Discussioni Italiane su Reddit: Analisi di Dinamiche e Sentiment

Repository Link: https://github.com/MassimoZarantonello2/RedditDebateAnalysis

Membri del gruppo:

- Pulerà Francesca 870005
- Zarantonello Massimo 866457

Indice

1. Descrizione del progetto

Il seguente progetto ha avuto l'obiettivo di analizzare la propensione degli italiani a discutere sotto post riguardanti vari temi di attualità (e non solo) e di caratterizzare la natura di queste discussioni. A tal fine, è stato creato un dataset tramite l'API di Reddit, estraendo tutti i commenti presenti sotto 9 post selezionati manualmente appartenenti al subreddit "italia".

Dopo aver definito per induzione il concetto di "discussione" sotto un post, è stata condotta un'analisi tramite grafi per valutare la frequenza e l'intensità delle discussioni relative ai post selezionati. L'analisi si è focalizzata sull'importanza dei nodi, ossia sulla centralità e il grado di connessione dei partecipanti alla discussione.

Successivamente, è stata effettuata un'analisi del sentiment per caratterizzare il tono delle discussioni identificate. L'analisi ha permesso di comprendere se i commenti fossero positivi, ad esempio elogiando la bellezza dei gatti, o negativi, come nel caso di insulti legati a visioni politiche divergenti.

Infine, è stata condotta un'analisi dei risultati per interpretare i dati raccolti e trarre conclusioni sulle dinamiche delle discussioni online.

I risultati ottenuti forniscono una panoramica preliminare sulla dinamica delle discussioni online in Italia, rivelando le tendenze comportamentali degli utenti e la natura delle interazioni sotto post di vario genere. Sebbene le considerazioni derivanti dalle analisi condotte non possano essere generalizzate a causa del numero limitato di post analizzati, il progetto ha posto le basi per estendere l'analisi a un numero maggiore di dati in futuro. Questo consentirebbe di fare affermazioni statisticamente significative, come ad esempio la tendenza degli italiani a discutere di più sotto post di calcio rispetto a quelli sulla violenza di genere, un dato che meriterebbe una riflessione approfondita.

1.1. Obiettivi

Si riassumono di seguito i principali obiettivi:

1. Creazione di un dataset specifico: realizzare un dataset a partire dalla raccolta dei commenti su Reddit di interesse. Verranno effettuate analisi preliminari per definire le proprietà fondamentali del dataset, essenziali per i successivi task.
2. Definizione del concetto di discussione: pensare e formalizzare una definizione di discussione, necessaria per l'analisi dei grafi. Questa definizione permetterà di identificare e isolare le diverse discussioni presenti nei commenti.

3. Analisi delle dinamiche di discussione: utilizzare tecniche di analisi dei grafi per individuare e analizzare le discussioni. Creare "sotto grafi" (con relativi dataset) per evidenziare le proprietà specifiche di tali discussioni.
4. Sentiment analysis: analizzare il sentiment e le emozioni prevalenti in ciascuna discussione. Confrontare i sentimenti legati a temi attuali e rilevanti, come politica e attualità, con quelli di temi più leggeri, come il calcio o i gatti.
5. Interpretazione dei risultati: interpretare i dati raccolti e trarre conclusioni sulle dinamiche delle discussioni online, fornendo una panoramica delle tendenze comportamentali degli utenti italiani su Reddit.
6. Fondamenti per future analisi: porre le basi per estendere l'analisi a un numero maggiore di post in futuro, consentendo di formulare affermazioni statisticamente significative sulle tendenze delle discussioni degli italiani su Reddit.

1.2. Struttura della relazione

La relazione sarà strutturata in modo da fornire una panoramica completa del progetto, includendo una descrizione dettagliata della metodologia utilizzata, i risultati ottenuti e le future direzioni della ricerca.

2. Dataset

Il processo di creazione del dataset ha avuto inizio con la selezione dei post di interesse, concentrati su diversi temi di attualità che spaziano dalla guerra tra Palestina e Israele alla violenza di genere. Questa scelta è stata motivata dall'interesse e dall'importanza che riteniamo rivestano tali argomenti, considerati estremamente attuali.

Abbiamo anche voluto includere post su argomenti più leggeri per poterli confrontare con quelli più seri e quindi per comprendere meglio cosa tende a suscitare maggiore discussione tra gli italiani. Questo approccio ci permette di esplorare se gli utenti dedicano più tempo a commentare tematiche leggere o questioni più profonde.

Inoltre, per affrontare il task di analisi del sentiment e comprendere il carattere delle discussioni generate sotto ciascun post, era essenziale avere commenti che esprimessero emozioni chiare e manifestassero un sentimento evidente.

2.1. Selezione dei post

Sono stati selezionati 8 differenti post (contenenti ciascuno dai 400 agi 600 commenti circa) relativi ai seguenti topic:

- Misoginia
  - https://www.reddit.com/r/Italia/comments/17z2hci/ho_visto_troppi_post_sulla_giulia_cecchettin/
- Surriscaldamento globale
  - https://www.reddit.com/r/Italia/comments/1bulhj9/il_problema_del_caldo_anomalocausato_dal/
- Guerra Palestina-Israele
  - https://www.reddit.com/r/Italia/comments/17lese9/4000_bambini_morti_in_palestina_possibile_che/
- Guerra Ucraina-Russia
  - https://www.reddit.com/r/Italia/comments/1b6cg4q/possibile_guerra_nato_russia_se_davvero_dovessero/
- Elezioni europee
  - https://www.reddit.com/r/Italia/comments/1cwqkqe/chi_voterete_alle_europee/
- Farina di grilli
  - https://www.reddit.com/r/Italia/comments/10v8sey/natura_s%C3%AC_niente_cibo_a_base_di_insettisintetici/
- Fuga di cervelli

    - https://www.reddit.com/r/Italia/comments/1d5h5h6/boom_di_giovani_in_fuga_dallitalia_sono_oltre_un/
- Gatti
    - https://www.reddit.com/r/Italia/comments/19aeo2k/eccovi_il_mio_gatto/

- Fedez e Ferragni
    - https://www.reddit.com/r/Italia/comments/197vo6o/altre_nubi_sui_ferragnez_settimana_dopo_settimana/

2.2. Estrazione dei commenti tramite API

Per estrarre i commenti dai reddit sopraelencati, si è scelto di utilizzare PRAW (Python Reddit API Wrapper), una libreria Python che fornisce un'interfaccia semplice e intuitiva per interagire con l'API di Reddit.

Essa è progettata per facilitare lo sviluppo di applicazioni che necessitano di accedere ai dati su Reddit, come l'estrazione di post, commenti, informazioni sugli utenti e altre operazioni consentite dall'API di Reddit.

2.3. Creazione del Dataset

Il dataset generato presenta la seguente struttura:

- Numero totale di istanze:
- Numero totale di attributi:
- Attributi:
    - Id del commento -> *comment_id*
    - Id dell'autore del commento -> *comment_author_id*
    - Nome dell'autore del commento -> *comment_author_name*
    - Id del commento genitore (a che commento sta rispondendo il soggetto in questione) -> *comment_parent_id*
    - Nome dell'autore del commento genitore -> *comment_parent_name*
    - Score del commento -> *comment_score*
    - Risposte al commento -> *comment_replies*
    - Data e ora del commento (in formato) -> *comment_posted_time*
    - Testo del commento -> *comment_body*
    - Id del post sotto cui è stato lasciato il commento in questione -> *post_id*
    - Titolo del post sotto cui è stato lasciato il commento in questione -> *post_title*
    - URL del post -> *post_url*
    - Id del subreddit del post -> post_subreddit

```
comment_id,comment_author_id,comment_author_name,comment_parent_id,comment_parent_name,comment_score,comment_replies,comment_posted_time,comment_body,post_id,post_title,post_url,post_subreddit
```

```
k9xjbzd,deleted,deleted,0,OP,310,3,1700425648.0,"Non è un idiota l'ex ragazzo, è un assassino, punto.",17z2hci,Ho visto troppi post sulla Giulia Cecchettin,https://www.reddit.com/r/Italia/comments/17z2hci/ho_visto_troppi_post_sulla_giulia_cecchettin/,t5_2rbm5

k9zies2,e5pmjzs5c,MainDelay9804,0,OP,22,0,1700459454.0,Sono sti ragazzetti sui social che ormai han perso la testa e ogni tipo di personalità e o fanno i cinici finti redpillati o le suffragette a tempo perso ormai ogni cosa nel mondo è A o B e sui social i ragazzetti ci vanno dietro senza pensare.,17z2hci,Ho visto troppi post sulla Giulia Cecchettin,https://www.reddit.com/r/Italia/comments/17z2hci/ho_visto_troppi_post_sulla_giulia_cecchettin/,t5_2rbm5

k9xb0xr,utwr8mto,AlessandroIT,0,OP,83,1,1700422472.0,"Non fatevi influenzare dai post, continuate ad essere uomini civili e dimostratelo. Le mele marce ci saranno SEMPRE",17z2hci,Ho visto troppi post sulla Giulia Cecchettin,https://www.reddit.com/r/Italia/comments/17z2hci/ho_visto_troppi_post_sulla_giulia_cecchettin/,t5_2rbm5
```

Figura 1: Rappresenta tre esempi di istanze del dataset in questione

Il numero limitato di post di Reddit analizzati è stato determinato dal breve periodo a disposizione e dai limiti di richieste imposti dall'API selezionata. Nonostante queste restrizioni, il dataset generato risulta significativamente ampio, comprendendo un totale di 8 istanze.

3. Grafi

3.1. Creazione dei Grafi

La fase iniziale nella creazione dei grafi prevede la trasformazione del dataset contenente i commenti scaricati da Reddit. Data la struttura del dataset, la generazione dei nodi e degli archi del grafo è un processo diretto. Ogni nodo è caratterizzato dai seguenti parametri:

- name: identificativo univoco per riconoscere il nodo all'interno del grafo.
- label: nome dell'utente associato al nodo; possono esserci ripetizioni e il label è usato solo a scopo di visualizzazione.
- post_id: ID univoco del post a cui il commento si riferisce.
- color: colore assegnato al nodo.

Analogamente, anche gli archi del grafo (edge) hanno dei parametri specifici:

- source: nodo di origine dell'arco, ossia l'utente che ha pubblicato il commento.
- target: nodo di destinazione dell'arco, ossia l'utente o il post a cui è diretto il commento.

Questo metodo permette di creare un grafo che rappresenta lo scambio di commenti avvenuto sotto ogni post. Il processo viene ripetuto per ogni post scaricato, e il grafo complessivo viene visualizzato in un plot.

Si osserva che esistono molti archi orientati tra le stesse coppie di nodi, il che è cruciale poiché indica che i due nodi hanno avuto un dibattito o, in generale, numerose interazioni.

3.2. Pulizia del Grafo

Durante la creazione dei grafi, ci sono due nodi particolari presenti in ogni grafo, caratterizzati dalle seguenti etichette:

- **OP (Original Post**): rappresenta il post originale sotto il quale sono stati scritti tutti i commenti. Poiché non è un utente vero e proprio, non può essere lasciato nel grafo. Tuttavia, molti commenti fanno riferimento a questo post principale.
- deleted: rappresenta commenti che sono stati cancellati dall'utente che li ha pubblicati o dai moderatori di Reddit per violazioni delle linee guida della community. Questi commenti non vengono utilizzati in questo progetto, poiché è impossibile risalire agli utenti e includerli nei dibattiti.

Rimuovendo questi nodi "placeholder", il grafo risulta composto esclusivamente da utenti che hanno effettivamente interagito tra loro.

Nell'ambito di questa ricerca, che mira ad analizzare la natura dei dibattiti online, è necessaria un'ulteriore fase di eliminazione dei nodi. Prima di qualsiasi analisi, vengono identificati ed eliminati i nodi che possono essere scartati immediatamente, ossia tutti quegli utenti di livello top-level che non hanno partecipato a nessuna interazione con altri utenti. Questi nodi rappresentano utenti che hanno commentato il post principale senza ricevere risposte da altri utenti. Sebbene tali commenti possano esprimere un'opinione e sia possibile calcolarne il sentiment, non rappresentano una parte attiva del dibattito.

Dal punto di vista del grafo, tutti i nodi che non hanno nessun vicino vengono immediatamente eliminati. Questo porta a una significativa riduzione del numero di nodi nei grafi, semplificando le fasi successive di elaborazione.

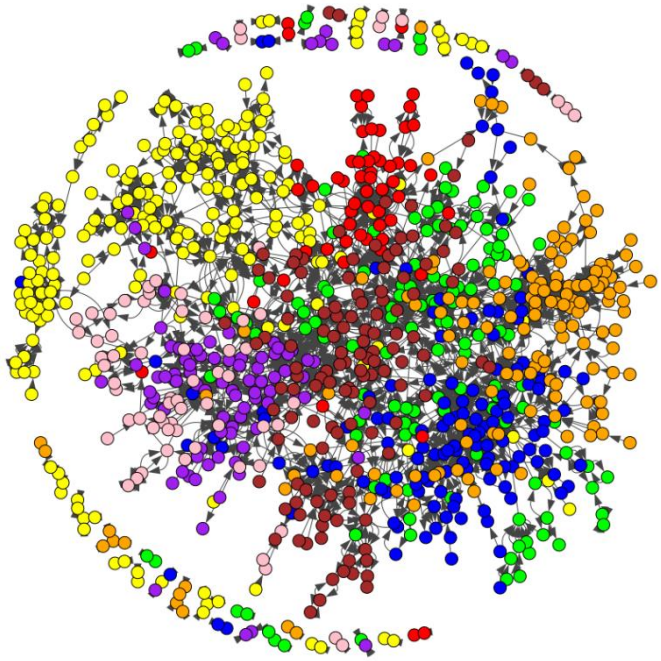


Figura 2: Rappresenta il grafo contenente tutti i commenti, il colore è stato assegnato sulla base del post

3.3. Creazione dei sottografi relativi ai post

Una volta generato il grafo contenente tutti i commenti relativi ad un post, è possibile osservare come i vari utenti interagiscono tra loro. Tuttavia, per gestire un numero maggiore di post, è probabile che lo stesso utente abbia partecipato a più discussioni sotto post diversi. Per isolare meglio le interazioni degli utenti e le relative discussioni, è stato deciso di separare il grafo principale nei suoi sottografi relativi ai vari post.

In questo modo, è possibile osservare con maggiore chiarezza il tipo di interazioni che si verificano tra i nodi.

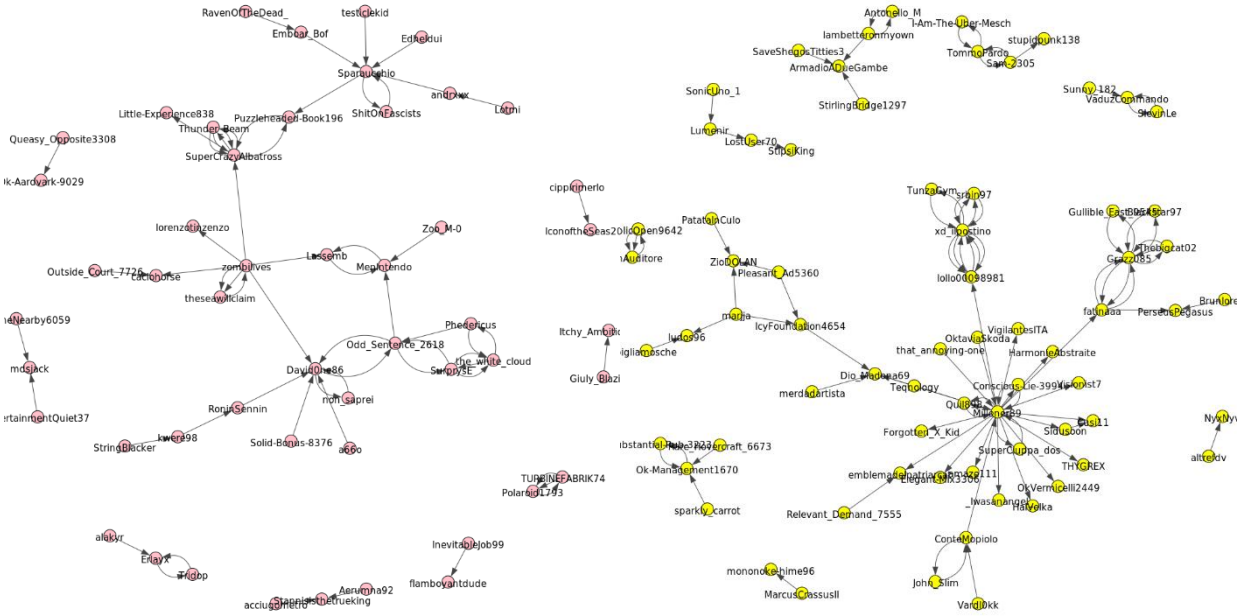


Figura 3,4: Rappresentano 2 grafi relativi ai commenti di 2 specifici post, sinistra post relativo ai Ferragnez, destra post sui gatti

3.4. Ricerca delle Discussioni

Una volta ottenuti i grafi dei singoli post, è possibile iniziare una fase di ricerca delle discussioni. Dal dizionario:

> *Esame approfondito di una questione, da parte di due o più persone che espongono ciascuna le proprie vedute (al livello di conversazione o a quello di dibattito): aprire, intavolare, rinviare, chiudere la discussione.*

Questa definizione, per quanto ben formulata, non è sufficiente per i nostri scopi. Infatti, risulta complesso creare un algoritmo su misura che riesca a individuare in maniera precisa una discussione a causa delle sue molteplici sfaccettature. Le discussioni non sono entità concrete: possono variare enormemente in forma, tono, contenuto e struttura. Possono manifestarsi attraverso scambi di singoli messaggi brevi o lunghe sequenze di commenti, essere esplicite o implicite, e divergere in vari sottotemi.

Per superare queste difficoltà, è stata sviluppata una definizione operativa di "dibattito" per mezzo dell'induzione. Utilizzando questa definizione e applicandola ai grafi relativi ai post, è possibile isolare le varie conversazioni che rispettano tali criteri.

Caso Base:

Consideriamo un nodo $n \in V$.

1.

$$InDegree(n) = 0 \bigwedge OutDegree(n) = 0 \rightarrow \text{n non è una discussione}$$

2.

$$InDegree(n) = 0 \bigwedge OutDegree(n) \geq 0 \rightarrow \text{n non è una discussione}$$

3.

$$\exists\, m \in V, n \neq m, (n,m) \in E, (m,n) \in E \Rightarrow \{n,m\} \rightarrow \text{ è una discussione}$$

Passo Ricorsivo:

Supponiamo $D \in V$ sia una discussione:

$$Se\ u \in V \backslash D\ e\ (u,v) \in E\ con\ v \in D, allora\ D' = D \cup \{u\}\ \text{è}\ una\ discussione$$

Utilizzando questa definizione, possiamo iterare attraverso i grafi dei singoli post per identificare e isolare le discussioni. Ogni volta che troviamo un nodo n che soddisfa i criteri del caso base o che può essere aggiunto a un insieme D già identificato come discussione, possiamo considerarlo parte di una conversazione. Questo approccio sistematico permette di tracciare e analizzare le discussioni in maniera più rigorosa e definita, facilitando l'analisi delle interazioni tra gli utenti.

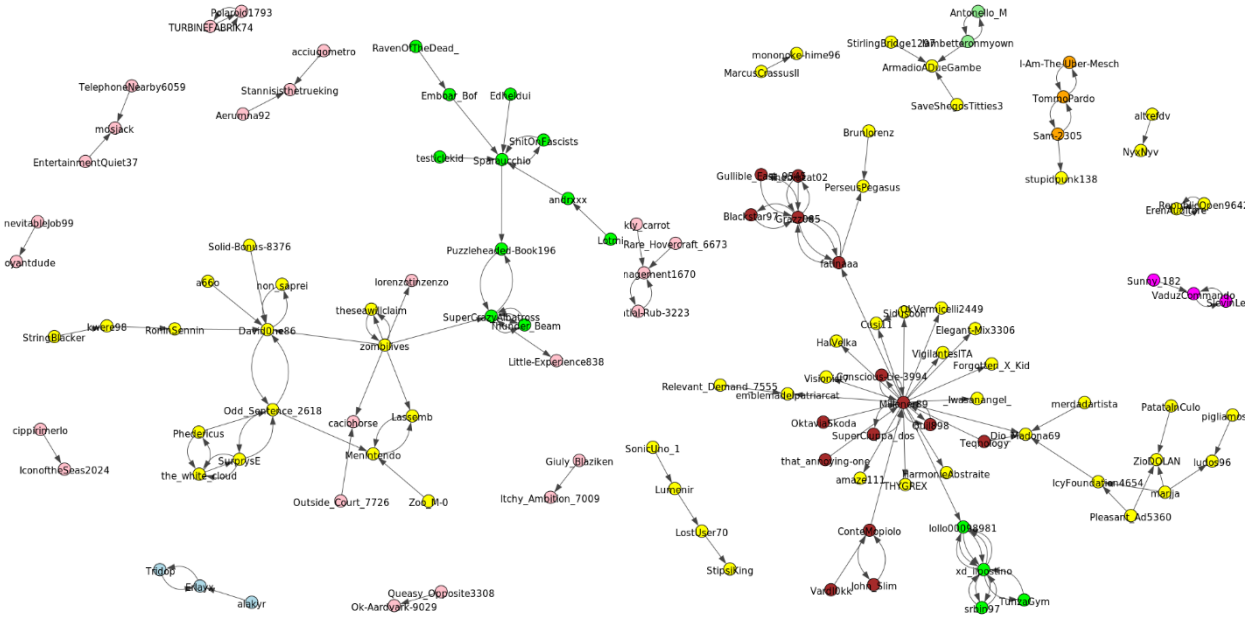


Figura 5,6: Rappresentano 2 grafi relativi ai commenti di 2 specifici post, sinistra post relativo ai Ferragnez, destra post sui gatti nella quale sono stati colorati i vari dibatti

3.5. Creazione del nuovo dataset

A questo punto, i dibattiti relativi ai vari post sono stati isolati e ora è possibile creare un nuovo dataset composto esclusivamente dai nodi che compongono questi dibattiti. Questo elimina

la necessità di eseguire il pre-processing dei grafi ogni volta, consentendo di fare riferimento direttamente a questo dataset consolidato.

Il dataset è strutturato come segue:

- **comment_id**: id univoco del commento
- **post_id**: post sotto la quale è stato scritto il commento
- **debate_group**: un id univoco che identifica all'interno di un post la discussione alla quale il commento fa parte
- **comment_user_name**: nome dell'utente che ha scritto il commento
- **commented_user_name**: nome dell'utente verso la quale il commento era indirizzato
- **comment_body**: il testo del commento
- **comment_score**: il punteggio che è stato assegnato al commento da altri utenti

comment_body	post_title	debate_group
Secondo me ha a che fare con il fatto che la maggior parte della gente è scettica in questo ambito, e quindi non vuole avere a che fare con cibo a base di insetti.	Natura sì: niente cibo a base di insetti/sintetici. Non rientrano nella nostra definizione di salute	1
il cliente da lei chiamato non è al momento raggiungibile...	Possibile guerra NATO - RUSSIA. Se davvero dovessero succedere e vi chiamassero al fronte per combattere ...	2
No way. Fino a qualche anno fa i ragazzi per non emigrare venivano a Milano. Ora che Milano è una cloaca, un misto tra un suq e Gotham city, andarsene via e l'unica scelta possibile. Non c'è volontà politica di sistemare la questione. Chi ci governa ha già nuova mano d'opera a basso costo da impiegare nelle loro realtà.	Boom di giovani in fuga dall'Italia, sono oltre un milione	6

Figura 7: estratto del dataset creato

È importante notare che i commenti associati a debate_group = 0 non appartengono a nessun dibattito specifico. Tuttavia, non sono stati eliminati durante la fase iniziale di pre-processing poiché contengono comunque informazioni rilevanti. Pertanto, sono stati conservati nel dataset.

Questo dataset consolidato permette di concentrarsi direttamente sull'analisi dei dibattiti senza dover ricreare i grafi da zero ogni volta, semplificando così le fasi successive di elaborazione e analisi dei dati.

3.6. Calcolo delle metriche sul grafo

Una volta ottenute tutte le discussioni sotto i vari post, diventa evidente che queste conversazioni variano significativamente sia in termini di dimensione che di densità. Questa variazione è cruciale quando si affronta la fase successiva di sentiment analysis, poiché non tutte le discussioni possono essere considerate allo stesso modo. Una breve interazione tra

due utenti non può avere lo stesso peso di una discussione più estesa coinvolgente decine di utenti con centinaia di messaggi scambiati.

Per affrontare questa sfida, è stato scelto di calcolare delle metriche sui grafi dei post e dei dibattiti al fine di stimare l'importanza relativa di ciascuna discussione. Questo approccio permette di valutare il livello di partecipazione degli utenti e il grado di interazione all'interno di ogni dibattito, fornendo una base più accurata per il calcolo del sentiment generale delle discussioni.

Le metriche selezionate per valutare l'importanza delle discussioni sono le seguenti, sia per il livello del post che per il singolo dibattito:

1. **Numero di nodi normalizzato**: rappresenta il numero di nodi che appartengono a un dibattito, diviso per il numero totale di nodi nel grafo del post. Questo fornisce una stima della partecipazione degli utenti rispetto alla totalità delle discussioni del post.

2**. Numero di archi normalizzato**: indica il numero di archi che appartengono a un dibattito, diviso per il numero totale di archi nel grafo del post. Questo aiuta a valutare il livello di interazione tra gli utenti all'interno del dibattito rispetto all'intero contesto del post.

3. **Grado medio dei nodi**: è l'average degree calcolato per ciascun dibattito, che rappresenta il numero medio di connessioni (archi) che ogni nodo ha all'interno del dibattito. Questa metrica offre una visione del livello di interazione media di ogni utente all'interno della discussione.

4. **Coefficiente di clustering medio**: indica quanto i vicini di ogni nodo sono interconnessi tra di loro. Un coefficiente di clustering più alto indica una maggiore coesione tra i vicini di un nodo, suggerendo una discussione più focalizzata o coesa.

Oltre a queste metriche, è stato anche creato un grafo che rappresenta la distribuzione dei gradi all'interno di ogni post. Questo grafico permette di visualizzare quanti nodi condividono lo stesso grado di connessione, identificare eventuali hub (nodi molto connessi) e spoke (nodi meno connessi), e ottenere una panoramica più dettagliata del livello di interazione degli utenti.

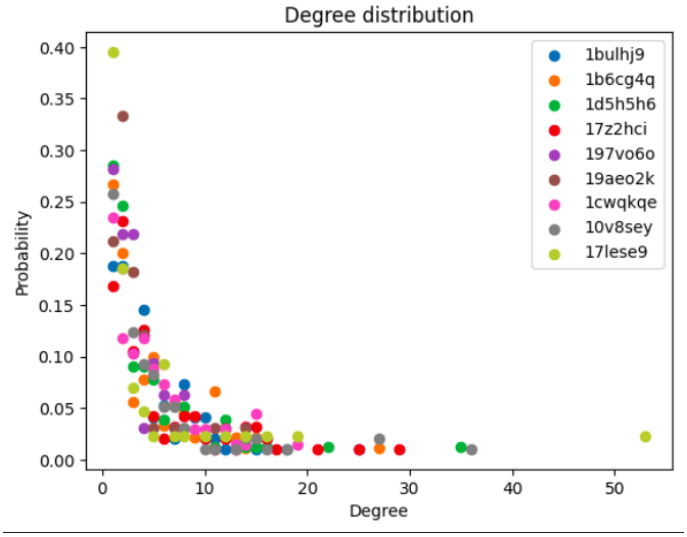


Figura 7: Rappresenta in un grafo la distribuzione di grado dei nodi di tutti i post

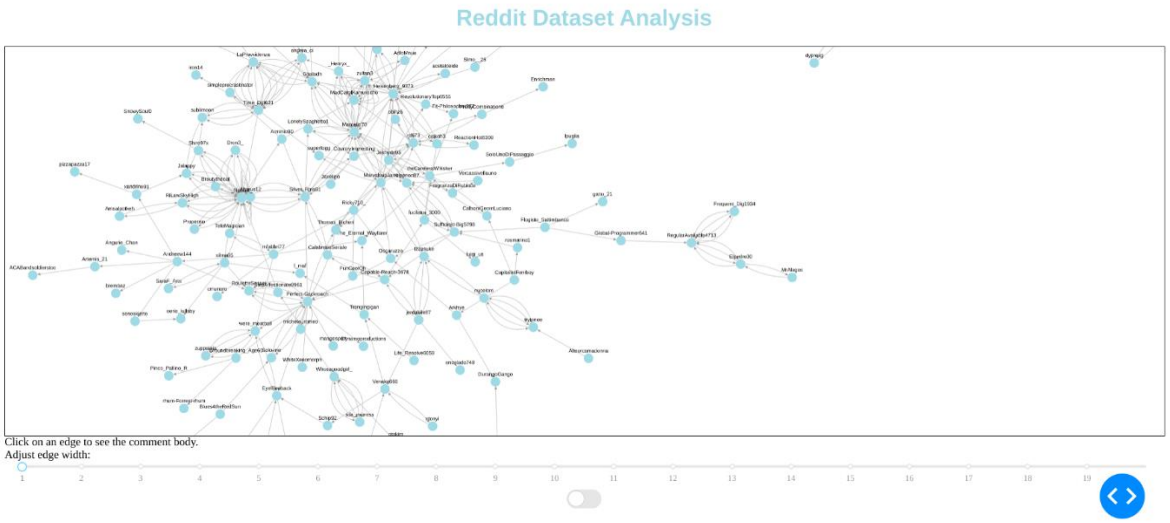
Questo approccio integrato di metriche sui grafi dei post e dei dibattiti fornisce una base analitica solida per comprendere meglio la dinamica e l'importanza delle discussioni online, facilitando una sentiment analysis più accurata e informativa.

3.7. Creazione della Demo

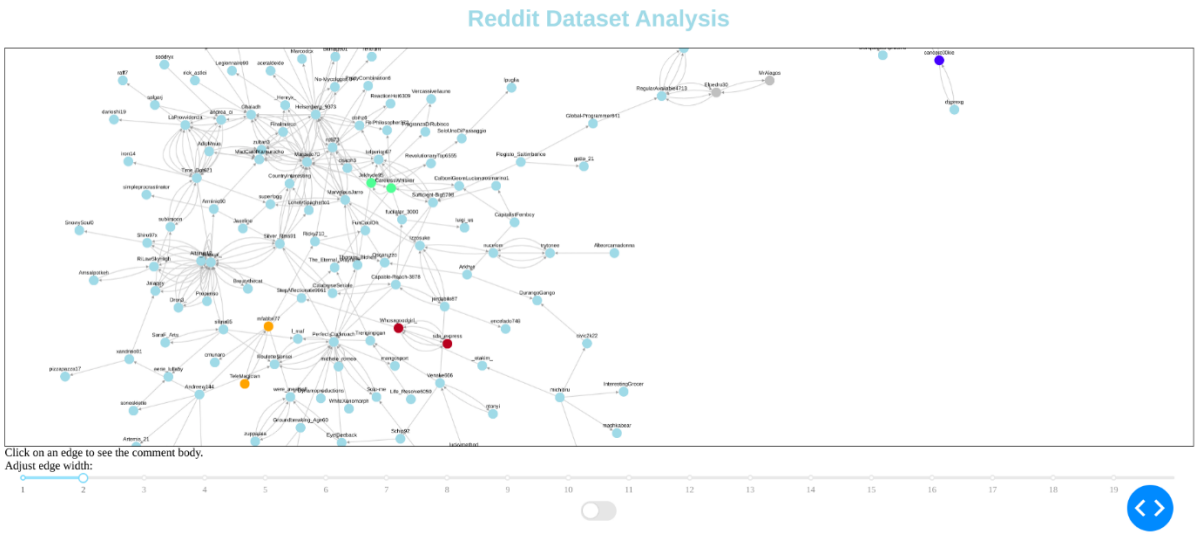
Una volta ottenuti i dati relativi ai vari dibattiti, è stata sviluppata una demo interattiva utilizzando il framework Python Dash per visualizzare in modo comprensivo il rapporto tra i nodi e la separazione dei dibattiti. L'applicazione presenta un grafo specifico relativo a tutti i commenti di un post, con funzionalità interattive che permettono agli utenti di esplorare e comprendere le relazioni tra i nodi.

3.7.1. Descrizione della Demo:

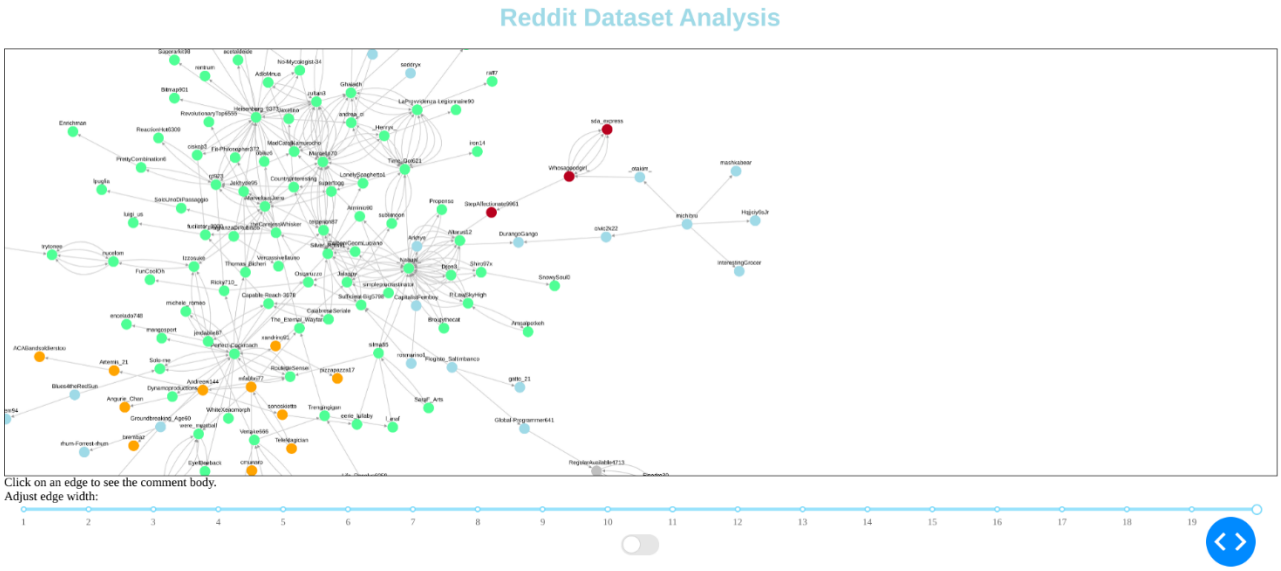
1- **Visualizzazione del Grafo**: L'applicazione mostra un grafo interattivo che rappresenta tutti i commenti di un post. Gli utenti possono interagire con il grafo cliccando sugli archi per visualizzare informazioni dettagliate, come il commento a cui è relativo, l'utente che lo ha scritto e l'utente a cui è stato indirizzato.



2- **Individuazione dei Dibattiti minimi**: Sotto la visualizzazione del grafo, è presente uno slider che consente agli utenti di regolare la profondità dei dibattiti. Gli indici vanno da 1 alla massima profondità dei dibattiti (ovvero il numero di passaggi ricorsivi che servono per individuarli), impostando il valore ad 1 è possibile evidenziare tutti quei nodi che costituiscono i dibattiti minimi, ovvero le coppie che rispettano le condizioni del caso base



3- **Visualizzazione dei Dibattiti in Formazione**: Utilizzando lo slider, gli utenti possono osservare come i dibattiti prendono forma. I nodi che soddisfano le caratteristiche del caso base vengono colorati per evidenziarli. Ogni incremento della slide mostra come il processo ricorsivo aggiunge nodi alle diverse discussioni, culminando nell'ultimo livello con tutti i dibattiti colorati in modo distinto.



4- Switch per Separare i Dibattiti: È presente uno switch che permette agli utenti di separare i dibattiti tra loro. Quando attivato, questo switch genera un grafo non connesso dove ogni componente connessa corrisponde a un dibattito distinto.

Reddit Dataset Analysis

Click on an edge to see the comment body.

Adjust edge width:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

4. Sentiment Analysis

Si è scelto di effettuare il calcolo del sentiment dei commenti che hanno generato discussione sotto ciascun post per diverse ragioni. Innanzitutto, comprendere il tono e l'emozione prevalenti nelle discussioni ha consentito di ottenere una visione più approfondita della natura delle interazioni tra gli utenti italiani di Reddit riguardo alcune tematiche.

Mentre l'analisi dei grafi ha permesso di individuare e isolare le discussioni, la sentiment analysis fornisce ulteriori informazioni qualitative riguardo al contenuto di queste discussioni: analizzare il sentiment dei commenti permette di identificare se le discussioni sono principalmente positive, negative o neutre e di comprendere meglio le reazioni emotive degli utenti rispetto a vari argomenti, siano essi di attualità o di interesse più leggero.

Un ulteriore obiettivo della sentiment analysis è stato quello di analizzare il carattere delle discussioni più importanti, ovvero quelle più discusse. È interessante vedere se i temi più dibattuti sono anche quelli che generano le emozioni più forti o se, al contrario, sono discussioni meno accese ma più partecipate. Questa distinzione aiuta a determinare come le diverse tematiche trattate nei post influenzano il tono delle conversazioni, e permette di confrontare il sentiment delle discussioni su argomenti seri, come la guerra tra Palestina e Israele o la violenza di genere, con quello su argomenti più leggeri, come il calcio o i gatti.

In questo modo, è possibile ottenere un quadro più completo delle dinamiche delle discussioni online tra gli utenti italiani di Reddit, evidenziando non solo il livello di partecipazione ma anche la qualità e il carattere delle interazioni.

L'analisi viene effettuata sul dataset generato al termine dell'analisi dei grafi, che include esclusivamente i commenti dei post selezionati inizialmente che hanno partecipato attivamente alle discussioni. Questa scelta metodologica consente di focalizzarsi sui commenti più rilevanti e di evitare rumori di fondo che potrebbero distorcere i risultati.

4.1. Metodologie e Modelli utilizzati

Il dataset utilizzato, essendo stato creato da noi, non era etichettato, il che ha reso necessario trovare un metodo per etichettarlo.

Si sono scelti due diversi approcci: uno più naif basato su una metodologia vista a lezione, l'altro che consiste nel mettere a paragone tre diversi moderni modelli di sentiment molto utilizzati.

4.1.1 Custom Lexicon Based

L'approccio iniziale scelto è stato quello del **custom lexicon based**, una metodologia discussa a lezione, che prevede l'uso di un lessico personalizzato in base al dominio.

Per questa fase, si è optato per un lessico già predisposto: l'**NRC Emotion Lexicon** che, nonostante contenga annotazioni influenti per le parole inglesi, è disponibile anche in numerose altre lingue, tra cui l'**italiano**. Pertanto, si è utilizzata la versione italiana di questo lessico.

4.1.2 Modelli preaddestrati di Sentiment Analysis a confronto

Successivamente, osservando che i risultati iniziali non erano soddisfacenti e studiando metodologie più moderne, si è deciso di utilizzare tre **modelli preaddestrati di sentiment analysis** recuperati su Hugging Face, selezionandoli tra i più utilizzati al momento.

È stato quindi **etichettato** il dataset finale sulla base del sentiment che appariva con maggiore frequenza tra i modelli selezionati, migliorando così l'accuratezza dell'analisi.

4.2. NRC Emotion Lexicon

Questo lessico delle emozioni comprende un elenco di parole e le loro associazioni con **otto emozioni** (rabbia, paura, anticipazione, fiducia, sorpresa, tristezza, gioia e disgusto) e **due sentimenti** (negativo e positivo). Le annotazioni sono state eseguite manualmente tramite Mechanical Turk di Amazon.

4.2.1. Preprocessing

Prima di passare i commenti ottenuti precedentemente al modello è necessario eseguire una fase di preprocessing, che consiste in:

1. **Tokenizzazione**: La tokenizzazione, nel contesto del trattamento del linguaggio naturale (NLP), è il processo di suddividere un testo in unità più piccole, chiamate token. Un token può essere una parola, una frase, un simbolo o qualsiasi altra unità significativa all'interno del testo.
2. **Rimozione delle stopword**: Le stopwords sono un insieme di parole utilizzate frequentemente nel linguaggio e presenti in tutti i testi, quali articoli, congiunzioni, pronomi etc... Questi non aggiungono nessuna informazione riguardo l'argomento del quale si parla in un documento, anzi, porterebbero problemi di stima vista l'alta frequenza con cui vengono utilizzati, dunque vengono rimossi dal corpus. Nel contesto specifico del nostro caso, sono state considerate le stopwords della lingua inglese poiché i documenti presenti nel dataset erano scritti in inglese.
3. **Rimozione della punteggiatura:** La rimozione della punteggiatura è spesso una pratica comune nell'elaborazione del linguaggio naturale (NLP) per diversi motivi, tra cui riduzione del rumore, uniformità del testo, dimensionalità ridotta e prevenzione dell'overfitting.

Comment_body	Comment_score	Preprocessed_comment	NRC
ci salveremmo sia dalla dipendenza	10	salveremmo dipendenza e	['negative', 'anger']
Penso quella energia si possa anch O entrambi, rompi il cazzo al vicinc	1	pensare energia potere sp	['negative', 'anger']
Non spariamo cazzate, potrei capir	-2	non sparare cazzare poter	['negative', 'anger']
Qualunque dato sulla Cina è al 99.9	-1	qualunque dato Cina 99.9	['negative', 'anger']
I padroni fanno il lavaggio del cerv	8	il padrone lavaggio cervell	['negative', 'anger']

4.3. Italian BERT Sentiment model

Questo modello esegue l'analisi del sentiment sulle frasi italiane. È stato addestrato a partire da un'istanza di bert-base-italian-cased, e messo a punto su un dataset italiano di tweets, raggiungendo su quest'ultimo l'82% di precisione.

https://huggingface.co/neuraly/bert-base-italian-cased-sentiment

4.3.1. Preprocessing

Nel caso di questo modello, il preprocessing non viene effettuato, in quanto il modello è stato fine-tunato su testi non preprocessati. Questo approccio è adottato perché BERT è particolarmente efficace nel catturare la semantica di sequenze di testo complesse.

4.4. FEEL-IT: Emotion and Sentiment Classification for the Italian Language

Utilizzo un altro classificatore di sentiment, dotato di un corpus di riferimento di post Twitter italiani annotati con quattro emozioni fondamentali: rabbia, paura, gioia, tristezza. Comprimendoli, possiamo anche fare l'analisi del sentiment.

PAPER: https://aclanthology.org/2021.wassa-1.8.pdf

https://huggingface.co/MilaNLProc/feel-it-italian-sentiment

4.4.1. Preprocessing

Il preprocessing effettuato è esattamente lo stesso del modello naive NRC Emotion Lexicon.

4.4.2. Sentiment and Emotion

Il modello FEEL-IT lavora sia sul sentiment che sulle emozioni, permettendo un'analisi più approfondita dei testi. Questo consente non solo di identificare se un commento è positivo o

negativo, ma anche di rilevare le emozioni specifiche come gioia, tristezza, rabbia e paura. Utilizzare questo modello arricchisce l'analisi del sentiment, offrendo una comprensione più dettagliata delle reazioni emotive degli utenti.

4.5. Multilingual

Si tratta di un modello bert-base-multilingue-uncased ottimizzato per l'analisi del sentiment sulle recensioni dei prodotti in sei lingue: inglese, olandese, tedesco, francese, spagnolo e italiano. Prevede il sentiment della recensione sotto forma di numero di stelle (tra 1 e 5).

https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

4.5.1. Preprocessing

Il preprocessing effettuato è esattamente lo stesso del modello naive NRC Emotion Lexicon.

4.6. Etichettatura del dataset

4.6.1. Sentiment dei commenti

Dopo aver ottenuto le previsioni da tutti e tre i modelli (BERT, FEEL_IT e MULTILINGUAL), si procede con l'etichettatura del vero sentiment del commento. Il metodo utilizzato consiste nell'utilizzare le previsioni di tutti e tre i modelli, e il sentiment prevalente tra di essi determinerà quello associato al commento. Nel caso in cui le previsioni siano divergenti (positivo, negativo e neutro), il sentiment assegnato sarà neutro.

4.6.2. Sentiment dei dibattiti

Una volta ottenuti i sentiment dei singoli commenti, è possibile calcolare il sentiment delle discussioni di cui fanno parte. Inizialmente, è stato considerato l'approccio di sommare gli score dei commenti per determinare quello prevalente all'interno del dibattito. Tuttavia, sin dalla fase di creazione del dataset, sono stati salvati anche gli score associati ai commenti, ovvero gli UpVote su Reddit, che indicano quanto un commento sia stato trovato interessante o condiviso dai lettori. Utilizzando questi score, è possibile pesare i commenti in base alla loro rilevanza, considerando anche il parere degli utenti che non hanno partecipato attivamente alla discussione ma hanno espresso un consenso con una delle parti.

È evidente come la stragrande maggioranza dei dibattiti presenti abbia un sentiment negativo. Questo risultato non sorprende, considerando che i dibattiti sono stati selezionati da contesti online, notoriamente più inclini a discussioni accese dove le persone possono esprimere liberamente le proprie opinioni. Inoltre, la tipologia di post scelti è stata deliberatamente orientata verso argomenti in cui era facile prevedere opinioni contrastanti, che naturalmente sfociano in dibattiti intensi e polarizzati.

Questo contesto contribuisce significativamente alla prevalenza di un sentiment negativo nei dibattiti analizzati. La libertà di espressione online può amplificare le tensioni e le divergenze di opinione, spesso portando a interazioni emotivamente cariche e controversie. Questa dinamica è riflessa nel sentiment dominante osservato nei dati analizzati, evidenziando come le piattaforme digitali siano spesso terreni fertile per discussioni intense e confronti vivaci tra diverse prospettive.

Questa comprensione del contesto aiuta a interpretare più accuratamente i risultati dell'analisi del sentiment, fornendo un quadro più completo delle dinamiche sociali e comunicative nelle piattaforme online.

4.6.3. Sentiment dei post

Dopo aver ottenuto gli score dai dibattiti, si calcola il sentiment dei vari post. Tuttavia, sommare semplicemente gli score dei dibattiti può risultare riduttivo, poiché non tutte le discussioni hanno la stessa importanza all'interno di un post. Oltre agli score di sentiment dei dibattiti, è pertanto necessario considerare altre metriche specifiche calcolate in precedenza, come il numero normalizzato di nodi, il numero normalizzato di archi, il grado medio dei nodi e il coefficiente di clustering medio. Queste metriche consentono di valutare l'importanza delle discussioni e, di conseguenza, di aggiustare l'impatto che il loro score ha sul post.

4.6.4. Commenti sui risultati

Nella tabella riportata di seguito sono elencati gli score dei vari dibattiti in base ai rispettivi post_id e post_sentiment:

Post_id	Post_sentiment
10v8sey	-184.24194841517496
17lese9	-83.36907841837238
17z2hci	-192.99597794495756
197vo6o	-52.085561497326196
19aeo2k	35.666666666666664
1b6cg4q	-188.61665824915823
1bulhj9	-171.99146645021645
1cwqkqe	-115.23436471362739
1d5h5h6	-112.51599326599329

Dalla tabella emerge che la maggior parte degli score assegnati ai dibattiti sono negativi, con l'unica eccezione del post 19aeo2k, relativo ai gatti. Questo risultato è in linea con le aspettative, poiché il post sui gatti è stato intenzionalmente inserito per verificare la possibilità di ottenere score positivi. I numeri forniti rappresentano un indicatore dell'intensità del sentiment associato ai post; in particolare, alcuni post presentano un sentiment più negativo rispetto ad altri.

5. Considerazioni finali e Sviluppi Futuri

Una volta completata la pipeline di lavoro e valutati tutti i dibattiti dei post assegnando loro uno score di sentiment, il progetto è concluso ma offre ancora molte opportunità di sviluppo ulteriore.

5.2. Espansione del numero di post

Per motivi di limitazioni temporali e di chiamate API, sono stati analizzati solo nove post, variati per genere ma tutti appartenenti allo stesso subreddit italiano. Questo ha permesso di ottenere una panoramica iniziale del sentiment all'interno di un contesto specifico, ma la varietà e la quantità dei dati rimane limitata.

Sarebbe interessante ampliare l'analisi del sentiment includendo un numero maggiore di post appartenenti allo stesso subreddit. Questo approccio consentirebbe di ottenere una visione più completa e dettagliata delle dinamiche di sentiment all'interno della comunità, permettendo di identificare pattern e tendenze che potrebbero non essere evidenti con un campione così ristretto.

5.3. Estrazione dei topic

Una volta ottenuto lo score di sentiment dai modelli, si potrebbe eseguire un'analisi di topic extraction sui post come sviluppo futuro. Questo permetterebbe di identificare chiaramente quali sono i temi che stimolano gli utenti a discutere attivamente e di comprendere il carattere delle loro discussioni. L'analisi di topic extraction consente di individuare automaticamente i temi principali presenti nei testi. Applicando questa tecnica ai post e ai commenti con sentiment analizzato, è possibile ottenere informazioni preziose su quali argomenti generano maggior interesse e discussione all'interno della community. Comprendere quali argomenti sono più frequentemente discussi permetterebbe di identificare le principali aree di interesse degli utenti e di osservare come i temi evolvono nel tempo, fornendo insight su trend emergenti o cambiamenti nelle preferenze degli utenti. Inoltre, la topic extraction potrebbe aiutare a collegare il sentiment dei commenti ai temi specifici, permettendo di capire se determinati argomenti generano reazioni più positive o negative. Questo tipo di analisi migliorerebbe la comprensione delle dinamiche delle discussioni e fornirebbe una base più solida per interventi mirati o ulteriori ricerche.

5.4. Sperimentazione con nuovi modelli e metriche di analisi del sentiment

Oltre ai modelli attuali come BERT, FEEL_IT e MULTILINGUAL, sarebbe utile esplorare nuovi modelli per l'analisi del sentiment come sviluppo futuro. Modelli di recente sviluppo basati su architetture transformer più avanzate potrebbero offrire miglioramenti significativi nella comprensione del contesto e delle sfumature emotive nei testi. L'esplorazione di modelli pre-addestrati su dataset specifici per determinate lingue o domini, così come la creazione di ensemble di modelli, potrebbe anche migliorare l'accuratezza delle analisi di sentiment.