accesso ai lyrics"

saveRDS(topic_info_merged, "topic_info_merged.rds")
saveRDS(topics_merged, "topics_merged_vector.rds")
saveRDS(topic_info_merged, "topic_info_final.rds")

#VISUALIZZAZIONE
topic_counts <- as.data.frame(table(topics_merged))
colnames(topic_counts) <- c("Topic", "Count_refined")
topic_counts$Topic <- as.integer(as.character(topic_counts$Topic))
topic_info_plot <- topic_info_merged %>%
  left_join(topic_counts, by = "Topic") %>%
  arrange(desc(Count_refined))

fig <- plot_ly(
  data = topic_info_plot,
  x = ~reorder(HumanLabel, Count_refined),
  y = ~Count_refined,
  type = "bar",
  marker = list(
    color = 'rgba(30, 215, 96, 0.7)',
    line = list(color = 'rgba(30, 215, 96, 1.0)', width = 1)),
  text = ~Words,
  hoverinfo = "text+y+x") %>%
  layout(
    title = "Distribuzione Recensioni per Topic (Post-Unione)",
    xaxis = list(title = "", tickangle = -45),
    yaxis = list(title = "Numero di Recensioni"),
    margin = list(b = 120))
fig

htmlwidgets::saveWidget(fig, "barchart_topics_merged.html")

#barchart (topic pre-merged)
topic_info <- model$get_topic_info()
n_topics <- sum(topic_info$Topic != -1)
n_topics <- as.integer(n_topics)
barchart <- model$visualize_barchart(
```

```
  top_n_topics = n_topics,
  height = 700L)
barchart$write_html("barchart_topics.html")

#mappa dei topic (topic pre-merged)
viz_topics <- model$visualize_topics(height = 700L)
viz_topics$write_html("topics_map.html")

#matrice similarità (topic pre-merged)
viz_heatmap <- model$visualize_heatmap(n_clusters = 10L)
viz_heatmap$write_html("topics_heatmap.html")

#tabella dei topic con i label
topic_info_df <- topic_info_merged %>%
  left_join(topic_counts, by = "Topic") %>%
  select(Topic, HumanLabel, Count_refined, Words, AutoLabel)

write.csv(topic_info_df, "topic_info_df.csv", row.names = FALSE)

#VALIDAZIONE MODELLO
#TOPIC COHERENCE
#eseguire py_install solo alla prima configurazione
#py_install("gensim")
gensim <- import("gensim")
gensim_corp <- import("gensim.corpora")
gensim_mod  <- import("gensim.models")

topics_raw <- topic_info_merged$Words
topics_cleaned <- lapply(topics_raw, function(x) {
  tokens <- unlist(strsplit(x, ", "))
  tokens_split <- unlist(strsplit(tokens, " "))
  return(unique(tokens_split))})

docs <- tokenizers::tokenize_words(testi)

dictionary <- gensim_corp$Dictionary(docs)
corpus <- lapply(docs, function(d) dictionary$doc2bow(d))

coherence_model <- gensim_mod$coherencemodel$CoherenceModel(
  topics = topics_cleaned,
  corpus = corpus,
  dictionary = dictionary,
  coherence  = 'u_mass')
umass_score <- coherence_model$get_coherence()
cat("Topic Coherence (U_mass):", umass_score, "\n")

#TOPIC DIVERSITY
all_words_raw <- unlist(lapply(topics_raw, function(x) {
  tokens <- unlist(strsplit(x, ", "))
  unlist(strsplit(tokens, " "))}))
all_words_clean <- all_words_raw[nchar(all_words_raw) > 1]
topic_diversity <- length(unique(all_words_clean)) / length(all_words_clean)
cat("Topic Diversity:", round(topic_diversity, 3), "\n")

#DISTRIBUZIONE TEMPORALE DELLE RECENSIONI NEI TOPIC
testi_clean  <- testi
dates <- recensioni_en$Date_clean

stopifnot(length(testi_clean)  == length(topics_merged))
stopifnot(length(dates) == length(topics_merged))

year_vec <- format(dates, "%Y")

df_base_time <- data.frame(
  Topic  = topics_merged,
  Year  = year_vec,
  stringsAsFactors = FALSE)
topic_time_distribution_merged <- df_base_time %>%
  filter(Topic >= 0) %>%
  count(Topic, Year) %>%
  group_by(Topic) %>%
  mutate(perc_within_topic = n / sum(n) * 100) %>%
  ungroup() %>%
  arrange(Topic, Year)
saveRDS(
  topic_time_distribution_merged,
  "topic_time_distribution_merged.rds")
```

```
write.xlsx(
  topic_time_distribution_merged,
  "topic_time_distribution_merged.xlsx",
  overwrite = TRUE)

#PAROLE DINAMICHE PER TOPIC STATICI
df_base <- data.frame(
  Review = testi_clean,
  Topic  = topics_merged,
  Year = year_vec,
  stringsAsFactors = FALSE)
stopifnot(!6 %in% unique(df_base$Topic)) #verifica assenza topic 6 (topic unito)

df_base <- df_base[df_base$Topic >= 0, ]

#tokenizzazione e pulizia
data("stop_words")

df_tokens <- df_base %>%
  unnest_tokens(word, Review) %>%
  filter(nchar(word) >= 4) %>%
  filter(!grepl("[^a-z]", word)) %>%
  anti_join(stop_words, by = "word") %>%
  mutate(
    word = lemmatize_words(word),
    word = wordStem(word, language = "en"))

domain_stopwords <- tibble(word = c(
  "spotify", "spotifi", "playlist", "song", "listen", "pause", "log"))

df_tokens <- df_tokens %>%
  anti_join(domain_stopwords, by = "word")

df_ctfidf_time <- df_tokens %>%
  count(Topic, Year, word) %>%
  filter(n >= 3) %>%
  group_by(Topic, Year) %>%
  mutate(tf = n / sum(n)) %>%
  ungroup()

df_ctfidf_time <- df_ctfidf_time %>%
  group_by(word) %>%
  mutate(df = n_distinct(interaction(Topic, Year))) %>%
  ungroup() %>%
  mutate(
    idf = log((1 + n_distinct(interaction(Topic, Year))) / (1 + df)),
    c_tfidf = tf * idf)

period_words <- df_ctfidf_time %>%
  group_by(Topic, Year) %>%
  slice_max(c_tfidf, n = 6, with_ties = FALSE) %>%
  summarise(
    period_words = paste(word, collapse = ", "),
    .groups = "drop")

topic_labels_final <- topic_info_merged %>%
  select(Topic, HumanLabel)

topic_words_over_time <- period_words %>%
  left_join(topic_labels_final, by = "Topic") %>%
  arrange(Topic, Year)

stopifnot(!any(is.na(topic_words_over_time$HumanLabel)))

write.csv(
  topic_words_over_time,
  "topic_words_over_time.csv",
  row.names = FALSE)
```

Appendice B

Si ripotano di seguito i grafici relativi alla frequenza delle recensioni appartenenti al topic negli anni e la tabella delle parole chiave nel tempo associate ai topic non discussi nel capitolo “Risultati dell’analisi dinamica”.

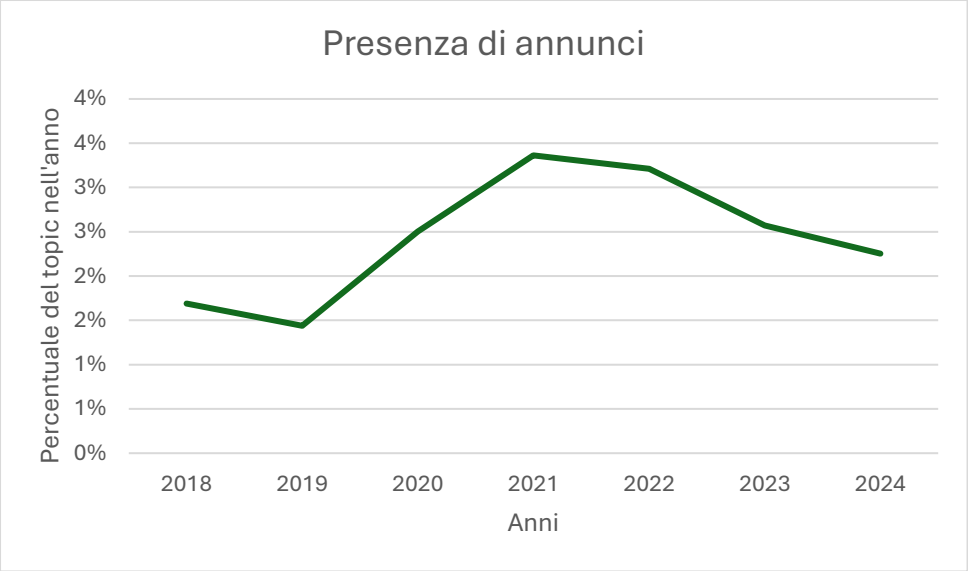

Figura 33 - Percentuale di recensioni appartenenti al topic "Presenza di annunci" negli anni

Topic	Year	period_words	HumanLabel
3	2018	tape, advertis, minut, annoy, short, premium	Presenza di annunci
3	2019	minut, commerci, free, wearabl, advertis, skip	Presenza di annunci
3	2020	minut, free, watch, premium, skip, commerci	Presenza di annunci
3	2021	minut, watch, uninterrupt, advertis, free, video	Presenza di annunci
3	2022	skip, minut, premium, free, annoy, advertis	Presenza di annunci
3	2023	minut, skip, premium, advertis, free, annoy	Presenza di annunci
3	2024	premium, minut, advertis, skip, free, lyric	Presenza di annunci

Figura 34 - Parole rappresentative del topic "Presenza di annunci” per anno

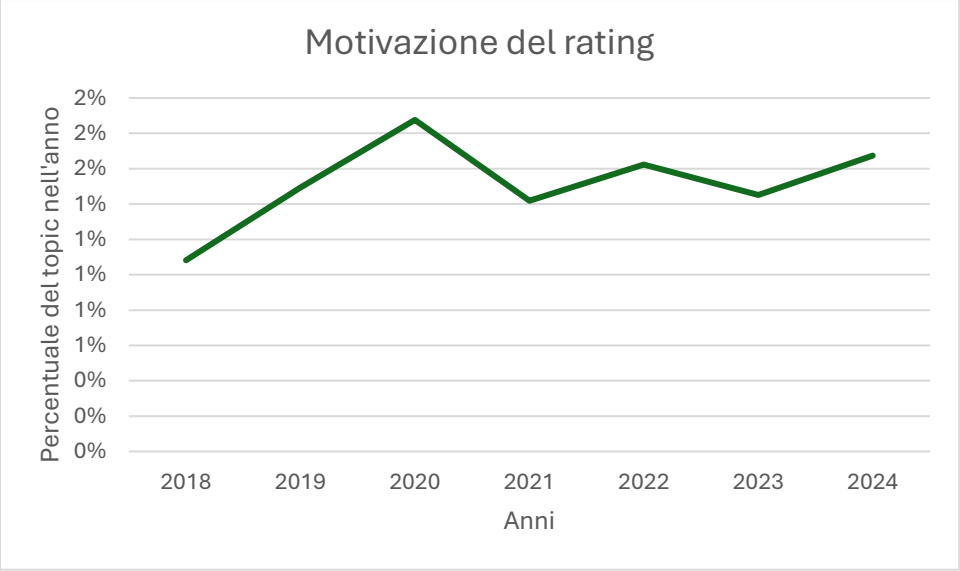

Figura 35 - Percentuale di recensioni appartenenti al topic "Motivazione del rating" negli anni

Topic	Year	period_words	HumanLabel
5	2018	star, offer, pick, shuffl, remov, button	Motivazione del rating
5	2019	star, rate, shuffl, skip, start, updat	Motivazione del rating
5	2020	star, rate, skip, watch, random, hour	Motivazione del rating
5	2021	star, reason, minut, rate, uniti, phone	Motivazione del rating
5	2022	star, rate, skip, reason, shuffl, start	Motivazione del rating
5	2023	star, rate, skip, premium, shuffl, random	Motivazione del rating
5	2024	star, rate, deserv, skip, lyric, premium	Motivazione del rating

Figura 36 - Parole rappresentative del topic "Motivazione del rating" per anno

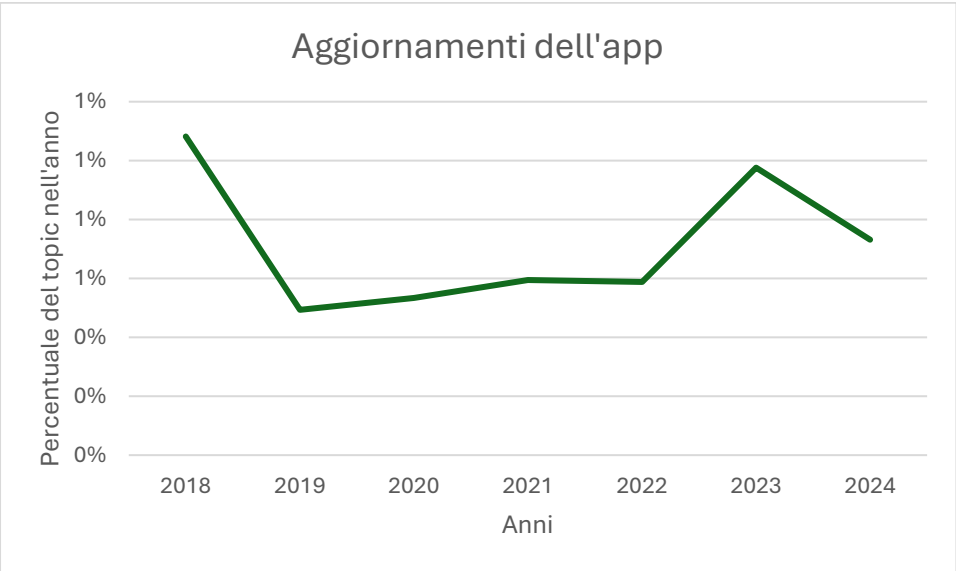

Figura 37 - Percentuale di recensioni appartenenti al topic "Aggiornamenti dell'app" negli anni

Topic	Year	period_words	HumanLabel
8	2018	star, opinion, support, expect, devic, rate	Aggiornamenti dell'app
8	2019	jett, joan, star, develop, softwar, save	Aggiornamenti dell'app
8	2020	star, review, premium, mode, haven, write	Aggiornamenti dell'app
8	2021	star, rate, minut, reason, review, content	Aggiornamenti dell'app
8	2022	star, korean, rate, review, reason, random	Aggiornamenti dell'app
8	2023	star, rate, premium, review, deserv, reason	Aggiornamenti dell'app
8	2024	star, rate, premium, money, deserv, lyric	Aggiornamenti dell'app

Figura 38 - Parole rappresentative del topic "Aggiornamenti dell'app" per anno

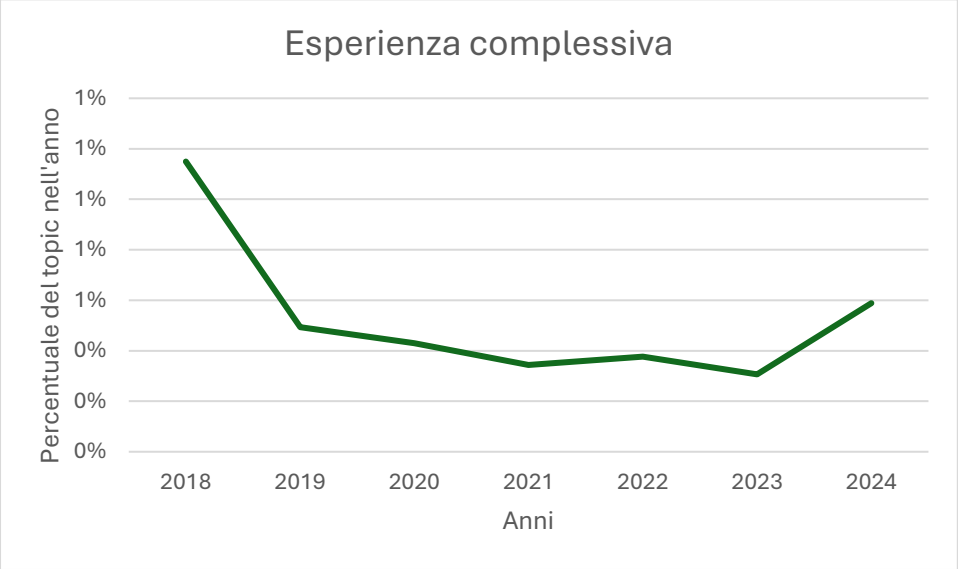

Figura 39 - Percentuale di recensioni appartenenti al topic "Esperienza complessiva" negli anni

Topic	Year	period_words	HumanLabel
10	2018	amaz, month, worth, premium, phone, search	Esperienza complessiva
10	2019	amaz, hassl, creat, premium, shuffl, fantast	Esperienza complessiva
10	2020	amaz, workout, clean, varieti, space, type	Esperienza complessiva
10	2021	amaz, podcast, premium, free, download, sound	Esperienza complessiva
10	2022	amaz, fave, podcast, instal, phone, origin	Esperienza complessiva
10	2023	amaz, premium, podcast, favorit, plan, skip	Esperienza complessiva
10	2024	amaz, premium, download, recommend, music, love	Esperienza complessiva

Figura 40 - Parole rappresentative del topic "Esperienza complessiva" per anno

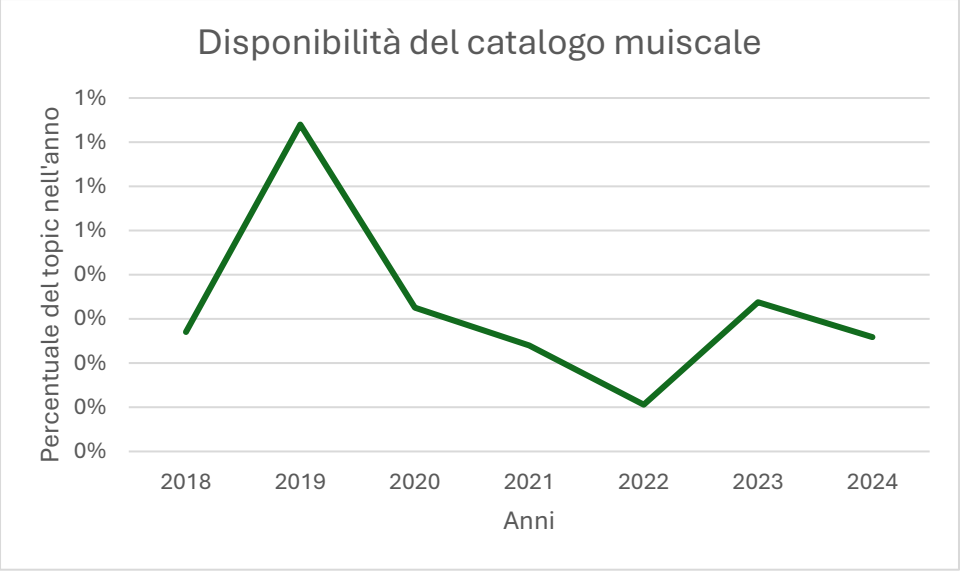

Figura 41 - Percentuale di recensioni appartenenti al topic "Disponibilità del catalogo musicale" negli anni

Topic	Year	period_words	HumanLabel
12	2018	languag, countri, live, user, chang, music	Disponibilità del catalogo musicale
12	2019	india, indian, collect, hindi, english, languag	Disponibilità del catalogo musicale
12	2020	india, punjabi, hindi, english, indian, languag	Disponibilità del catalogo musicale
12	2021	languag, hindi, intern, india, english, countri	Disponibilità del catalogo musicale
12	2022	chines, spanish, countri, english, languag, user	Disponibilità del catalogo musicale
12	2023	india, bollywood, indian, marathi, hindi, countri	Disponibilità del catalogo musicale
12	2024	telugu, hindi, indian, india, kannada, languag	Disponibilità del catalogo musicale

Figura 42 - Parole rappresentative del topic "Disponibilità del catalogo musicale" per anno