Università degli Studi di Milano - Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea in Informatica

Sulle relazioni tra notizie e trending social: Un'analisi descrittiva tra Google News e Twitter

Relazione di stage di:
Tommaso Redaelli
Matricola 830442

Anno Accademico 2023-2024

Abstract

Questo studio presenta un'analisi descrittiva volta a identificare le relazioni semantiche tra i contenuti pubblicati su Google News e i trending topics di Twitter. I dati sono stati raccolti tramite scraping dalle due piattaforme e sono stati impiegati tre diversi approcci di rappresentazione del testo: TF-IDF, Word2Vec e BERT. Le relazioni vengono estratte calcolando la similarità coseno, sia tra testi presi singolarmente, sia sfruttando l'aggregazione fornita dalle piattaforme. La visualizzazione dei collegamenti è stata realizzata mediante un diagramma di Sankey, si rende efficace per evidenziare le connessioni tra i contenuti. Infine, il lavoro discute i risultati ottenuti per i diversi eseperimenti, descrivendo le limitazioni degli approcci utilizzati e suggerendo possibili miglioramenti relativi. Questo studio rappresenta un tentativo di descrizione delle relazioni tra media tradizionali e social media ed offre un possibile punto di partenza verso successive analisi dei flussi informativi tra piattaforme diverse.

Indice

1 Introduzione **1**
1.1 Obiettivi e motivazione 1
1.2 Struttura del progetto 1

2 Descrizione dati **2**
2.1 Piattaforme analizzate 2
2.2 Processo di raccolta 2

3 Metodologia **4**
3.1 Algoritmi di rappresentazione del testo 4
3.2 Tecniche di preprocessing 4
3.3 Estrazione delle relazioni 6
3.3.1 Ottenimento degli embed 6
3.3.2 Analisi di similarità 6

4 Conclusioni **8**
4.1 Visualizzazione delle relazioni 8
4.1.1 Discussione dei metodi e relativi risultati 8
4.2 Limiti e miglioramenti futuri 10

1 Introduzione

1.1 Obiettivi e motivazione

Questo studio si pone l'obiettivo di analizzare e confrontare i contenuti notiziari con quelli virali pubblicati sui social media, al fine di individuare possibili relazioni tra gli argomenti trattati su questi due diversi canali di informazione. L'interesse verso la comparazione di questi nasce dalla crescente preoccupazione riguardo al fenomeno delle fake news, per cui le informazioni false o parzialmente false trovano terreno fertile sui social media, diffondendosi a una velocità molto superiore rispetto ai contenuti verificati [1] e pubblicati dai canali informativi tradizionali.

La motivazione principale di questa ricerca risiede nel fatto che la domanda di verifica delle informazioni è strettamente legata all'attualità degli argomenti trattati. Le notizie recenti e di grande rilevanza tendono a generare un elevato volume di discussioni sui social media, creando un terreno fertile per la diffusione di disinformazione. Pertanto, esplorare le relazioni tra i contenuti notiziari e quelli social può essere di aiuto per identificare rapidamente i temi su cui è più urgente concentrarsi per successivi processi di verifica della veridicità dell'informazione.

1.2 Struttura del progetto

Abbiamo scelto di confrontare i contenuti notiziari provenienti da Google News con quelli social ricavabili dalla sezione Trending di Twitter. La loro descrizione e quella dei dataset costruiti a partire da queste è discussa nel capitolo 2. Nel capitolo 3 vengono presentate la metodologia di preprocessing, di rappresentazione del testo e le tecniche con le quali si cerca di estrarre delle relazioni tra le due fonti. Nel capitolo 4 viene proposta una rappresentazione dei risultati ottenuti con relativa discussione di questi e descrizione delle limitazioni delle tecniche adottate e possibili conseguenti miglioramenti adottabili in futuro.

2 Descrizione dati

2.1 Piattaforme analizzate

La piattaforma da cui otteniamo i contenuti notiziari è Google News, in particolare la sezione dedicata all'Italia, che offre una panoramica delle notizie, aggregando contenuti da varie testate giornalistiche. La piattaforma da cui vengono ottenuti i contenuti social virali è Twitter, in particolare la sezione dedicata ai Trending Topics nel nostro paese che rappresentano gli argomenti più discussi sulla piattaforma in tempo reale, mostrando i temi che stanno attirando la maggiore attenzione del pubblico.

2.2 Processo di raccolta

Il processo di raccolta dei dati è stato eseguito utilizzando tecniche di web scraping, in particolare attraverso la libreria Selenium [2] combinata con ChromeWebDriver [3]. Questa metodologia ha permesso di automatizzare l'acquisizione dei dati sia da Google News che da Twitter. Di seguito vengono descritti i dettagli relativi alla raccolta sulle due diverse piattaforme:

1. **Google News:**
   Per i contenuti provenienti da Google News, vengono raccolti diversi dati rilevanti, tra cui la data della notizia, l'agenzia che ha pubblicato l'articolo, il link web e la headline, che è stata utilizzata come testo rappresentativo della notizia. Google News presenta le notizie in modo aggregato per argomento. Questo aspetto è stato registrato e conservato per un utilizzo successivo nelle analisi. Ogni processo di raccolta ha prodotto circa 50 blocchi di argomenti, ciascuno composto da 3 o 4 headline di diverse agenzie.

2. **Twitter Trending Topics:**
   Per i contenuti provenienti dalla sezione Trending Topics di Twitter, sono stati raccolti il nome del trend (spesso rappresentato da un hashtag), la data di pubblicazione del tweet, il link e il testo associato al tweet. Ogni sessione di raccolta ha ottenuto almeno 30 tweet per ciascuno dei 10 top-trend proposti dalla piattaforma.

Al termine della raccolta, queste sono le informazioni relative alle quantità dei dati ottenuti dal 2 Agosto 2024 al 17 Agosto 2024:

1. **Google News:**
    - 424 argomenti
    - 1693 singole headline
2. **Twitter Trending Topics:**
    - 80 trending topic
    - 2003 singoli tweet

Di seguito vengono mostrate le anteprime dei due dataset relativi rispettivamente alle notizie provenienti da Google News, figura 2.1, e ai tweet provenienti da Twitter Trending Topic, figura 2.2.

	text	date	agency
0	Strage di Bologna: Meloni, "Ingiustificati gli attacchi a me e al governo". Schlein: "Fa vittimismo, deplorevole"	2024-08-02T15:25:00Z	Agenzia ANSA
1	Strage di Bologna. Meloni: "Grave e pericoloso dire che le radici dell'attentato sono nella destra di governo". Bolognesi: "Finisca di fare la vittima"	2024-08-02T12:46:05Z	La Repubblica
2	Prarolo ricorda Rossella Marceddu, morta nella strage di Bologna: "Dolore scolpito nella memoria"	2024-08-02T15:26:00Z	La Stampa
3	La strage in stazione, quel dolore mobile e il dovere di ricordare	2024-08-02T04:00:00Z	BolognaToday
4	Roma, due ragazze stuprate in un'abitazione alla Borghesiana: arrestati gli aggressori, uno stava per scappare	2024-08-02T14:08:09Z	ilmessaggero.it
...	...	...	...
1688	Bus fuoristrada a Verona, autista è dipendente ditta bus Mestre	2024-08-16T10:49:00Z	Agenzia ANSA
1689	Suicidio dopo essere stato licenziato per 280 euro, la collega al funerale: «Sfruttati e sacrificati». Il prete la interrompe: «Non siamo al sindacato»	2024-08-17T04:30:09Z	Corriere della Sera
1690	Dipendente di Metro si toglie la vita dopo 27 anni di lavoro: era stato licenziato per un ammanco di 280 euro	2024-08-14T18:52:00Z	La Stampa
1691	Licenziato dopo 27 anni per un presunto danno all'azienda di 280 euro, dopo un mese si suicida: il caso a Marghera	2024-08-14T16:02:17Z	Open
1692	Licenziato dopo 27 anni per errore da 280 euro, dipendente modello si toglie la vita	2024-08-14T14:58:00Z	Adnkronos

Figura 2.1: Preview del dataset contentenente le headline provenienti da Google News.

	trend_name	text	date
0	#ImaneKhelif	Se leggo ancora una volta un'altra di quelle idiozie propagandistiche su #ImaneKhelif. \n\n#giochiolimpici #parigi2024 #olimpiadi	2024-08-02T09:03:03.000Z
1	#ImaneKhelif	Io ho la sindrome dell'ovaio policistico, produco quindi più testosterone del normale.\nHo la vagina e tutto il resto, ho pure partorito due figli. \n\nSono un uomo e non lo sapevo?\n\n#parigi2024 #GIOCHIOLIMPICI #ImaneKhelif	2024-08-01T13:03:30.000Z
2	#ImaneKhelif	I soliti decerebrati di una parte politica, lanciano in pasto all'odio una donna algerina (per chi non lo sapesse, in Algeria essere #LGBTQIA è reato), non conoscendo nulla, #ImaneKhelif è nata donna, ha la vagina, l'utero, produce solo tanto testosterone. Non è una #trans.	2024-08-01T06:53:37.000Z
3	#ImaneKhelif	Angela #Carini ha lasciato il ring olimpico senza salutare #ImaneKhelif. Un gesto che ha fatto discutere, ma che Emanuele Renzini, direttore tecnico della squadra italiana, ha spiegato ai nostri microfoni. Un racconto che chiarisce quanto avvenuto. #olimpiadi #paris2024	2024-08-01T14:05:15.000Z
4	#ImaneKhelif	La merda che si sta buttando addosso a #ImaneKhelif è qualcosa di ignobile.	2024-08-02T20:14:07.000Z
...	...	...	...
2259	Miretti	#Calciomercato #Juventus: possibile rinnovo per #McKennie. #Miretti potrebbe lasciare in quel caso: su di lui il #Genoa\n(@DiMarzio)\n#Transfers	2024-08-16T11:41:15.000Z
2260	McKennie	Il polacco ha preso la buonuscita, McKennie rinnova, gli altri esuberi un posticino lo trovano. L'unico babbo di minchia che si farà tutto l'anno in tribuna è lui.\n\nBabbo di minchia.	2024-08-16T09:26:41.000Z
2261	Maurizio Mosca	Non so perché sia in tendenza Maurizio Mosca, ma approfitto per postare questa perla indimenticabile	2024-08-17T08:26:38.000Z
2262	Maurizio Mosca	Ricomincia la #SerieA e c'è Maurizio Mosca in tendenza perché, si, un po' manca quella sua ingenuità e leggerezza nel seguire il calcio prima che i social (e pure le stesse TV) ne avvelenassero e ne polarizzassero il racconto...	2024-08-17T10:31:22.000Z
2263	Maurizio Mosca	Maurizio Mosca in tendenza. Lo ricordo come un animale mitologico, un minotauro metà giornalista e metà comico, con quel retrogusto del "ma ci è o ci fa". Un'immagine indelebile nella mia TV domenicale del prepartita alle ore 14 (si, un prepartita c'era. uno solo)	2024-08-17T12:22:01.000Z

Figura 2.2: Preview del dataset contentenente i tweet provenienti da Twitter Trending Topics.

3 Metodologia

3.1 Algoritmi di rappresentazione del testo

Per l'analisi comparativa dei contenuti notiziari e social, sono state adottate tre diverse tecniche di rappresentazione del testo:

1. **TF-IDF (Term Frequency-Inverse Document Frequency):**
   Misura l'importanza di una parola in un documento rispetto al corpus totale, combinando la frequenza del termine con l'inverso della frequenza del documento, per evidenziare i termini più rappresentativi di esso.

2. **Word2Vec — GloVe (Global Vectors for Word Representation) [4]:**
   Utilizza vettori continui per le parole, catturando relazioni semantiche basate sui contesti di utilizzo per i quali parole con contesti simili tendono ad avere vettori simili.

3. **BERT (Bidirectional Encoder Representations from Transformers) [5] [6]:**
   Modello basato su Transformer per generare rappresentazioni di testi, considerando sia il contesto precedente che successivo dei termini che li compongono, per catturare sfumature semantiche più profonde.

Queste tre tecniche di rappresentazione rappresentano tre diversi approcci all'estrazione del significato: TF-IDF si basa su una rappresentazione statistica, Word2Vec sfrutta relazioni semantiche tra le parole mentre BERT, e comunque la famiglia di LLM, è in grado di catturare contesti complessi e relazioni più profonde tra i termini, motivo per cui a differenza di Word2Vec, le rappresentazioni delle parole possono differire a seconda dei contesti in cui sono utilizzate.

3.2 Tecniche di preprocessing

Per le prime due tecniche di rappresentazione del testo (TF-IDF e Word2Vec), è stata eseguita una fase di preprocessing per pulire e preparare i testi conservandone solo gli aspetti più significativi. Le operazioni svolte sono le seguenti:

1. **Rimozione di codice HTML:** Eliminazione di eventuali tag HTML presenti nei testi.

Originale	**Preprocessed**
Morti marito e moglie: la loro Porsche si è schiantata contro la vetrina di un negozio	Morti marito moglie Porsche schiantato vetrina negozio
Violenze in carcere: la procura chiede 29 misure cautelari, il gip le rigetta	Violenze carcere Procura chiedere 29 misura cautelare gip rigettare
#2agosto è il 44° anniversario della strage alla stazione di #Bologna. #Meloni anche questa volta non sarà a Bologna. Il 2 agosto ha sempre di meglio da fare e non può mai partecipare al ricordo delle vittime della #strage fascista.	2agosto 44° anniversario strage stazione Bologna Meloni volta Bologna 2 agosto avere fare partecipare ricordo vittima strage fascista
#Paris2024 Purtroppo Beppe Grillo già eliminato nelle batterie dei 100 metri (me lo ricordavo più forte a nuoto).	Paris2024 Beppe Grillo eliminare batteria 100 metro ricordare forte nuoto

Tabella 1: Alcuni esempi di testi prima e dopo l'operazione di preprocessing. Le prime due righe sono headlines estratte da Goole News, la terza e quarta riga contengono invece testi estratti da Twitter.

2. **Rimozione di link:** Eliminazione di URL (non rilevanti nella definizione del contenuto).
3. **Rimozione di emoji:** Rimozione di emoji per mantenere il focus sul contenuto testuale.
4. **Rimozione di caratteri speciali:** Rimozione o sostituzione di caratteri speciali (ad esempio, "-" sostituito con uno spazio).
5. **Filtro sui PoS-Tag [4]:** Mantenimento di parole con tag grammaticali rilevanti (nomi propri, nomi, verbi, aggettivi, numeri), escludendo parole non significative come articoli e congiunzioni.
6. **Lemmatizzazione [4]:** Riduzione delle parole alla loro forma base per uniformare le varianti di uno stesso termine.

Nella tabella 1 vengono mostrati degli esempi di trasformazione del testo dopo l'applicazione di questa fase di preprocessing.

3.3 Estrazione delle relazioni

Di seguito viene descritto il processo di ottenimento degli embed testuali per entrambi le fonti di informazione attraverso i tre algoritmi descritti sopra. Da questi sarà poi possibile estrarre delle relazioni basate sulla similarità delle rappresentazioni.

3.3.1 Ottenimento degli embed

Per ciascuna delle tecniche di rappresentazione, abbiamo generato vettori che codificano i testi dei tweet e delle notizie attraverso le seguenti procedure:

- **TF-IDF:** Viene calcolata la matrice TF-IDF, quindi dai vettori riga di questa vengono estratte le rappresentazioni testuali corrispondenti a ogni tweet e headline. Poiché la dimensione della matrice TF-IDF è determinata dal numero di termini presenti nel vocabolario, viene applicata una decomposizione SVD per ridurre la dimensionalità dei vettori a uno spazio più gestibile, passando dalla dimensione iniziale del vocabolario (9581) a una dimensione ridotta pari a 512 consentendoci di mantenere la maggior parte delle informazioni rilevanti, semplificando al contempo il successivo calcolo delle similarità tra i testi.
- **Word2Vec:** Viene ricavato il vettore di ogni parola di ogni testo (300 dimensioni), quindi partendo da questi è ottenuto il vettore complessivamente rappresentativo dell'intero testo attraverso la media dei vettori delle parole che lo compongono.
- **BERT:** Dal modello BERT, estraiamo i vettori di embedding tramite il pooler-output layer, che fornisce una rappresentazione complessiva dell'intera frase attraverso una aggregazione delle informazioni estratte dalle parole di questa nel corso dei diversi strati del modello.

3.3.2 Analisi di similarità

I vettori ricavati al passaggio precedente rappresentano diversi criteri per confrontare e analizzare le similitudini tra le due fonti di informazioni. Il passaggio successivo è stato l'adozione di due metodologie distinte per esplorare potenziali relazioni tra i dati, motivate dalla necessità di cogliere le somiglianze sia a livello individuale che a livello di contenuti aggregati:

1. **Matrice di similarità basata sulla distanza coseno tra ogni tweet e ogni headline:**
   Questa tecnica permette di confrontare direttamente ciascun tweet con ogni notizia, calcolando il grado di somiglianza tra i vettori rappresentativi tramite

la distanza coseno. Questo approccio è utile per identificare somiglianze a livello di singoli contenuti, verificando quali tweet siano affini alle notizie.

2. **Matrice di similarità tra la rappresentazione media dei trending topics e la media degli argomenti delle notizie:**
   Questo secondo approccio si concentra su una rappresentazione più globale, calcolando la media dei vettori relativi a tutti i tweet associati a un determinato trend, e confrontandola con la media dei vettori delle notizie riguardanti uno specifico argomento. L'idea è quella di ridurre il rumore generato dalle singole variazioni nei tweet e nelle notizie, cercando di cogliere una somiglianza più a livello di temi generali. Questo metodo dovrebbe consentire di esaminare le somiglianze tra i macro-argomenti fornendo una visione più ampia di questi.

Entrambi gli approcci forniscono prospettive complementari: il primo permette un'analisi dettagliata e puntuale dei singoli contenuti, mentre il secondo aiuta a cogliere le tendenze più generali. Utilizzando poi una soglia arbitraria, possiamo determinare quando la somiglianza tra le rappresentazioni testuali di Twitter e quelle di Google News è sufficientemente alta da suggerire che le due fonti trattino argomenti simili.

4 Conclusioni

In questo capitolo vengono presentati i risultati ottenuti attraverso l'analisi delle relazioni tra i testi, utilizzando le tre tecniche di rappresentazione e le due metodologie di computazione della similarità.

4.1 Visualizzazione delle relazioni

Per la visualizzazione delle relazioni tra i dati, è stato impiegato il diagramma di Sankey, figure 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, che si è dimostrato particolarmente efficace nel rappresentare in modo intuitivo i collegamenti tra i tweet e le notizie. I colori dei collegamenti si riferiscono ai macro-topic relativi alle headline, e sono estratti tramite un classificatore basato su un modello BERT preaddestrato [7].

Il diagramma di Sankey, nella sua versione digitale, consente inoltre di esplorare i collegamenti in dettaglio, permettendo di visualizzare la forza delle relazioni individuate tra le diverse fonti. Con questa rappresentazione grafica, dovrebbe risultare intuitvo osservare come i singoli tweet siano collegati a specifiche notizie. Ogni flusso nel diagramma rappresenta una connessione tra un tweet e una headline nel caso di confronto singolo oppure tra un Trending Topic ed un argomento notiziario nel caso di confronto aggregato, con la larghezza del flusso che indica la forza della similarità e la possibilità di osservare quali dei tweet e delle headline sono più rappresentativi del collegamento. Inoltre, la colorazione dei collegamenti, aiuta a comprendere quali siano le tematiche più influenti nelle discussioni social.

4.1.1 Discussione dei metodi e relativi risultati

TF-IDF

La tecnica di rappresentazione basata su TF-IDF consente di individuare relazioni forti grazie alla frequenza dei termini, con collegamenti chiari e inerenti. Tuttavia, il limite principale di TF-IDF riesiede nel non considerare la semantica del testo, rendendo difficile individuare relazioni più complesse quando contenuti simili utilizzano termini diversi.

Sfruttanto l'aggregazione tematica fornita dalle piattaforme, si ottiene una rappresentazione più completa del vocabolario, aumentando la probabilità di identificare

collegamenti coerenti. Questo approccio sfrutta la complementarità dei testi riducendo il rischio di identificare relazioni casuali tra frasi che condividono termini uguali ma all'interno di contesti differenti.

Word2Vec

L'approccio basato su Word2Vec è efficace nel catturare relazioni semantiche, poiché i vettori delle parole vengono costruiti considerando il contesto in cui ogni parola appare, permettendo di cogliere somiglianze tra termini che non condividono lo stesso significato letterale, ma che appaiono in contesti simili. Per questo motivo Word2Vec può risultare più utile nel riconoscere relazioni concettuali tra testi che trattano lo stesso argomento, anche quando utilizzano un vocabolario diverso. Tuttavia, questo approccio ha dei limiti. Essendo basato sulla media dei vettori delle singole parole, Word2Vec perde parte della struttura sintattica e delle interazioni tra parole. Questo riduce la capacità di rappresentare accuratamente frasi più complesse o catturare relazioni dipendenti dall'ordine delle parole.

Inoltre, con l'approccio Word2Vec, poiché la rappresentazione di ciascun testo deriva dalla media dei vettori delle parole che lo compongono, esiste il rischio che, attraverso la rappresentazione aggregata, in particolare nel caso dei Tweet di un trending topic, la rappresentazione media vada ad appiattirsi eccessivamente. Questo fenomeno riduce la capacità del modello di preservare la diversità tra contenuti che, pur trattando argomenti diversi, finiscono per avere una rappresentazione troppo simile.

BERT embeddings

BERT ottiene collegamenti migliori rispetto ad altre tecniche di rappresentazione perché è progettato per comprendere l'intero contesto di una frase, analizzando ogni parola in relazione a tutte le altre parole nel testo. Questa capacità di catturare non solo il significato delle singole parole ma anche le interazioni tra di esse nel contesto della frase consentendo a BERT di costruire una rappresentazione complessiva più accurata del testo. In questo modo possiamo rilevare collegamenti semantici più profondi tra contenuti anche quando non vi è una corrispondenza diretta dei termini.

Tuttavia, proprio per questa sensibilità al contesto, bisogna prestare attenzione alle relazioni estratte. La similarità tra due testi può risultare elevata non perché trattano lo stesso argomento, ma anche perché gli argomenti possono essere simili o correlati. In questi casi, BERT potrebbe identificare connessioni tra contenuti che in realtà descrivono fatti o temi diversi che comunque appartengono ad un unico macro-argomento, portando a collegamenti fuorvianti. Questo richiede un'analisi

più attenta e una soglia critica nella valutazione della similarità, per evitare di considerare erroneamente correlazioni tra temi distinti.

4.2 Limiti e miglioramenti futuri

Quella presentata ad ora, può essere intesa come una prima analisi esplicativa del fenomeno, che descrive i pregi e difetti delle diverse tecniche adottabili. Tuttavia, emergono alcuni limiti che aprono la strada a potenziali miglioramenti futuri:

1. **Raccolta dati limitata a poche piattaforme:**
   Uno dei limiti principali riguarda la raccolta dati, che si è concentrata su una singola piattaforma di notizie e su un solo social network. Ampliando la raccolta a più piattaforme si potrebbe generalizzare meglio non solo sui contenuti, ma soprattutto anche sugli stili linguistici offrirendo una visione più rappresentativa e diversificata del fenomeno studiato.

2. **Aggiunta di metriche di popolarità dei contenuti social:**
   Ottenere informazioni supplementari, come le statistiche di popolarità (ad esempio, like, commenti, condivisioni) sui contenuti social potrebbe contribuire a identificare con maggiore precisione i contenuti più influenti.

3. **Considerazione di elementi speciali nei testi:**
   Prestare maggiore attenzione agli elementi distintivi nei testi social, come gli hashtag, potrebbe migliorare la capacità di estrarre relazioni semantiche e rafforzare i collegamenti esistenti. Gli hashtag non solo contribuiscono alla categorizzazione dei contenuti, ma spesso sono rappresentativi dell'intero significato e potrebbero aiutare a rilevare connessioni tra argomenti che altrimenti potrebbero non essere evidenti.

Questi miglioramenti possono rappresentare non solo un raffinamento dell'approccio adottato, ma anche contribuire a una comprensione più profonda e complessiva delle dinamiche tra social media e notizie, con un impatto di maggiore rilievo per eventuali analisi future dei flussi informativi.

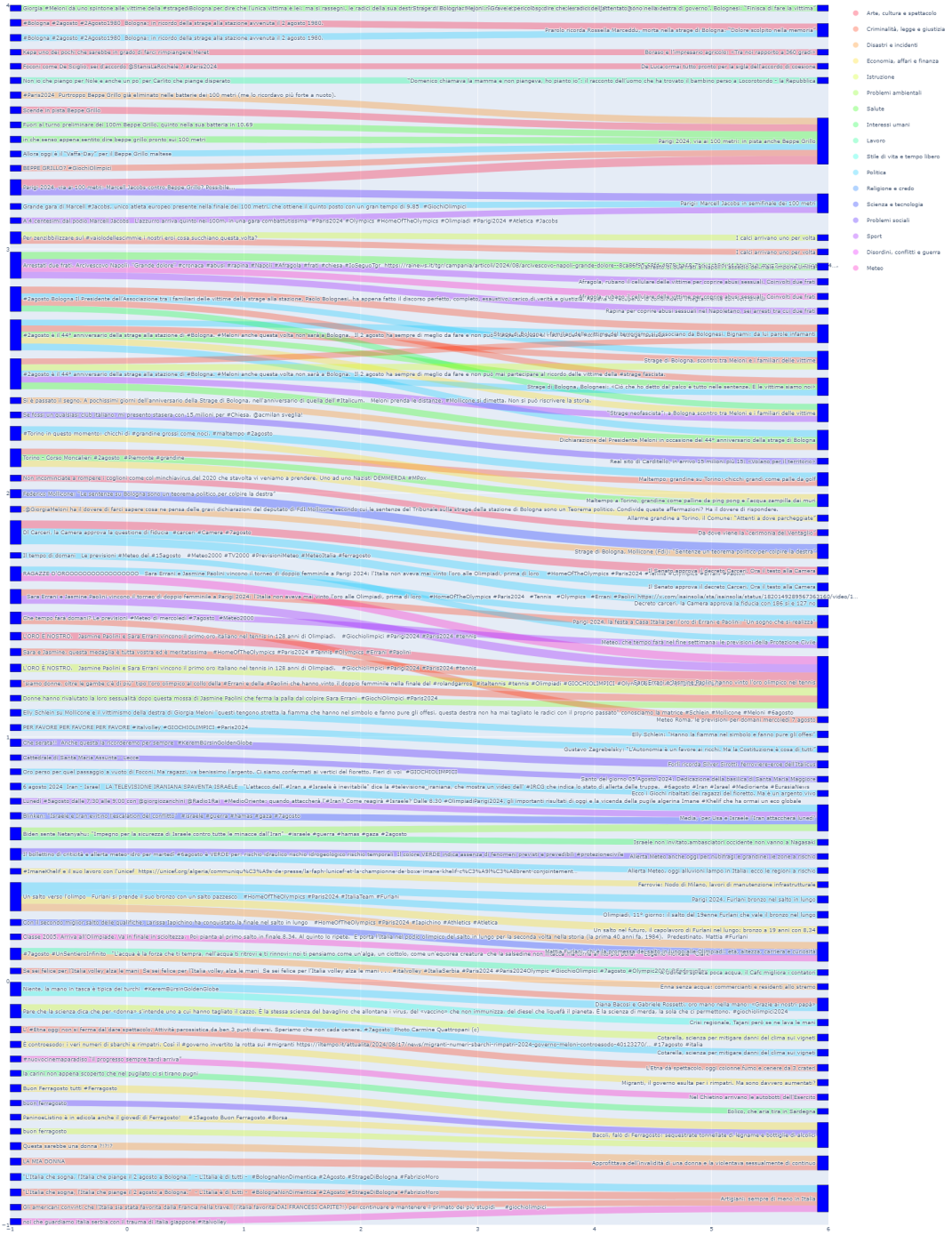

Figura 4.1: Top 100 relazioni ottenute attraverso rappresentazione TF-IDF e confronto di similarità singolo tra ogni testo. Threshold pari a 0.58

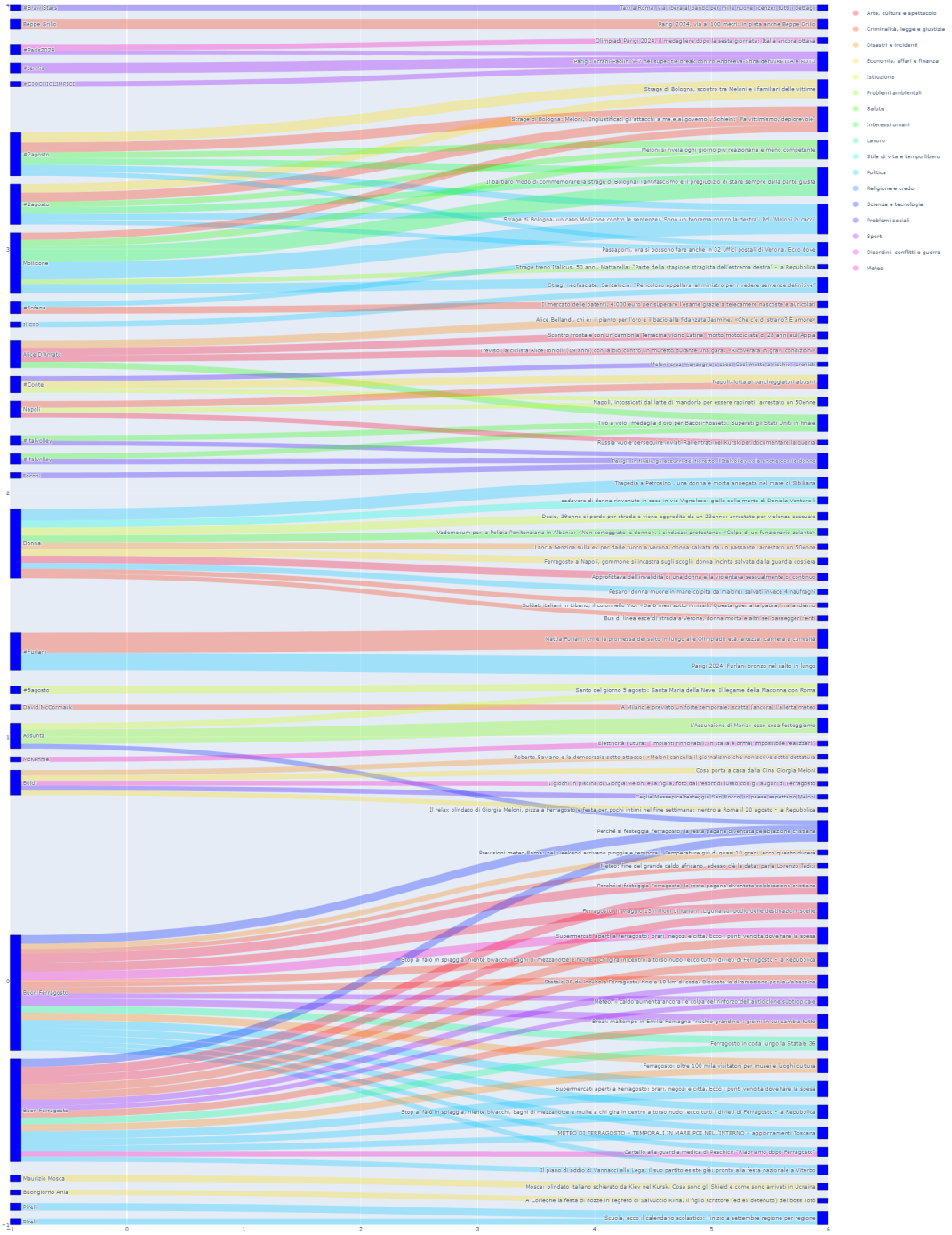

Figura 4.2: Top 100 relazioni ottenute attraverso rappresentazione TF-IDF e confronto di similarità in modo aggregato. Threshold pari a 0.17

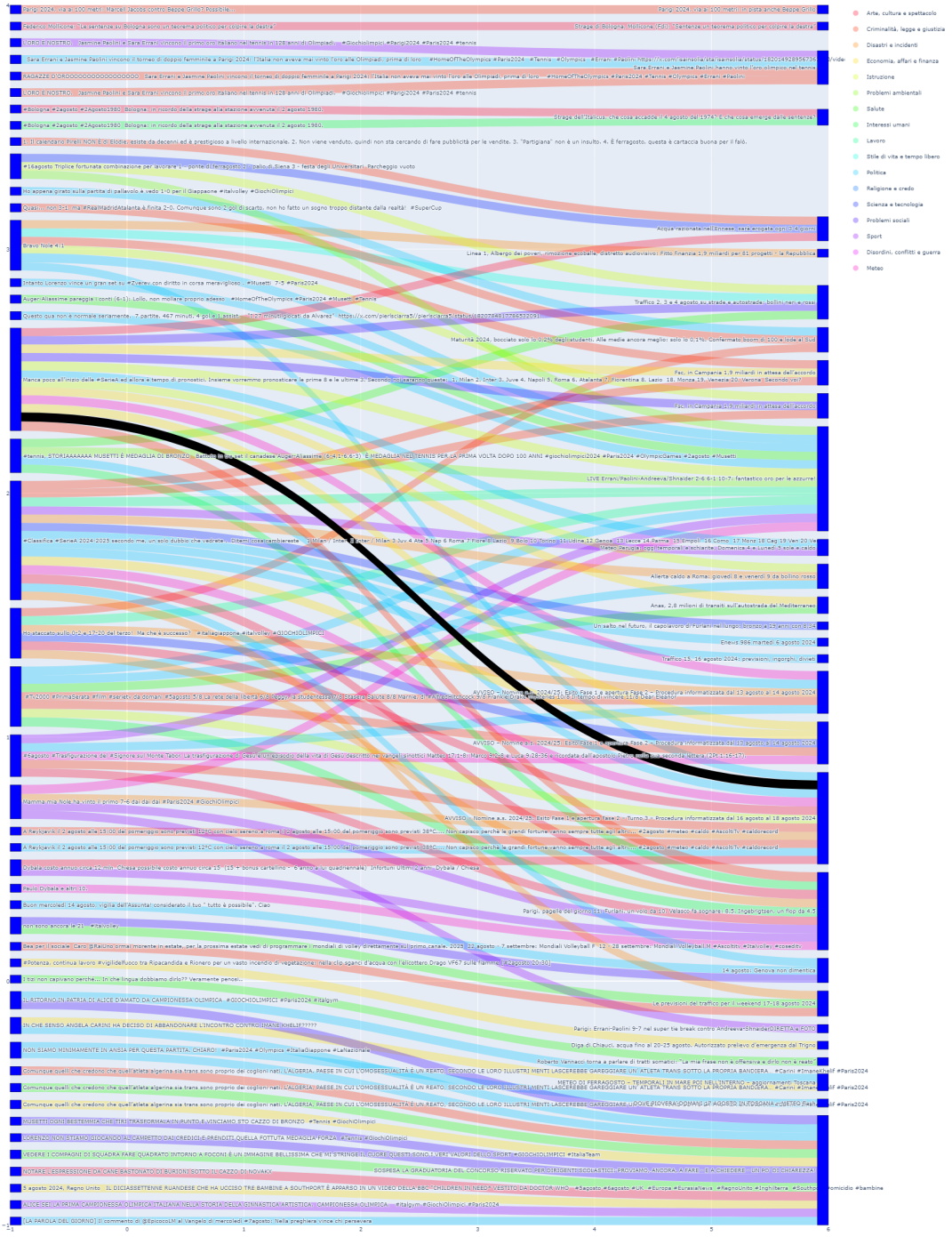

Figura 4.3: Top 100 relazioni ottenute attraverso rappresentazione Word2Vec e confronto di similarità singolo tra ogni testo. Threshold pari a 0.81

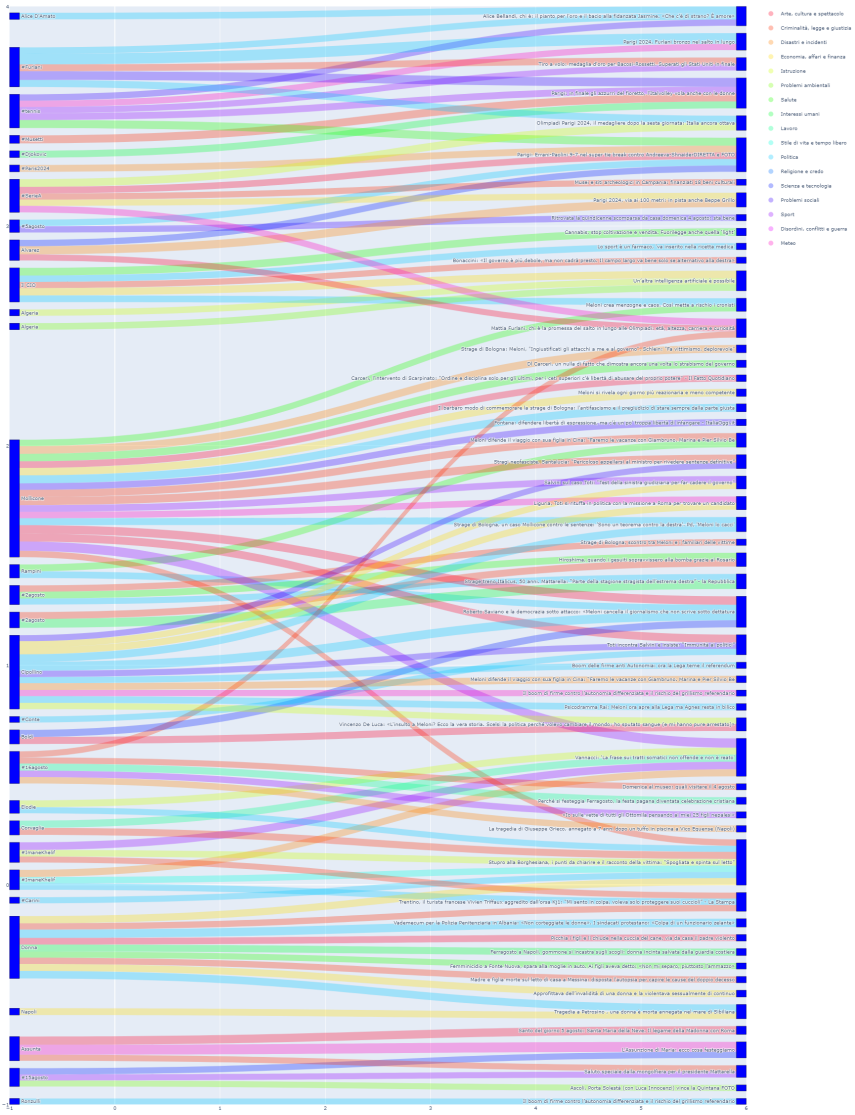

Figura 4.4: Top 100 relazioni ottenute attraverso rappresentazione Word2Vec e confronto di similarità in modo aggregato. Threshold pari a 0.42

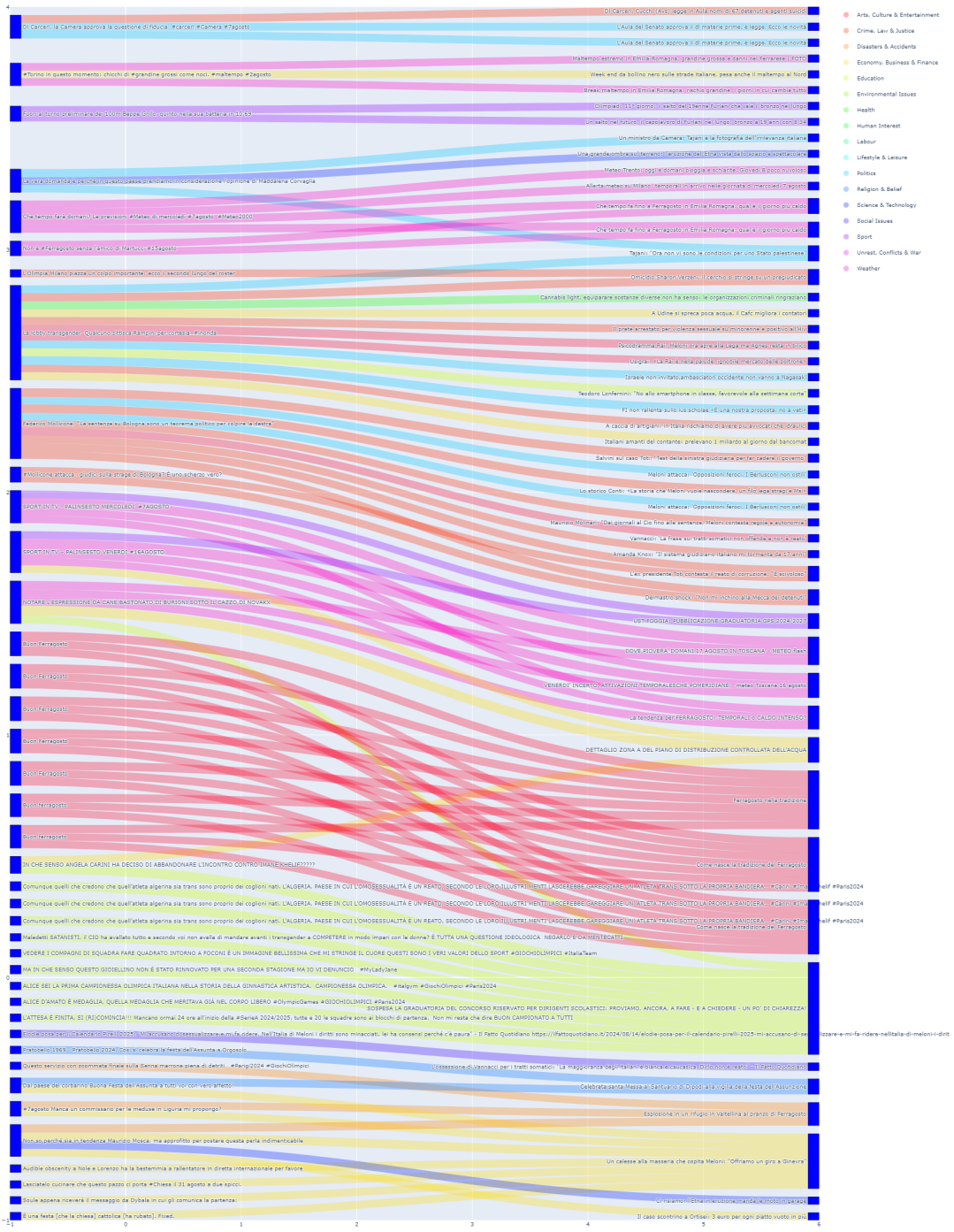

Figura 4.5: Top 100 relazioni ottenute attraverso rappresentazione tramite BERT embeddings e confronto di similarità singolo tra ogni testo. Threshold pari a 0.59

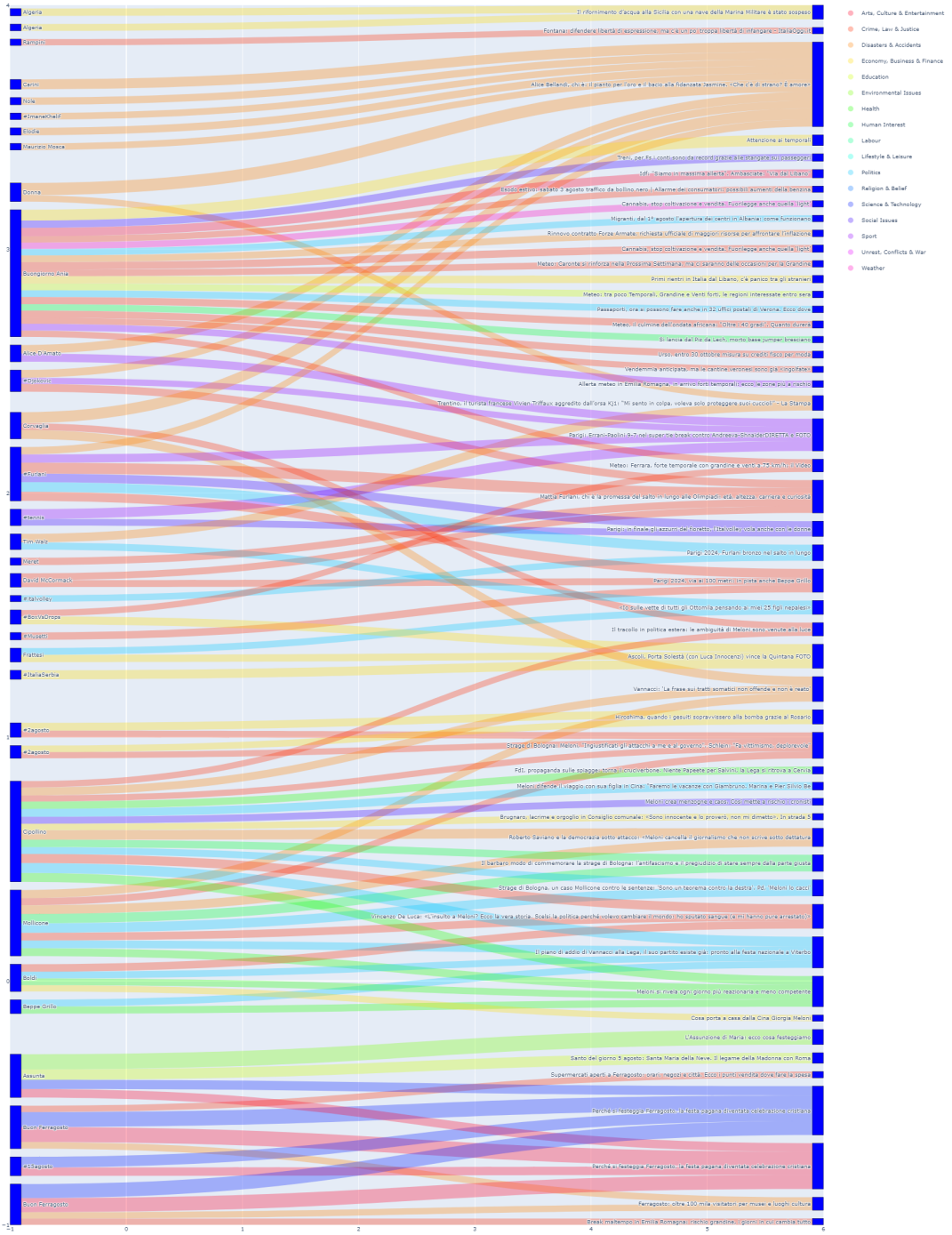

Figura 4.6: Top 100 relazioni ottenute attraverso rappresentazione tramite BERT embeddings e confronto di similarità in modo aggregato. Threshold pari a 0.20

Bibliografia

[1] Soroush Vosoughi, Deb Roy e Sinan Aral.
«The spread of true and false news online».
In: *science* 359.6380 (2018), pp. 1146–1151.

[2] *Selenium with Python 2014; Selenium Python Bindings 2 documentation — selenium python.readthedocs.io.*
`https://selenium-python.readthedocs.io/`. [Accessed 06-09-2024].

[3] *ChromeDriver - WebDriver for Chrome - Downloads — sites.google.com.*
`https://sites.google.com/chromium.org/driver/downloads`.
[Accessed 06-09-2024].

[4] *Italian · spaCy Models Documentation — spacy.io.*
`https://spacy.io/models/it#it_core_news_lg`. [Accessed 06-09-2024].

[5] Jacob Devlin. «Bert: Pre-training of deep bidirectional transformers for language understanding». In: *arXiv preprint arXiv:1810.04805* (2018).

[6] *google-bert/bert-base-uncased · Hugging Face — huggingface.co.*
`https://huggingface.co/google-bert/bert-base-uncased`.
[Accessed 22-08-2024].

[7] *nlpodyssey/bert-italian-uncased-iptc-headlines · Hugging Face — huggingface.co.* `https://huggingface.co/nlpodyssey/bert-italian-uncased-iptc-headlines`. [Accessed 06-09-2024].