UNIVERSITÀ DEGLI STUDI
DI SALERNO

NaLA

Natural Analysis of **L**anguage **A**ttitudes in BlueSky Conversations

Luigina Costante

Annalaura Miglino

Angelo Nazzaro

Dipartimento di Informatica
Laurea Magistrale in Data Science & Machina Learning
Natural Language Processing

Fisciano, June 2025

UNIVERSITÀ DEGLI STUDI
DI SALERNO

NaLA

Natural Analysis of **L**anguage **A**ttitudes in BlueSky Conversations

Luigina Costante

Annalaura Miglino

Angelo Nazzaro

Dipartimento di Informatica
Laurea Magistrale in Data Science & Machina Learning
Natural Language Processing

Fisciano, June 2025

Contents

1 Introduzione **2**

2 Stato dell'arte **3**

3 Metodologia **5**

3.1 Dataset . . . 5

3.1.1 Preprocessing . . . 5

3.1.2 Sentiment-Analysis . . . 9

3.2 Modelli "from scratch" . . . 10

3.2.1 Pre-processing . . . 10

3.2.2 RandomForest . . . 11

3.2.3 Classificatore Multi-layer Perceptron . . . 11

3.2.4 Naive Bayes . . . 12

3.2.5 RNN . . . 13

3.3 Modelli preaddestrati . . . 15

3.3.1 BERTweet . . . 15

3.3.2 RoBERTa . . . 16

4 Risultati e Discussioni **17**

4.1 Setup Sperimentale . . . 17

4.2 Sentiment Analysis . . . 18

4.2.1 Topic: Social Media . . . 18

4.2.2 Topic: Politics . . . 19

4.2.3 Topic: Gaming . . . 21

4.2.4 Topic: Relationship . . . 22

4.2.5 Topic: Religion . . . 24

4.2.6 Topic: Climate Change . . . 24

4.2.7 Topic: Immigration . . . 26

4.2.8 Topic: War & Conflicts . . . 27

4.3 Metriche . . . 29

4.4 Modelli "from scratch" . . . 30

4.4.1 RandomForest . . . 30

4.4.2 Classificatore Multi-layer Perceptron . . . 30

4.4.3 Naive Bayes 31
4.4.4 RNN 32
4.5 Modelli preaddestrati 34
4.5.1 BERTweet 34
4.5.2 RoBERTa 36
4.6 Confronto Prestazionale 37

5 Conclusioni **39**

Bibliografia 42

Questa pagina è intenzionalmente bianca.

1 Introduzione

Negli ultimi anni, la sentiment analysis su post dei social media si è affermata come uno strumento chiave per comprendere l'umore collettivo, la diffusione di opinioni e la percezione pubblica in tempo reale. Questo tipo di analisi si è rivelato particolarmente utile in una vasta gamma di contesti, tra cui il monitoraggio delle reazioni a eventi politici, campagne di comunicazione, crisi sanitarie e campagne di marketing. La disponibilità di grandi quantità di dati generati dagli utenti ha alimentato lo sviluppo di approcci sempre più sofisticati, che spaziano da metodi statistici tradizionali a modelli linguistici di tipo neurale, fino ad arrivare a modelli di linguaggio preaddestrati di grandi dimensioni.

In questo contesto, il presente lavoro si propone di esplorare e confrontare diversi approcci di sentiment analysis, caratterizzati da livelli crescenti di complessità modellistica e computazionale. L'analisi è condotta su un sottoinsieme di testi estratti da post provenienti dal social media Bluesky[1] [M, 2023], una piattaforma emergente che sta rapidamente guadagnando popolarità come alternativa decentralizzata ai social tradizionali. L'obiettivo è valutare le prestazioni dei modelli rispetto a vincoli computazionali, accuratezza e semplicità di implementazione, con particolare attenzione al trade-off tra qualità della classificazione e costo computazionale. In particolare, verranno messi a confronto:

- Metodi di *machine learning* classici, quali: Random Forest e Naive Bayes;
- Modelli di *deep learning*, quali: MLP, RNN bidirezionali, BERTweet e RoBERTa.

[1] https://bsky.app

2
Stato dell'arte

Nello stato dell'arte attuale, i lavori di sentiment analysis, come riassunto da Sudhir et al., 2021 e Kapur et al., 2022, si distinguono fondamentalmente in quattro macro-categorie di approccio:

1. **Metodi lessicali** basati su vocabolari di parole positive e negative, in cui ogni testo viene valutato come somma dei punteggi lessicali; tali tecniche sono semplici ma non tengono conto nè del contesto sintattico né della negazione strutturale.
2. **Classificatori di tipo statistico/machine learning tradizionali**, che adottano rappresentazioni quali Bag-of-Words o TF–IDF e algoritmi supervisionati come Naive Bayes, Support Vector Machine, Random Forest o Logistic Regression, offrendo un buon compromesso tra prestazioni e requisiti computazionali.
3. **Modelli di deep learning** (CNN, RNN, LSTM, GRU) che sfruttano embedding statici (Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014]) e tecniche di gestione delle sequenze per catturare relazioni a lungo raggio e ordine delle parole.
4. **Architetture transformer** (BERT, RoBERTa, BERTweet, XLNet) pre-addestrate su grandi corpora e successivamente fine-tuned su task di sentiment classification, che rappresentano attualmente lo stato dell'arte grazie alla loro capacità di modellare il contesto bidirezionale e l'attenzione su intere frasi.

Gli autori evidenziano come i modelli basati su transformer superino sistematicamente le performance dei metodi classici e delle reti ricorrenti, raggiungendo F1-score tra 0.90 e 0.95 su benchmark standardizzati, a discapito però di costi computazionali significativamente maggiori.

Di particolare rilievo, nell'ambito della ricerca sulla sentiment analysis nei social media, sono i lavori che si concentrano sulla creazione e l'analisi di dataset dedicati.

Uno dei contributi più rilevanti in questo ambito è il dataset BlueTempNet [Jeong et al., 2024], che raccoglie dinamiche temporali delle interazioni tra oltre 150.000 utenti

pubblici su Bluesky. Il dataset integra tre dimensioni principali: interazioni user-to-user (follow, block), interazioni user-to-community (join, create) e relazioni tra utenti e feed. Queste informazioni sono modellate come grafi firmati e affiliazione, e includono timestamp con precisione al millisecondo.

Gli autori evidenziano che, grazie alla natura decentralizzata della piattaforma e alla trasparenza delle API, è possibile studiare le dinamiche di rete in assenza di interferenze dovute ad algoritmi di raccomandazione controllati dalla piattaforma stessa. Inoltre, il dataset ha mostrato che eventi strutturali, come il passaggio a un modello di accesso pubblico (febbraio 2024), generano impatti misurabili nelle interazioni tra utenti, specialmente in termini di aumento di blocchi e creazione di feed. Infine, l'analisi di rete rivela che le comunità su Bluesky esibiscono proprietà di tipo small-world, con una struttura sociale che favorisce l'emergere di cluster tematici e linguistici, contribuendo a una maggiore granularità nell'analisi delle dinamiche sociali.

In linea con questo approccio, si colloca anche il lavoro di Failla et al., 2024, in cui gli autori presentano un dataset molto ampio e dettagliato, comprendente oltre 235 milioni di post da 4 milioni di utenti (circa l'81% della piattaforma) raccolti tramite l'API ufficiale di Bluesky dal 17 febbraio 2023 al 25 marzo 2024. La raccolta è stata eseguita in tre fasi: esplorazione a partire dall'account ufficiale Bluesky, raccolta del grafo seguiti/follower, e infine interazioni (like/bookmark). Quindi, è stato eseguito l'hashing degli identificativi e rimozione di dati sensibili (bio, immagini, date di registrazione), con normalizzazione dei tag linguistici secondo ISO 639-2. I post in inglese sono stati etichettati con un modello RoBERTa pre-addestrato su circa 129 milioni di tweet, mappando i sentimenti su una scala positivo/neutrale/negativo. Tuttavia, in questo lavoro non ci si concentra esclusivamente sulla sentiment analysis, piuttosto mira a fornire un dataset robusto e un'analisi sugli utenti di tale piattaforma.

3 Metodologia

3.1 Dataset

Il dataset selezionato per questa analisi è *bluesky-posts* [M, 2023], disponibile sulla piattaforma Hugging Face. Si tratta di una raccolta pubblica di 7.880.000 post estratti dal social Bluesky, una piattaforma decentralizzata di microblogging.

Il dataset è costituito da 20 file in formato *JSON Lines*, partizionati cronologicamente e ognuno contenente 393.883 post differenti. La struttura dell'archivio dati presenta sei campi fondamentali, descritti nella Tabella 3.1.

Feature	Tipo	Descrizione
`text`	Stringa	Contenuto testuale del post. Sebbene la lingua predominante sia l'inglese, il campo può contenere testi in varie lingue, emoji, hashtag e riferimenti (@).
`created_at`	Stringa ISO 8601	Timestamp di creazione in UTC.
`author`	Stringa	Handle univoco dell'utente, nel formato `nomeutente.bsky.social`.
`uri`	Stringa	Identificativo AT-URI univoco, ad esempio `at://did:plc:<ID>/app.bsky.feed.post/<ID>`.
`has_images`	Booleano	Flag che indica la presenza (`true`) o l'assenza (`false`) di immagini allegate.
`reply_to`	Stringa/Null	URI del post originale se si tratta di una risposta, `null` altrimenti.

Table 3.1: *Descrizione delle feature nel dataset* bluesky-posts

3.1.1 Preprocessing

Per ridurre i tempi di elaborazione e gestire in modo efficiente la grande mole di dati, l'analisi è stata effettuata su un sottoinsieme del dataset complessivo, costituito da cinque file `.jsonl`, ognuno contenente 393.883 post. Ogni file è stato considerato mantenendo inalterate le sei feature originarie (contenuto testuale, timestamp, autore, URI,

flag immagini e riferimento a eventuali risposte). Prima di procedere con le fasi successive, è stata eseguita una fase di preprocessing per verificare la correttezza dei dati a disposizione e per estrarre feature utili alle successive elaborazioni. La fase di preprocessing ha previsto: filtraggio linguistico, estrazione delle features e topic modeling. Ognuna di queste fasi verrà approfondita nelle sezioni successive.

Filtraggio linguistico

Dal momento che l'obiettivo principale dell'analisi è focalizzato su contenuti in lingua inglese (*"en"*), è stato applicato un filtro linguistico basato su un sistema di rilevamento automatico della lingua. In particolare, si è utilizzata una procedura che prova a identificare il codice ISO 639-1 associato al testo di ciascun post, assegnando la lingua "sconosciuta" nei casi in cui il testo fosse vuoto o generasse errori durante il riconoscimento. Successivamente, tutti i post la cui lingua rilevata non risultava inglese sono stati scartati.

Nel dettaglio, il processo di filtraggio ha portato ai seguenti risultati di numeri di post in lingua inglese per ciascuno dei cinque file analizzati: nella prima parte sono stati conservati 275.091 post, nella seconda 228.153, nella terza 240.642, nella quarta 242.366 e nella quinta 234.573. Complessivamente, il sottoinsieme filtrato comprende 1.220.825 post, un campione considerato sufficientemente rappresentativo e bilanciato per l'analisi successiva.

Estrazione delle features

Per ciascun post nei dataset di training, validation e test, sono state calcolate sette categorie di feature distinte che catturano diverse dimensioni informative.

La prima feature corrisponde alla *lunghezza del testo*, espressa come numero di token ottenuti tramite tokenizzazione linguistica. Parallelamente, sono state identificate le *parole tematiche* specifiche associate all'argomento del post mediante un dizionario esterno predefinito, registrandone sia la presenza effettiva nel testo che la frequenza assoluta. Per rilevare la polarità lessicale, sono stati implementati contatori per parole positive e negative basati sull'Opinion Lexicon di Hu et al., 2004, un repertorio lessicale standard nel dominio dell'analisi del sentiment. È stato inoltre introdotto un marcatore binario $(0/1)$ per segnalare la *presenza del termine "no"* nel testo, considerata la sua frequenza in frasi di negazione.

La componente semantica è stata catturata attraverso *embedding testuali* generati dal transformer preaddestrato `all-MiniLM-L6-v2` [Sentence-Transformers Team, 2023], che produce rappresentazioni vettoriali di 384 dimensioni. Questo modello è stato applicato con normalizzazione L2 e processing in batch da 1024 campioni.

L'output del processo è costituito da tre dataset arricchiti in formato CSV con 393 feature ciascuno, comprendenti sia indicatori sintattico-lessicali che rappresentazioni semantiche.

Topic modeling

Alla lista delle sette feature descritte in precedenza ne è stata aggiunta un'ottava, la quale rappresenta il tema (*topic*) affrontato nel testo di ciascun post. Nei paragrafi seguenti viene illustrato l'approccio adottato per individuarla.

Esperimenti preliminari e criticità riscontrate Prima di stabilire la pipeline definitiva, abbiamo condotto diverse analisi esplorative volte a verificarne l'efficacia sul nostro corpus. Nei primi test, basati su un pre-processing minimale e sulle impostazioni standard di BERTopic (UMAP e HDBSCAN "out-of-the-box"), i cluster risultavano generici e con marcata sovrapposizione lessicale. Per tali motivi si è deciso di introdurre un controllo esplicito sulla dimensione minima dei cluster e tentato un raggruppamento automatico tramite matching di parole-chiave; sebbene il numero di topic aumentasse, l'approccio lessicale si è rivelato troppo approssimativo, producendo assegnazioni semanticamente incoerenti.

Da questi riscontri emergono due esigenze fondamentali:

- integrare meccanismi di diversificazione delle parole-chiave (ad esempio MMR) per ridurre la ridondanza;
- definire soglie di similarità per le operazioni di matching, evitando decisioni basate unicamente sulla corrispondenza lessicale.

Estrazione e assegnazione dei topic La pipeline finale combina BERTopic con una fase di mappatura "zeroshot" su un insieme di 39 categorie predefinite. In primo luogo, si è costruito un vocabolario di topic generali e associato a ciascuno un set di parole-chiave rappresentative. Successivamente, ogni documento viene trasformato in embedding semantico tramite il modello `thenlper/gte-small` [Alibaba DAMO Academy, 2024], e proiettato in uno spazio a bassa dimensione con UMAP per preservare le relazioni topologiche.

Il clustering è gestito da HDBSCAN, il quale separa automaticamente i topic da eventuali outlier, mentre un modello basato su *Maximal Marginal Relevance* migliora la selezione delle parole-chiave aggiungendo varietà e rilevanza. Infine, da ogni cluster sono state estratte le parole-chiave principali e confrontate con il dizionario di topic generali, assegnando ciascun gruppo alla categoria con cui condivide il maggior numero di corrispondenze. Ogni topic riceve infine un'identità univoca, permettendo di etichettare i documenti con una label testuale e un ID numerico coerenti.

In Figura 3.1 è riportato il grafico relativo alla mappa di similarità ottenuta considerando i 39 topic definiti manualmente. La matrice presenta sia gruppi molto coesi sia relazioni inaspettate. Ad esempio, Social Media, Gaming e Journalism formano un blocco compatto ($sim \geq 0,98$), mentre Climate Change si comporta da ponte tra Urban Life e Agriculture. Legami insoliti emergono ad esempio tra Religion e Programming ($sim\ 0,94$) e tra Law & Justice e Organic Food ($sim\ 0,92$).

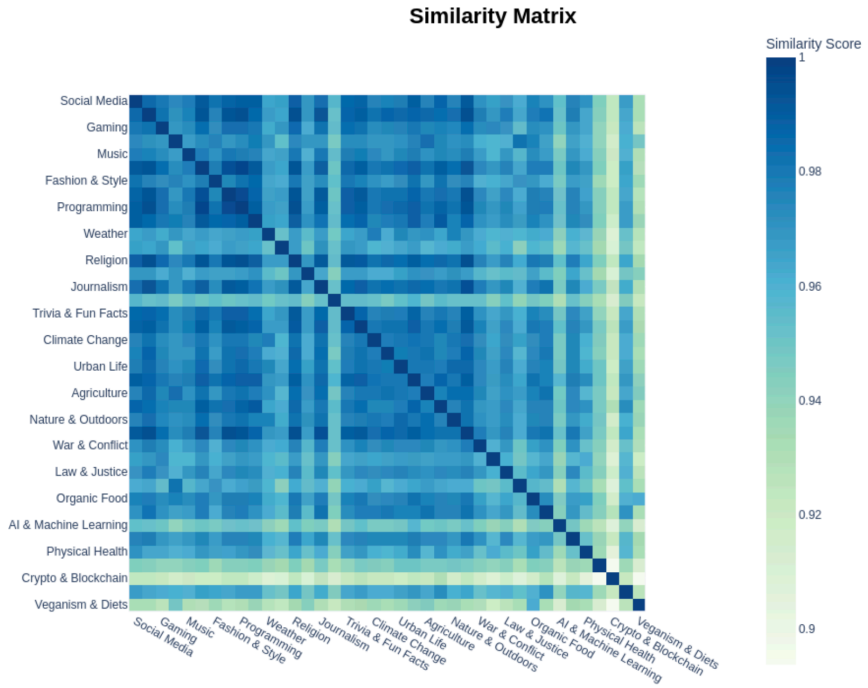

Figure 3.1: *Matrice di similarità semantica tra topic.*

Da queste osservazioni potrebbero nascere perplessità sulla capacità degli embedding di cogliere davvero le relazioni contestuali, ma il dendrogramma in Figura 3.2 ci aiuta a chiarirle. Esso mette in evidenza una struttura gerarchica caratterizzata da: un primo raggruppamento che include temi tecnologico-scientifici affiancati a questioni ambientali, un secondo associa questioni socio-culturali a tendenze di lifestyle, mentre un terzo connette nicchie specializzate con temi di attualità. A livello più fine, emergono distanze gerarchiche particolarmente ridotte tra argomenti come "Social Issues" e "Politics" (< 0.2) o "Mental Health" e "Self-improvement" (< 0.4), mentre distanze elevate (oltre 0.8) tra "War & Conflict" e "Traveling" sottolineano nette separazioni concettuali.

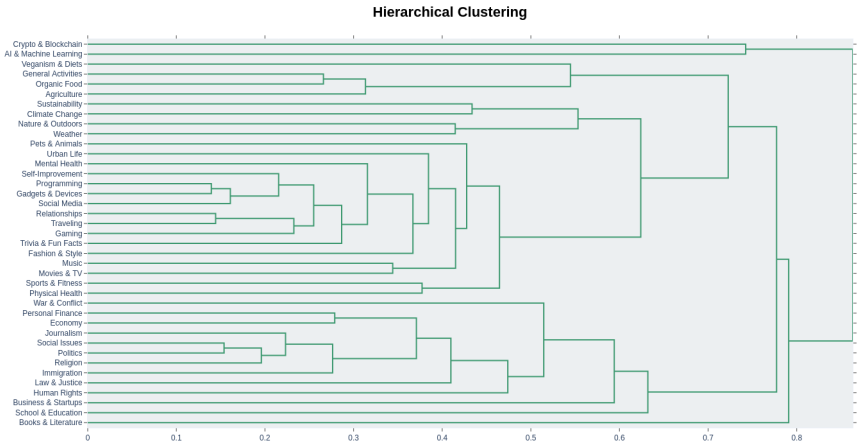

Figure 3.2: *Struttura gerarchica dei cluster.*

Analogamente, la mappa bidimensionale in Figura 3.3 restituisce una disposizione spaziale dei 39 topic che ne evidenzia la coesione per quadranti: in basso a sinistra trovano posto i temi tecnico-scientifici, in alto a sinistra quelli legati a sostenibilità e ambiente, in alto a destra la cultura digitale, e in basso a destra le questioni sociali. Le distanze minime si riscontrano tra argomenti affini come ”Gaming” e ”Fashion & Style”; quelle massime separano, per esempio, ”Social Media” da ”Veganism & Diets” o ”Programming” da ”Law & Justice”.

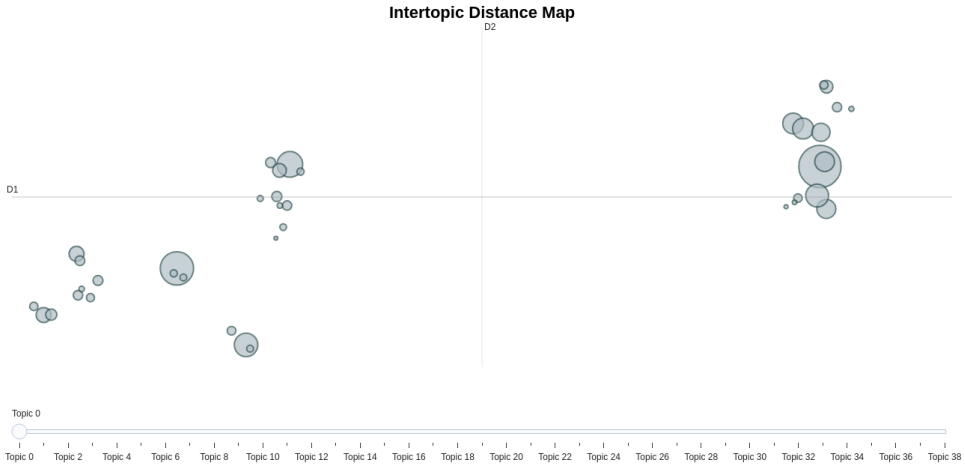

Figure 3.3: *Proiezione bidimensionale delle distanze tra i topic.*

3.1.2 Sentiment-Analysis

Al fine di rendere il dataset applicabile a un task di sentiment analysis, i testi sono stati etichettati utilizzando un modello pre-addestrato basato su BERT [Devlin et al., 2019; TabularisAI, 2024] che ha classificato i testi in cinque classi: *Positive*, *Negative*, *Very Positive*, *Very Negative* e *Neutral*. Nel grafico 3.4 è riportata la distribuzione iniziale delle classi di sentimento, come risultante dall'etichettatura automatica. È possibile osservare un marcato sbilanciamento, con una predominanza delle classi *Neutral* e *Very Positive*, a discapito delle altri classi.

Per mitigare tale squilibrio e rendere la distribuzione delle classi più bilanciata, è stata effettuata una fusione semantica: le classi *Positive* e *Very Positive* sono state unite, raggiungendo complessivamente il 38.5% dei campioni; analogamente, *Negative* e *Very Negative* sono state accorpate, totalizzando il 25%. La nuova distribuzione, a seguito di questa aggregazione, è mostrata in Figura 3.5.

Limiti computazionali A causa di risorse computazionali limitate, è stato considerato il 50% delle istanze totali, corrispondente a circa 630.000 istanze selezionate tramite campionamento stratificato. Successivamente, tali dati sono stati suddivisi nelle seguenti partizioni: 70% per il training set, 15% per il validation set e 15% per il test set.

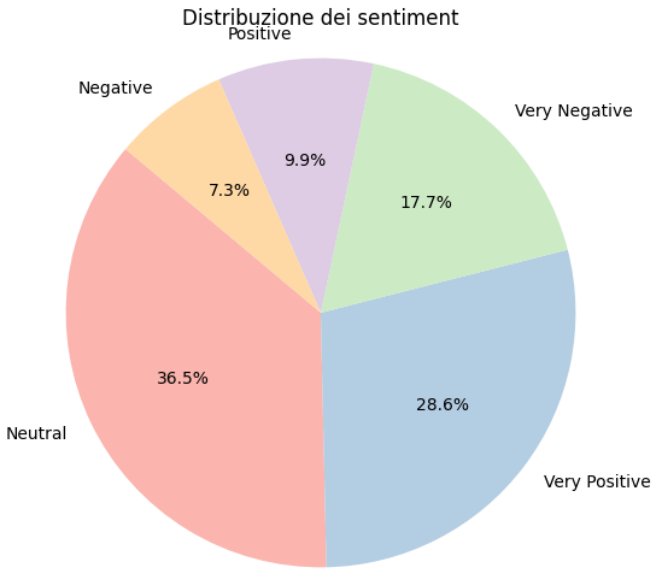

Figure 3.4: *Distribuzione iniziale delle classi di sentimento.*

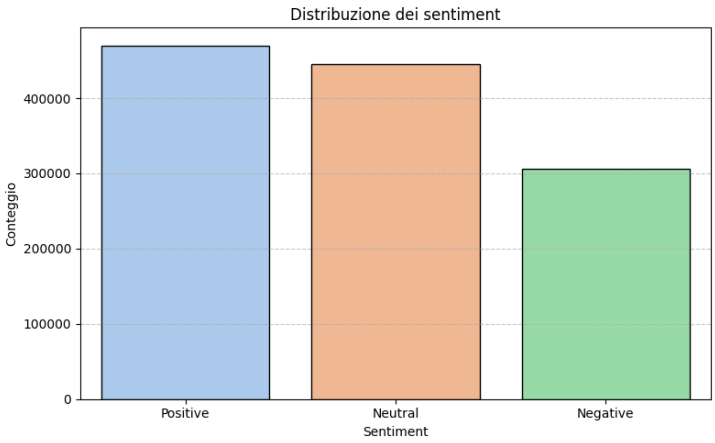

Figure 3.5: *Distribuzione delle classi di sentimento dopo l'unione.*

3.2 Modelli "from scratch"

3.2.1 Pre-processing

Prima di effettuare l'addestramento, tutti i modelli considerati sono stati sottoposti alla medesima fase di pre-processing, descritta di seguito:

1. **rimozione dei valori nulli**: sono state eliminate le righe contenenti valori nulli, poiché sono risultate in misura minima rispetto al totale delle istanze;
2. **tokenizzazione**;
3. **rimozioni delle *stop words***;
4. **lemmatizzazione del testo tramite WordNet** [**Miller, 1995**; **Fellbaum, 1998**];
5. **gestione delle negazioni**: le parole che seguono una negazione (es. *don't*, *never*, *didn't*) vengono prefissate con il token `NOT_` fino al successivo segno di punteggiatura.

Ad esempio, la frase *"I didn't like it, the movie I mean"* viene trasformata in *"I didn't* `NOT_like` `NOT_it`*, the movie I mean"*. Questo dovrebbe consentire ai modelli di distinguere tra il significato affermativo e quello negato di una parola, permettendo, ad esempio, a termini come `NOT_bored` di essere interpretati in senso positivo. Tale strategia migliora la sensibilità del modello verso le inversioni di polarità semantica, che sono frequenti nei testi soggettivi.

Si precisa che non sono state effettuate operazioni relative al casing delle parole in quanto quest'ultimo potrebbe rivelarsi fondamentale ai fini della corretta classificazione del sentimento, in quanto questo è tipicamente utilizzato dagli utenti per esprimere emozioni forti.

Il risultato di tutto il processo descritto è, quindi, una lista di token normalizzati. Per quanto riguarda le feature numeriche (come lunghezza del testo, conteggio di parole topic e sentiment ed embeddings), per garantire la coerenza delle scale e velocizzare la convergenza del modello, sono state normalizzate con lo StandardScaler per normalizzare i valori a media zero e deviazione standard unitaria. Le feature categoriche (come il topic) vengono codificate tramite one-hot-encoding per trasformarle in variabili binarie. Inoltre, per ogni modello, sono stati applicate delle fasi *ad-hoc* dove necessario che verranno descritte nelle sezioni succesive.

3.2.2 RandomForest

Random Forest è un algoritmo di machine learning supervisionato, utilizzato sia per problemi di classificazione che di regressione. È uno dei metodi più popolari e potenti, grazie alla sua accuratezza, robustezza e capacità di evitare l'overfitting. Alla base di Random Forest c'è un insieme ("foresta") di alberi decisionali ("decision trees") in cui ciascun albero contribuisce con un voto o una stima per determinare la previsione finale. In pratica, viene generato un insieme di alberi decisionali, ciascuno addestrato su un sottoinsieme casuale del dataset. Ogni albero, quindi, durante la costruzione, considera solo un sottoinsieme casuale di feature a ogni nodo, aumentando la diversità tra gli alberi. Per la classificazione ogni albero vota e si prende la classe più votata (majority vote).

Il Random Forest non accetta feature letterali, per cui, in questo contesto, sono state eliminate tutte le colonne contenenti valori testuali e liste, sostituendole con altre feature relative alla lunghezza di tali liste oppure a quella dei dati testuali. L'addestramento è stato eseguito con la K-Fold Cross Validation con $k = 10$.

3.2.3 Classificatore Multi-layer Perceptron

In questa sezione è descritto il classificatore basato su una rete neurale di tipo *multi-layer perceptron* (MLP) progettato per assegnare a ciascun testo una delle tre possibili categorie di sentimento: positivo, negativo o neutro. Il modello integra informazioni

testuali con feature numeriche e embedding semantici, con l'obiettivo di superare i limiti delle tecniche tradizionali basate esclusivamente su bag-of-words.

L'architettura MLP è composta da uno strato di input, un unico hidden layer e uno strato di output. In dettaglio, il vettore di input ha dimensione $D = N_{\text{num}} + N_{\text{one\_hot}} + 384$, dove N_{num} rappresenta il numero di feature numeriche di base precedentemente elencate, $N_{\text{one\_hot}}$ è il numero di categorie di topic e 384 è la dimensione dell'embedding semantico. L'hidden layer è costituito da 50 neuroni, ciascuno dotato di funzione di attivazione ReLU, definita come $\text{ReLU}(z) = \max(0,\, z)$. Sul layer di output, che conta un neurone per ogni classe di sentiment, è stata applicata la funzione *softmax* per ottenere una distribuzione di probabilità fra le tre classi:

$$\sigma(\mathbf{z})_i \;=\; \frac{e^{z_i}}{\sum_{j=1}^{3} e^{z_j}}, \quad i = 1, 2, 3.$$

Il modello è stato addestrato minimizzando la *cross-entropy loss* per classificazione multiclasse, integrata da un termine di regolarizzazione ℓ_2 con peso $\alpha = 10^{-4}$.

L'ottimizzazione dei pesi è effettuata mediante l'algoritmo *Adam* mentre il tasso iniziale η_0 è fissato a 10^{-3} nella modalità *adaptive*, in modo che, se la perdita di validazione smette di diminuire per dieci epoche consecutive, il learning rate venga dimezzato.

3.2.4 Naive Bayes

Il classificatore Naive Bayes è un modello probabilistico *semplice ma efficace* per la classificazione di testi. Esso rappresenta i documenti attraverso il paradigma della bag-of-words, ovvero come collezioni *non ordinate* di parole, ciascuna associata alla propria frequenza nel documento. Questa rappresentazione ignora la struttura sintattica e la posizione delle parole, concentrandosi esclusivamente sulla loro presenza e frequenza.

Naive Bayes predice la classe più probabile per un documento d, dato un insieme di classi C, calcolando la probabilità a posteriori tramite la seguente formulazione:

$$\hat{c} := \arg\max_{c \in C}; P(c|d) \tag{3.1}$$

Applicando il teorema di Bayes, la probabilità a posteriori può essere riscritta come:

$$\hat{c} := \arg\max_{c \in C}; \frac{P(d|c)P(c)}{P(d)} = \arg\max_{c \in C}; P(d|c)P(c) \tag{3.2}$$

Poiché $P(d)$ è costante rispetto alla classe, può essere ignorata durante la fase di classificazione.

Addestramento e Pre-processing

Per l'addestramento del classificatore, è stata utilizzata la Laplace smoothing (o *add-one smoothing*) al fine di evitare probabilità nulle per parole non osservate in combinazione

con una certa classe durante la fase di training. Inoltre, in linea con la letteratura, le parole non presenti nel test set sono state rimosse dal training set, così da ridurre l'impatto di termini irrilevanti e garantire una valutazione più realistica del modello.

Sulla stessa linea di principio, in aggiunta a quanto descritto precedenetemente il preprocessing, è stato applicato il conteggio binario: anziché conteggiare ogni occorrenza di una parola, è stato adottato un conteggio binario. In questo modo, una parola contribuisce una sola volta per documento, indipendentemente da quante volte vi compare. Questa scelta consente di enfatizzare i termini che risultano realmente caratteristici a livello documentale, riducendo l'influenza di ripetizioni meccaniche e mettendo in risalto parole che offrono indizi più significativi sul sentimento realmente espresso nel testo.

3.2.5 RNN

Le Recurrent Neural Networks (RNN) [Elman, 1990] costituiscono una particolare classe di reti neurali caratterizzate dalla presenza di connessioni ricorrenti, in cui l'output di alcune unità viene reintrodotto come input in istanti temporali successivi. Questa struttura permette alle RNN di mantenere una *forma di memoria*, risultando particolarmente adatte alla modellazione di sequenze temporali, come i testi —motivo per cui trovano ampio impiego nel campo del NLP.

Le RNNs sono dette *ricorrenti* perché, data una sequenza $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, applicano la stessa operazione ad ogni elemento $\mathbf{x}_t$, con $t = 1 \ldots, n$. I componenti principali di una RNN sono:

- **Tre matrici di peso condivise**: le operazioni delle RNNs sono definite da tre matrici condivise
  - **input-to-hidden** ($\mathbf{U}$): connette l'input all'hidden state;
  - **hidden-to-hidden** ($\mathbf{V}$): consente alla rete di manetenere una forma di memoria tra i vari istanti temporali;
  - **hidden-to-output** ($\mathbf{W}$): connette l'hidden state all'output.
- **hidden state**: l'**hidden state** $\mathbf{h}_t$ agisce come *la memoria della rete* all'istante t. Dipende dall'input corrente $\mathbf{x}_t$ e l'hidden state precedente $\mathbf{h}_{t-1}$:

$$\mathbf{h}_t := g(\mathbf{U} \cdot \mathbf{x}_t + \mathbf{V} \cdot \mathbf{h}_{t-1} + \mathbf{b}) \tag{3.3}$$

  dove g è una funzione di attivazione non lineare.
- **Output**: ad ogni istante temporale, il vettore di output $\mathbf{o}_t$ è calcolato sulla base dell'hidden state corrente:

$$\mathbf{o}_t := f(\mathbf{W} \cdot \mathbf{h}_t + \mathbf{b}) \tag{3.4}$$

  dove f è una funzione di attivazione adatta al task (ad esempio, la softmax per la classificazione multi-classe).

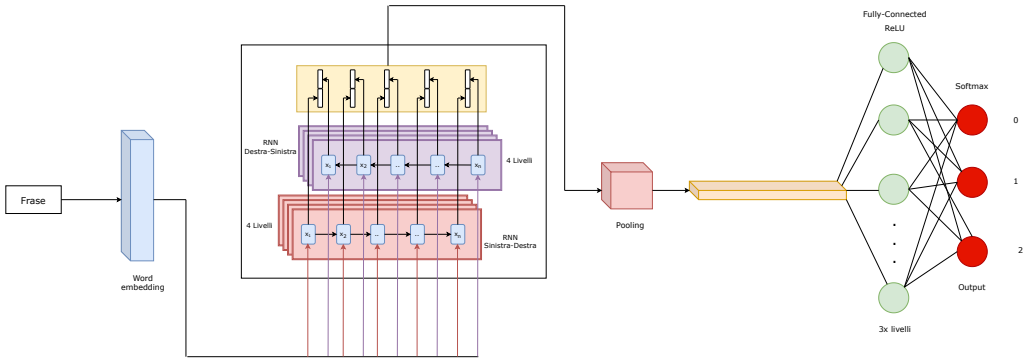

Figure 3.6: *Architettura del modello di sentiment-analysis basato su RNN.*

Architettura

Per affrontare il compito di sentiment analysis, abbiamo adottato una RNN stacked bidirezionale. Questa scelta è stata guidata dalla necessità di catturare informazioni contestuali provenienti sia dal passato (contesto sinistro) che dal futuro (contesto destro) della sequenza testuale. Le RNN standard, infatti, trattano le sequenze in una sola direzione, limitando la loro capacità di cogliere relazioni dipendenti da parole future rispetto a una data posizione nella frase. Le RNN bidirezionali [Schuster et al., 1997] superano tale limitazione elaborando la sequenza in entrambe le direzioni.

Nel dettaglio, abbiamo addestrato due reti RNN distinte, ciascuna costituita da 4 livelli: una che elabora la sequenza in ordine cronologico, denotata come $\text{RNN}_{\text{forward}}$, e l'altra in ordine inverso, denotata come $\text{RNN}_{\text{backward}}$. Per ogni posizione temporale t nella sequenza, i rispettivi hidden states $\mathbf{h}_t^f$ (forward) e $\mathbf{h}_t^b$ (backward) vengono concatenati per ottenere una rappresentazione completa del contesto bidirezionale:

$$\mathbf{h}_t := \mathbf{h}_t^f \oplus \mathbf{h}_t^b \tag{3.5}$$

Vengono considerati esclusivamente i vettori di stato nascosto $\mathbf{h}t^f$ e $\mathbf{h}t^b$ provenienti dall'ultimo strato di entrambe le reti. La loro concatenazione fornisce la rappresentazione finale $\mathbf{h}_t$, che viene quindi utilizzata come input per un classificatore fully-connected. Quest'ultimo impiega una funzione di attivazione softmax per produrre una distribuzione di probabilità sulle classi possibili. La classe predetta viene selezionata applicando l'operazione di `argmax` sulla distribuzione risultante:

$$c := \underset{i \in \{0,1,2\}}{\arg\max} \left(\text{Softmax}(\mathbf{W} \cdot \mathbf{h}_t + \mathbf{b})\right) \tag{3.6}$$

Dove $\mathbf{W}$ e $\mathbf{b}$ rappresentano rispettivamente i pesi e il bias del classificatore fully-connected. In Figura 3.6 è mostrata l'architettura della rete.

Addestramento

Per contenere la complessità del modello e garantire una rappresentazione efficace, la rete è stata addestrata utilizzando un vocabolario limitato alle $20,000$ parole più frequenti nel set di addestramento. Le parole non comprese in questo dizionario vengono mappate su un token speciale `<UNK>` per rappresentare termini fuori vocabolario.

Come funzione di loss, è stata adottata la cross-entropy, comunemente utilizzata nei problemi di classificazione multiclasse. Essa misura la discrepanza tra la distribuzione predetta $\hat{y}$ e quella reale y:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{3.7}$$

dove C è il numero di classi, y_i è la variabile indicatrice (1 se l'osservazione appartiene alla classe i, 0 altrimenti), e $\hat{y}_i$ è la probabilità predetta per la classe i.

3.3 Modelli preaddestrati

3.3.1 BERTweet

Per l'analisi del sentiment, si è scelto di adottare *BERTweet*, un modello linguistico su larga scala preaddestrato specificamente per il testo proveniente da Twitter in lingua inglese, come presentato da Nguyen et al., 2020. A differenza dei modelli generici, questo sistema è ottimizzato per le peculiarità del linguaggio utilizzato nei social media, caratterizzato da un linguaggio informale ricco di abbreviazioni, errori ortografici e slang, nonché da elementi distintivi come hashtag, menzioni di utenti e emoticon. Inoltre, BERTweet è progettato per gestire strutture testuali spesso frammentarie, tipiche dei messaggi brevi e dalla sintassi non sempre regolare.

Il modello si basa sull'architettura di BERT-base (Devlin et al., 2019), comprendente dodici livelli di Transformer, con hidden states di dimensione pari a 768 e dodici attention-head, gestendo sequenze di lunghezza massima pari a 128 token. Il preaddestramento è stato eseguito su un corpus di circa 850 milioni di tweet in inglese, applicando la metodologia di RoBERTa (Liu et al., 2019).

Per l'applicazione specifica dell'analisi del sentiment, si è fatto ricorso a una versione fine-tuned del modello, resa disponibile su *Hugging Face*. Questo adattamento prevede una tokenizzazione specializzata che consente di gestire elementi caratteristici dei social media, quali emoticon, menzioni e URL, trasformandoli in rappresentazioni testuali standardizzate. Il modello è configurato per effettuare una classificazione su tre classi corrispondenti alle etichette Negative, Neutral e Positive.

L'adozione di BERTweet è stata ritenuta adatta per l'analisi del sentiment su Bluesky Social, poiché i contenuti pubblicati su questa piattaforma presentano caratteristiche linguistiche simili a quelle dei tweet, in termini di informalità e presenza di emoji. Inoltre, essendo stato addestrato su sequenze relativamente brevi e su testi affetti da ru-

more linguistico, dimostra robustezza anche nell'interpretare messaggi con grammatica irregolare o ortografia non convenzionale, condizioni frequentemente riscontrate nelle comunicazioni su Bluesky.

3.3.2 RoBERTa

In aggiunta a BERTweet, è stato considerato un ulteriore modello basato su RoBERTa per la fase di valutazione, diverso da quello utilizzato in fase di etichettature del dataset per evitare possibili bias e fornire una valutazione quanto più imparziale possibile. In particolare, è stato utilizzato il modello `twitter-roberta-base-sentiment-latest` [Barbieri et al., 2023], disponibile sulla piattaforma *Hugging Face*. Questo modello è stato pre-addestrato su circa 124 milioni di tweet pubblicati tra gennaio 2018 e dicembre 2021, e successivamente fine-tunato per il compito di sentiment analysis utilizzando il benchmark TweetEval [Barbieri et al., 2020]. Le motivazioni alla base dell'adozione di questo modello coincidono con quelle già illustrate nella sezione precedente.

4
Risultati e Discussioni

4.1 Setup Sperimentale

Iper-parametri A causa delle medesime limitazioni computazionali descritte precedentemente (Sezione 3.1.2), non è stato possibile effettuare il tuning degli iperparametri per i modelli più complessi (MLP e RNN) mediante tecniche come la grid search. Di conseguenza, per questi ultimi sono stati adottati valori di iperparametri già noti e consolidati nella letteratura. Gli iperparametri utilizzati per l'MLP e l'RNN sono riportati nella Tabella 4.1. L'unico modello su cui è stato possibile applicare la grid search è il Random Forest, essendo computazionalmente meno oneroso. La grid search è stata effettuata sui seguenti iper-parameteri:

- *max_depth*: None, 5, 20, 30;
- *max_features*: 'sqrt', 'log2';
- *min_samples_leaf*: 1, 2, 4;
- *min_samples_split*: 2, 5, 10;
- *n_estimators*: 100, 200, 500.

Tuttavia, è emerso che le combinazioni migliori coincidono con i parametri di default forniti da sickit-learn, ovvero:

- *criterion*: 'gini';
- *max_depth*: None;
- *max_features*: 'sqrt';
- *min_samples_leaf*: 1;
- *min_samples_split*: 2;
- *n_estimators*: 100.

Altre considerazioni È stato applicato l'early stopping sia durante l'addestramento dell'MLP che della RNN.

Modello	Learning Rate	N. Livelli	Embedding Size	Alpha
RNN	$1 \cdot 10^{-5}$	4	768	–
MLP	$1 \cdot 10^{-3}$	2	–	$1 \cdot 10^{-4}$

Table 4.1: *Iper-parametri utilizzati per ciascun modello. Per il RandomForest, sono stati usati i valori di default forniti dall'implementazione di sickit-learn.*

4.2 Sentiment Analysis

Diventa ora interessante esplorare in che modo determinati topic vengano percepiti su Bluesky, attraverso l'analisi del sentiment ad essi associato. I topic considerati sono 39, e ciascuno di essi ha una propria frequenza all'interno del dataset; quest'ultima viene riassunta per tutti all'interno della Figura 4.1. In questo caso, si osserva che il

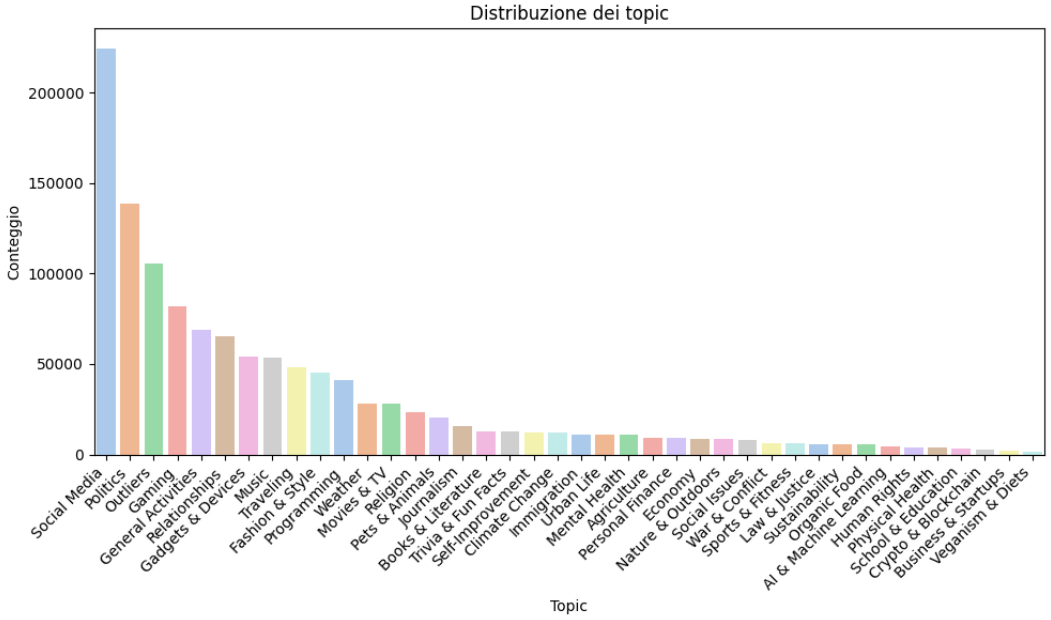

Figure 4.1: *Istogramma dei topic.*

topic *Social Media* è quello con il maggior numero di post associati; tuttavia, è importante sottolineare che, sebbene i topic posizionati più a destra sull'asse X presentino frequenze inferiori, il loro volume resta comunque significativo, nell'ordine delle migliaia, considerando che il primo intervallo sull'asse Y raggiunge fino a 50.000 istanze. A questo proposito, si approfondiscono i topic che superano tale soglia, per capire comé il sentimento generale dei post su BlueSky, e anche topic di interesse generale, *Religion*, *Climate Change*, *Immigration* e *War & Conflicts*.

4.2.1 Topic: Social Media

Il numero totale di post che presentano questo topic sono 224.046, e di questi:

- 87.053 (38.8%) hanno sentiment *Neutral*;
- 82.202 (36.6%) hanno sentiment *Positive*;

- 54.791 (24.6%) hanno sentiment *Negative*.

Questo viene visualizzato nell'istogramma in Figura 4.2.

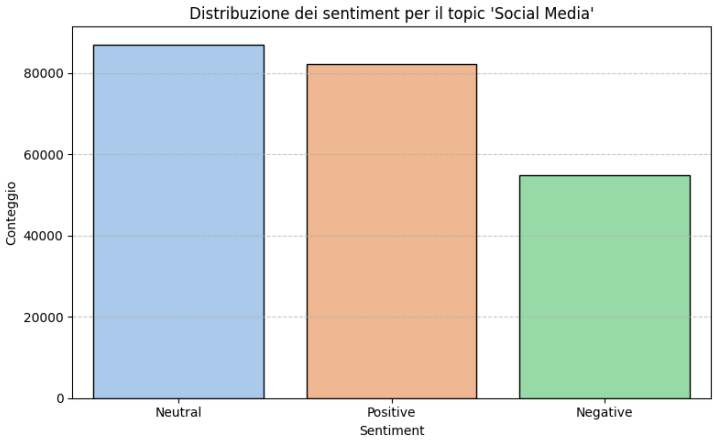

Figure 4.2: *Sentiment del topic Social Media.*

Di conseguenza, si può concludere che la maggior parte dei post inerenti a tale topic mostra neutralità o positività. Per quanto riguarda i post neutrali, sono rilevanti a constatazioni che vengono fatte sui social network, come ad esempio "*Still posting on X?*"; invece, i post negativi sono, per la maggior parte, relativi a lamentele su BlueSky oppure utenti che non hanno comportamenti corretti. I post positivi, invece, sono per la maggior parte ringraziamenti o auguri, o apprezzamenti riguardo particolari tematiche espresse da altri utenti. Volendo riportare un esempio di post positivo:

> I hit 50 followers. Thanks gang.

Si può, inoltre, valutare anche la lunghezza dei post relativamente al topic *Social Media*, che si può osservare in Figura 4.3. Da tale grafico, si nota subito una distribuzione asimmetrica, essendo che la maggior parte dei post è concentrata tra 0 e 40 parole (con picco intorno alle 20 parole). Inoltre, è presente una coda lunga verso destra, poichè pochi post superano le 60-80 parole, infatti il 75% dei post contiene un numero minore a 30 parole. Quasi nessun post raggiunge le 100+ parole, suggerendo che i contenuti su questo topic sono tipicamente concisi.

4.2.2 Topic: Politics

Il numero totale di post che presentano questo topic sono 138.541, e di questi:

- 56.126 (40.5%) hanno sentiment *Neutral*;
- 47.077 (33.9%) hanno sentiment *Negative*;
- 35.338 (25.6%) hanno sentiment *Positive*.

Questo viene visualizzato nell'istogramma in Figura 4.4.

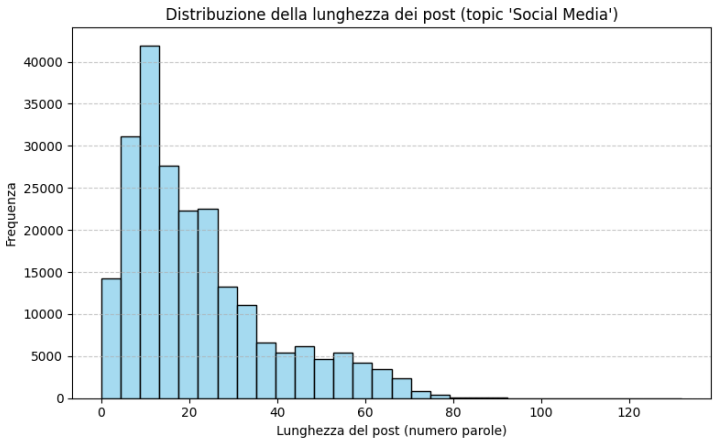

Figure 4.3: *Lunghezza dei post del topic Social Media.*

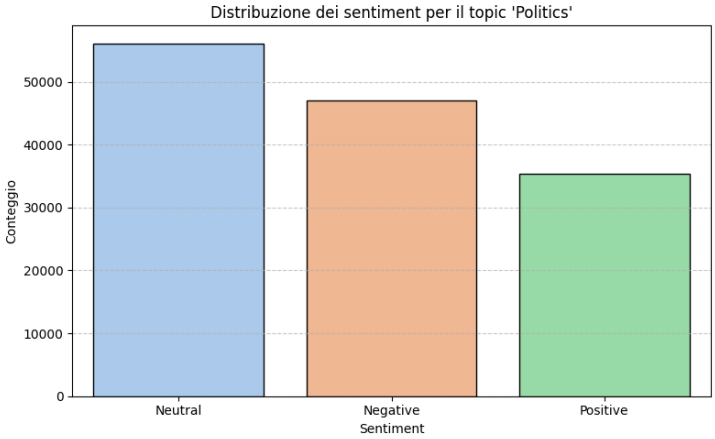

Figure 4.4: *Sentiment del topic Politics.*

Di conseguenza, si può concludere che la maggior parte dei post inerenti a tale topic mostra neutralità o negatività, al contrario del topic precedente in cui i post con sentiment *Negative* costituivano solo il 24.6%. I post neutrali sono relativi, per lo più, a petizioni, breaking news o constatazioni generali, in cui si fa evidenza di un particolare evento. I post positivi sono relativi a ringraziamenti per volontariato, oppure a gioia relativamente a cambiamenti politici, come l'elezione di alcune cariche dello Stato. Per quanto riguarda i post negativi, invece, si parla principalmente di lamentele riguardo alcuni principi Costituzionali non rispettati, e scontento riguardo alcuni diritti che vengono violati, sia da civili che da cariche dello Stato. Si riporta un esempio di un post con sentiment *Negative*:

> It is not the job of marginalized groups to explain to people why we should not be belittled, oppressed, and killed. Imagine asking a black person to explain to a white person why we aren't subhuman.

Si può, inoltre, valutare anche la lunghezza dei post relativamente al topic *Politics*,

che si può osservare in Figura 4.5. Da tale grafico, si nota che la maggior parte dei post hanno una lunghezza intorno alle 25 parole. Sono presenti, però, anche degli aumenti intorno a 50 parole e tra 50 e 75, indicando post più lunghi. Addirittura, alcuni post raggiungono le 150 parole, indicando dibattiti più lunghi (rari ma presenti). Il 50% di tali post contiene 21 parole o meno, mentre il 75% ne contiene 37. I post del topic *Politics* sono mediamente più lunghi di quelli di *Social* Media (dove il picco era a 20 parole). Inoltre, la coda è più estesa, essendo che comunque in politica si scrivono più contenuti complessi.

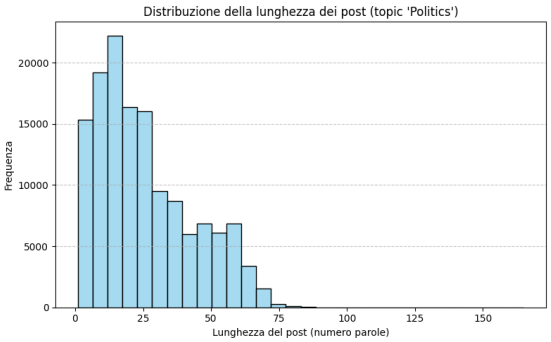

Figure 4.5: *Lunghezza dei post del topic Politics.*

4.2.3 Topic: Gaming

Il numero totale di post di tale topic sono 82.030, di cui:

- 34.045 (41.5%) hanno sentiment *Positive*;
- 26.318 (32%) hanno sentiment *Neutral*;
- 21.667 (26.5%) hanno sentiment *Negative*.

La Figura 4.6 mostra tale situazione. Come si può notare, i sentiment positivi e neu-

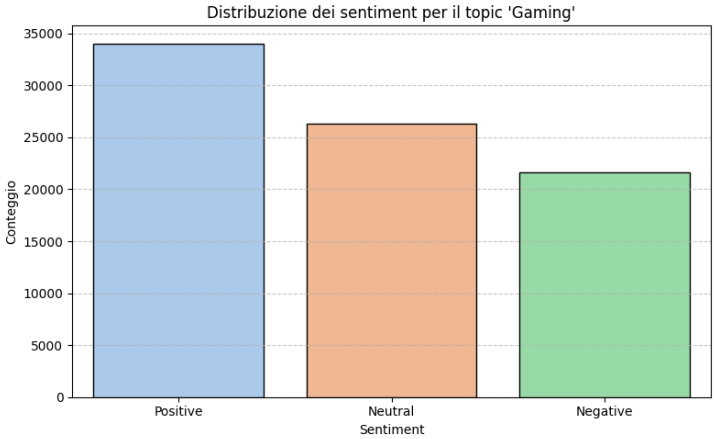

Figure 4.6: *Sentiment del topic Gaming.*

trali sono quelli che prevalgono, ricoprendo il 73.5% delle istanze totali del topic considerato. Per quanto riguarda i post neutrali, sono relativi principalmente a prestazioni

su GPU, occupazione della RAM di alcuni giochi e tutorial YouTube, oltre alla narrazione di alcune situazioni che vengono a crearsi nell'ambiente di gioco. I post negativi, invece, sono relativi a lamentele su alcuni giochi, e narrazioni di situazioni in cui il protagonista (autore del post) perde contro l'antagonista. I post positivi, invece, sono relativi ad entusiasmo verso alcuni giochi, oltre che loro performance su vari dispositivi. Si riporta un esempio di un post con sentiment *Positive*:

> Guys, I'm very late to the party, but Zelda: Breath Of The Wild is an absolutely phenomenal game. I'm only half way through it (hard to tell with Zelda games) but I'm totally blown away by it in so many ways.

Volendo valutare anche la lunghezza dei post relativamente al topic *Gaming*, si può visualizzare in Figura 4.7. Si nota che la maggior parte delle istanze è concentrata prima di 40, con picchi intorno a 20. Quasi nessun post supera le 75-100 parole (a differenza del topic *Politics* che arrivava a 150). L'istogramma mostra, inoltre, un picco molto stretto e alto, tipico di messaggi ultra-brevi e standardizzati. Il 50% dei post, infatti, ha meno di 19 parole, mentre il 75% un numero minore o uguale di 31 parole.

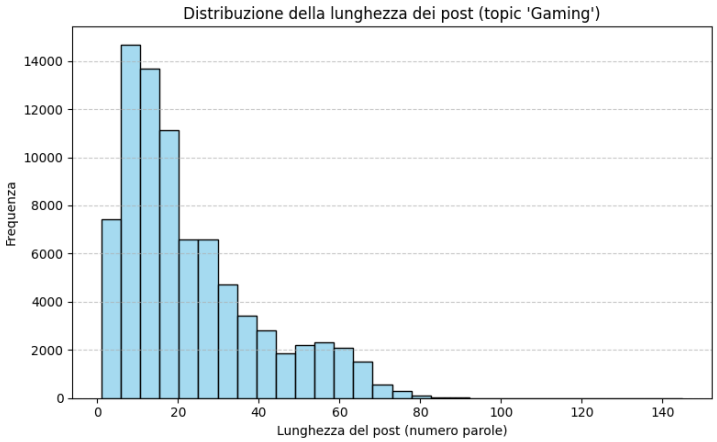

Figure 4.7: *Lunghezza dei post del topic Gaming.*

4.2.4 Topic: Relationship

I post in totale relativi a questo topic sono 65.085; di questi:

- 27.753 (41.6%) hanno sentiment *Positive*;
- 20.366 (31.2%) hanno sentiment *Neutral*;
- 16.966 (27.2%) hanno sentiment *Negative*.

Ciò viene mostrato in Figura 4.8, in cui risulta evidente che le istanze negative costituiscono solo una piccola parte del totale. Analizzando anche i post negativi, risultano relativi per lo più a situazioni di rottura dei rapporti, sia d'amore che di amicizia. In-

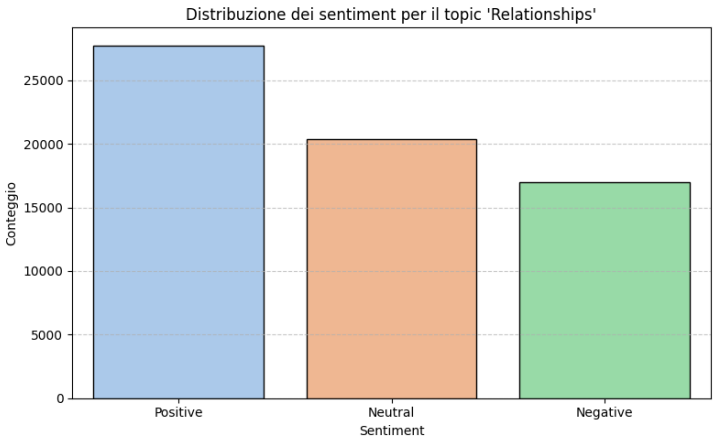

Figure 4.8: *Sentiment del topic Relationships.*

vece, i post positivi sono relativi ad auguri in occasioni festive, e alla gratitudine di avere alcune persone accanto. Ad esempio, un post *Positive* è il seguente:

> Today, I am truly thankful for a supportive family and for great friends. I am thankful to be working with great teams full of amazing people. And thankful for all of you new friends on Bluesky. Thank you!

Volendo valutare anche la lunghezza dei post, si rimanda alla Figura 4.9. Come si può notare, la maggior parte dei post contiene 20 parole o poco più. Questo può essere attribuito al fatto che i messaggi di auguri o comunque di gratitudine sono brevi e concisi, e raramente si dilungano oltre 60 parole. Infatti, il 75% delle istanze è concentrato prima delle 32 parole.

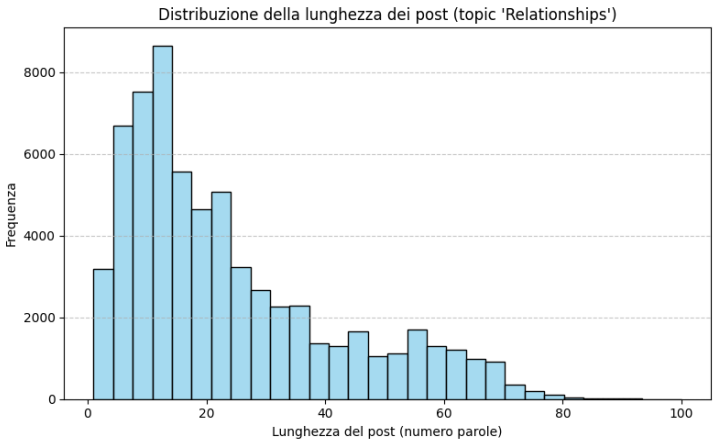

Figure 4.9: *Lunghezza dei post del topic Relationships.*

4.2.5 Topic: Religion

Per quanto riguarda il topic *Religion*, il sentiment non si sbilancia molto verso uno piuttosto che un altro, come si nota anche dalla Figura 4.10. Infatti, si ha che, su 23.633 istanze:

- 8.325 (35.2%) hanno sentiment *Negative*;
- 8.181 (34.6%) hanno sentiment *Neutral*;
- 7.127 (30.2%) hanno sentiment *Positive*.

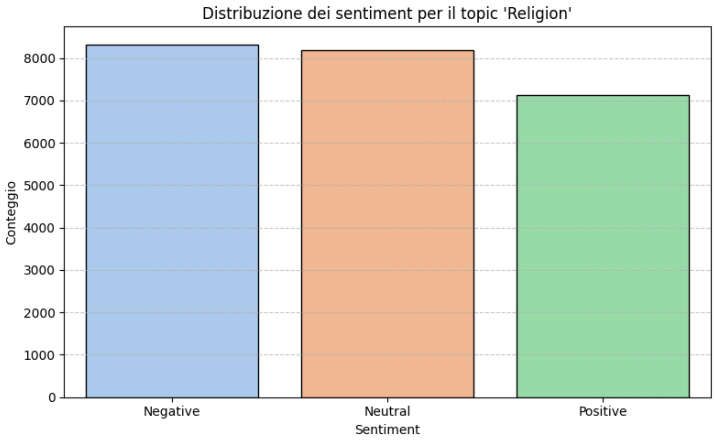

Figure 4.10: *Sentiment del topic Religion.*

I post con sentiment *Negative* sono, in genere, relativi all'insoddisfazione per alcuni dogmi religiosi oppure per alcune usanze, oltre che alle contaminazioni religiose. Molti post, come quello riportato per esempio, sono relativi a critiche verso altre religioni:

> Your deranged religion has a bigger body count than any other religion on earth from the invasion of Europe, Americas, Asia and Africa (in the tens of millions if not more). Not only that but your religion destroyed civilization in Europe for 1500 years and helped facilitate two different plagues.

Si riporta anche la lunghezza dei post per quanto riguarda il topic in esame in Figura 4.11. Anche in questo caso, come per il topic precedente, buona parte dei post si concentrano intorno alle 20 parole. Infatti, il 50% presenta un numero minore o uguale a 19 parole, mentre il 75% dei post arrivano da 0 a 34 parole. Tuttavia, la lunghezza massima è di 120 parole, segno che questo topic presenta dibattiti maggiori.

4.2.6 Topic: Climate Change

In questo caso, la maggior parte del sentiment collettivo è neutrale, ma una buona parte è negativo. Infatti, su un totale di 11.976 post, come si evidenzia anche nella Figura 4.12, si ha che:

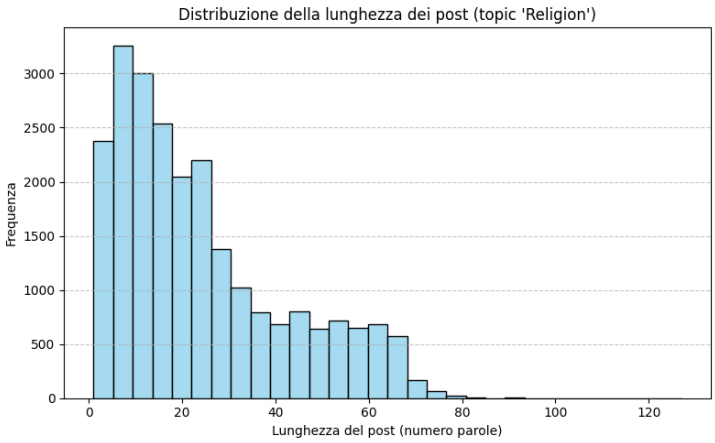

Figure 4.11: *Lunghezza dei post del topic Religion.*

- 5.566 (46.4%) hanno sentiment *Neutral*;
- 3.893 (32.5%) hanno sentiment *Negative*;
- 2.517 (21.1%) hanno sentiment *Positive*.

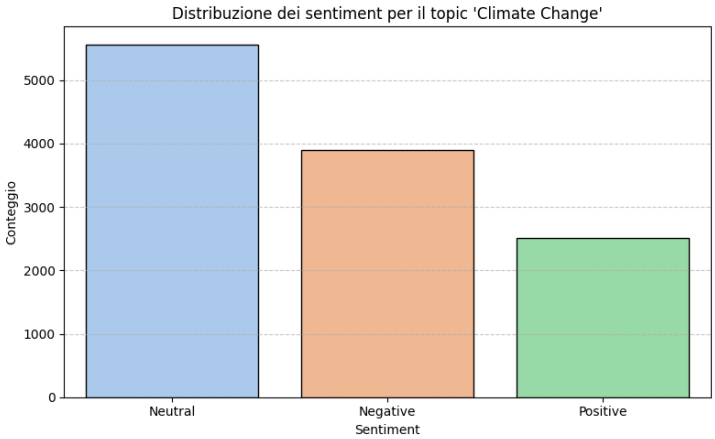

Figure 4.12: *Sentiment del topic Climate Change.*

I post con sentiment *Neutral* non hanno una particolare enfasi nell'esprimere preoccupazione riguardo il cambiamento climatico, piuttosto sembra che dichiarino dati di fatto: ad esempio, un post di questo tipo è

> Worth noting an announcement hidden in the small print on "last resort business interventions" (corporate bailouts). Companies receiving support will need to agree appropriate conditions, including on climate change. That's potentially very significant for high carbon emitters.

Un esempio di post con sentiment *Negative*, invece, è il seguente:

> Polluting industries want you to equate nominal steps to reduce the harm they do with them doing good even as their business model eats our futures. Don't buy it. Don't buy it when oil & gas companies pretend tiny slivers of their capex going to CCS will stop the climate crisis.

In quest'ultimo, è chiara l'enfasi dello scrittore, molto più accentuata rispetto al primo considerato.
Volendo valutare anche la lunghezza dei post relativamente a questo topic, si può osservare in Figura 4.13. In questo caso, l'andamento delle frequenze non è in ripida discesa, ma presenta picchi intorno ai 20 e tra le 40 e le 60 parole. Nello specifico, si ha che il 50% dei post hanno una lunghezza minore o uguale a 21 parole, mentre il 75% dei post ha una lunghezza minore o uguale a 37 parole. Anche la lunghezza massima (90 parole) non è molto elevata, in quanto i messaggi sono per lo più brevi e concisi, oppure trattano di constatazioni come gli esempi riportati.

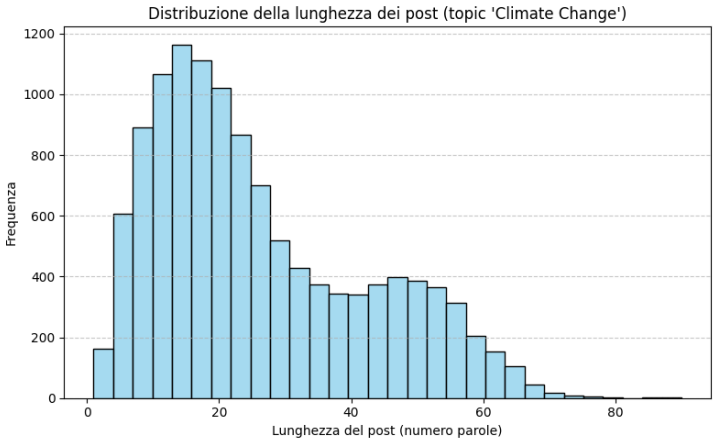

Figure 4.13: *Lunghezza dei post del topic Climate Change.*

4.2.7 Topic: Immigration

I post relativi a questo topic sono 11.249, e di questi:

- 4.855 (43.1%) hanno sentiment *Neutral*;
- 3.705 (32.9%) hanno sentiment *Negative*;
- 2.689 (24%) hanno sentiment *Positive*.

Come si vede anche nella Figura 4.14, non vi sono grandi sbilanciamenti tra i sentiment, ma la maggior parte dei post sono neutrali o pessimisti riguardo tale fenomeno (76%). Osservando i post, si intende che alcuni post neutrali sono sarcastici, oppure evidenziano di non essere contro all'immigrazione, ma di fatto lo sono, come in questo caso:

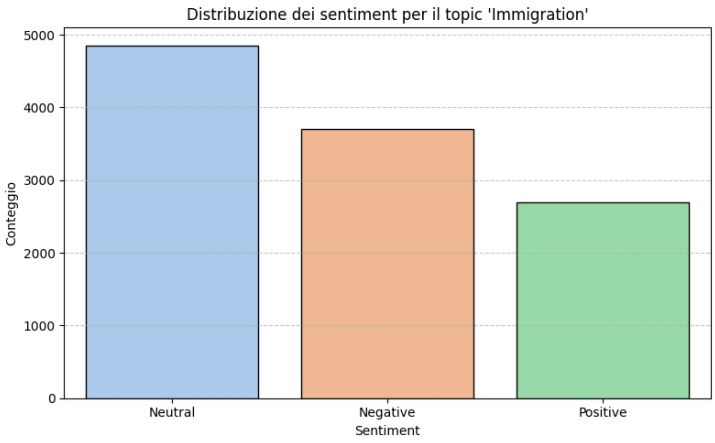

Figure 4.14: *Sentiment del topic Immigration.*

> I don't have a problem with these people having legitimate jobs, and coming in legitimately. But you know that's not happening currently and most of them coming illegally are criminals.

I post negativi, invece, riguardano principalmente i limiti imposti da Trump negli Stati Uniti d'America ed episodi di razzismo, sia del presente che del passato, sia noti che personali. Un esempio è il seguente:

> I'm here to get away from you racist intolerant tRump humpers

Per quanto riguarda, invece, la lunghezza dei post, viene riportato il grafico in Figura 4.15. Come nel caso precedente, sono presenti sia picchi più alti prima delle 20 parole e a ridosso di queste, che dopo tra le 40 e le 60 parole. Questo topic viene molto discusso nei media contemporanei, e infatti la lunghezza dei post risulta maggiore a quella di tutti i topic considerati fino a questo momento, raggiungendo le 40 parole per il 75% dei post.

4.2.8 Topic: War & Conflicts

Di questo topic fanno parte 6.507 post di cui, come si osserva anche dalla Figura 4.16:

- 2.613 (40.1%) hanno sentiment *Neutral*;
- 2.261 (34.7%) hanno sentiment *Negative*;
- 1.633 (25.2%) hanno sentiment *Positive*.

Analizzando i post, si osserva che i post neutrali discutono principalmente delle cause delle guerre e dei conflitti, dando delle motivazioni e cercando di dare spiegazioni a ciò. Per quanto riguarda i post negativi, parlano ovviamente degli orrori della guerra; alcuni post sono anche raccolte fondi. Un esempio preso da questi post è il seguente:

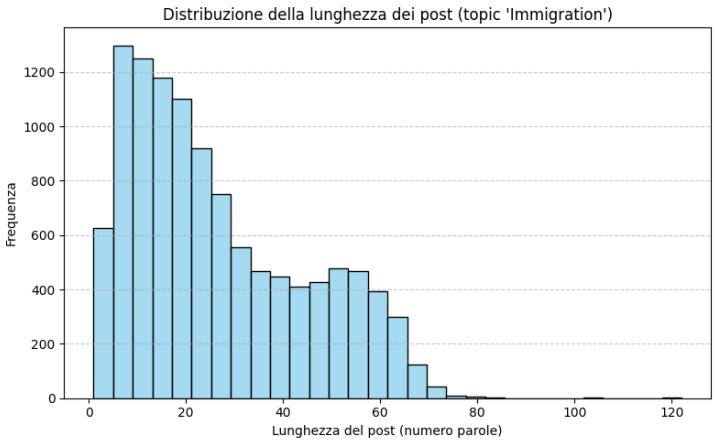

Figure 4.15: *Lunghezza dei post del topic Immigration.*

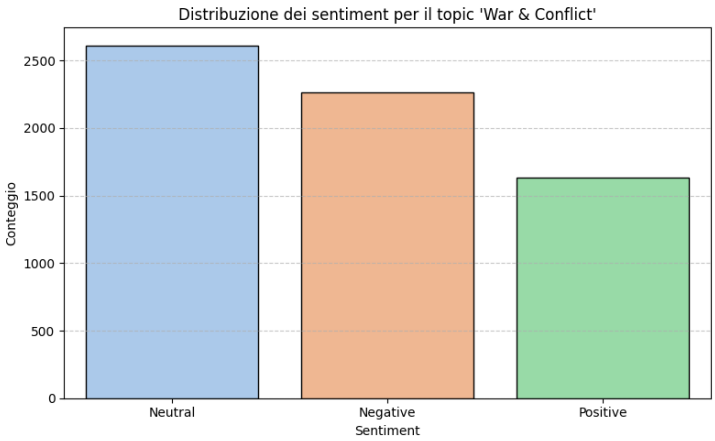

Figure 4.16: *Sentiment del topic War & Conflicts.*

> They never slaughtered each other. Just peaceful people.

Come questo post, ce ne sono altri in cui si racconta di persone pacifiche che sono costrette ad andare a combattere per volere dei loro superiori, e questo fa parte degli orrori della guerra. Per quanto riguarda i post positivi, parlano principalmente di ringraziamenti (per lo più ai soldati per il loro servizio) e speranza.
Analizzando anche la lunghezza di tali post (Figura 4.17), si nota che segue un andamento molto diverso da quelli visti finora. Infatti, ci sono picchi in salita fin quasi alle 60 parole, indicando la guerra come altro argomento molto discusso. La lunghezza massima raggiunta è di 83 parole che, sebbene non sia il numero più alto riscontrato, dimostra una maggiore omogeneità nelle lunghezze dei post. La maggior parte dei post (il 75%) presenta una lunghezza minore o uguale di 42 parole, il che significa che comunque viene discusso questo topic in maniera abbastanza esaustiva.

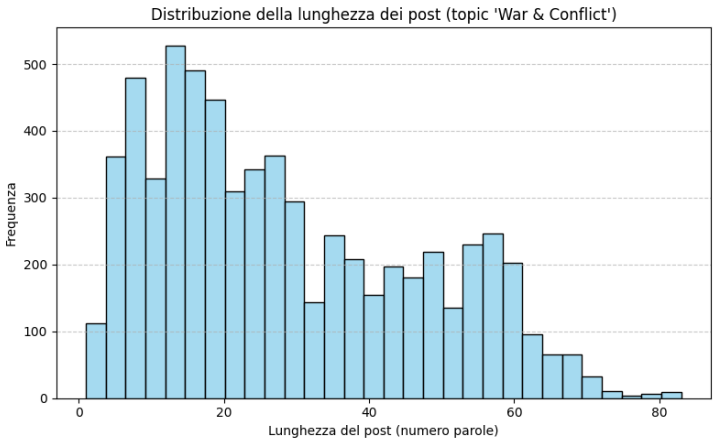

Figure 4.17: *Lunghezza dei post del topic War & Conflicts.*

4.3 Metriche

Per valutare le prestazioni dei modelli nel task di sentiment analysis, sono state utilizzate le seguenti metriche:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

dove:

- TP (True Positives) è il numero di istanze positive correttamente classificate,
- TN (True Negatives) è il numero di istanze negative correttamente classificate,
- FP (False Positives) è il numero di istanze negative classificate erroneamente come positive,
- FN (False Negatives) è il numero di istanze positive classificate erroneamente come negative.

L'accuracy misura la proporzione complessiva di classificazioni corrette, ma può essere influenzata da dataset sbilanciati.

$$\text{precision} = \frac{TP}{TP + FP} \tag{4.2}$$

La precisione indica la percentuale di istanze classificate come positive che risultano effettivamente positive.

$$\text{recall} = \frac{TP}{TP + FN} \tag{4.3}$$

La recall rappresenta la capacità del modello di identificare correttamente tutte le istanze positive presenti nel dataset.

$$\text{f1-score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \tag{4.4}$$

L'F1-score è la media armonica tra precision e recall, fornendo una misura bilanciata

che considera entrambi gli aspetti.

4.4 Modelli "from scratch"

4.4.1 RandomForest

I risultati sono stati valutati su un test set costituito da 100.000 istanze campionate in modo stratificato. I valori delle metriche riscontrati sono i seguenti:

- Accuracy media: 0.6014;
- F1-Score media: 0.5965;
- Precision media: 0.6050;
- Recall media: 0.6014.

L'accuratezza pari a 0.60 mostra una capacità predittiva discreta, e un'analisi più approfondita viene riportata nella Tabella 4.2.

Table 4.2: *Report di classificazione sul test set per il classificatore Random Forest*

Classe	Precision	Recall	F1-score	Supporto
0 (Negativo)	0.59	0.41	0.48	2.503
1 (Neutro)	0.55	0.66	0.60	3.646
2 (Positivo)	0.67	0.67	0.67	3.851
Media macro	0.60	0.58	0.59	91 562
Media pesata	0.61	0.60	0.60	91 562

Volendo visualizzare una matrice di confusione, viene riportata quella relativa al test set (Figura 4.18), essendo quella del validation set molto simile. Come si evince dalla matrice di confusione, le prestazioni sono buone in linea generale. Sicuramente il modello è più abile a classificare correttamente il sentiment *Neutrale* o *Positivo*, essendo molto spesso che predice la classe neutrale anche in caso di post con sentiment negativo. Il motivo può riscontrarsi nel fatto che i post neutrali e negativi sono comunque molto simili, essendo che i primi possono essere relativi comunque a lamentele o scontento da parte dell'autore del post. Volendo riepilogare il numero di istanze classificate correttamente sul totale per ciascuna classe, si ha che:

- per la classe *Negative*, sono state predette correttamente il 40.79% delle istanze totali;
- per la classe *Neutral*, ne sono state predette correttamente il 66.45%;
- per la classe *Positive*, sono state predette in maniera esatta il 67.43%.

4.4.2 Classificatore Multi-layer Perceptron

Il classificatore MLP, descritto nella Sezione 3.2.3, è stato valutato su un insieme di test costituito da circa 100 000 istanze campionate in modo stratificato. I risultati ottenuti mostrano una discreta capacità predittiva caratterizzata da un'**accuratezza pari**

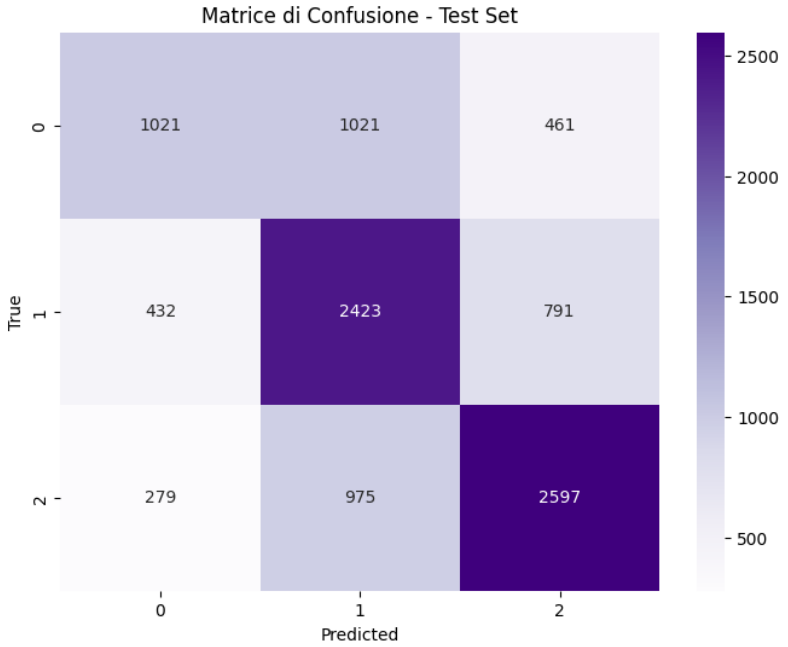

Figure 4.18: *Matrice di confusione per il Random Forest sul test set.*

a 0.6559. L'analisi più approfondita delle prestazioni per ciascuna classe di sentiment è riportata nella Tabella 4.3, mentre la Figura 4.19 mostra la corrispondente matrice di confusione.

Table 4.3: *Report di classificazione sul test set per il classificatore MLP*

Classe	Precision	Recall	F1-score	Supporto
0 (Negativo)	0.60	0.56	0.58	22 918
1 (Neutro)	0.63	0.64	0.63	33 388
2 (Positivo)	0.72	0.74	0.73	35 256
Media macro	0.65	0.64	0.65	91 562
Media pesata	0.65	0.66	0.65	91 562

Dalla matrice di confusione in Figura 4.19 e dalle metriche elencate nella tabella soprastante si osserva una buona capacità del modello nel riconoscere le istanze della classe positiva (etichetta 2), con oltre 28 000 predizioni corrette su 38.506 esempi e tassi di precision e recall superiori al 0.7. Le prestazioni risultano invece più contenute per le classi negativa e neutra, che presentano una maggiore confusione reciproca. In particolare, molte istanze negative vengono erroneamente classificate come neutre, e viceversa, come suggerito dalla presenza significativa di valori fuori diagonale nelle prime due righe della matrice.

4.4.3 Naive Bayes

Il classificatore Naive Bayes, descritto nella Sezione 3.2.4, ha mostrato prestazioni sorprendentemente solide, nonostante la sua semplicità e l'assunzione di indipendenza

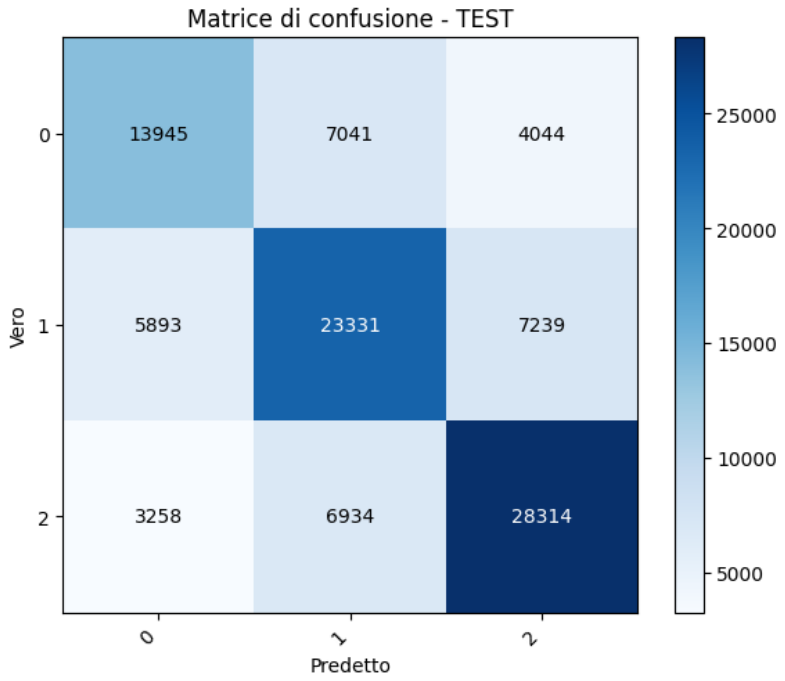

Figure 4.19: *Matrice di confusione per il classificatore MLP sul test set.*

tra le caratteristiche. I risultati riportati in Tabella 4.4 mostrano valori comparabili a quelli ottenuti dal classificatore MLP: media della precision pari a 0.63 contro 0.65, media del recall pari a 0.62 contro 0.64 e media dell'F1-score di 0.62 contro 0.65.

Table 4.4: *Report di classificazione sul test set per Naive Bayes.*

Classe	Precision	Recall	F1-score	Supporto
0 (Negativo)	0.58	0.54	0.56	22 918
1 (Neutro)	0.62	0.58	0.60	33 388
2 (Positivo)	0.67	0.75	0.71	35 256
Media macro	0.63	0.62	0.62	91 562
Media pesata	0.63	0.64	0.63	91 562

Dall'analisi della matrice di confusione (Figura 4.20) emerge una maggiore efficacia del modello nel classificare correttamente le istanze appartenenti alla classe "positiva". Questo comportamento può essere spiegato, almeno in parte, dal leggero sbilanciamento del dataset a favore di questa classe, che consente una stima più accurata delle probabilità condizionate utilizzate dal modello.

4.4.4 RNN

L'RNN, descritta nella Sezione 3.2.5, ha mostrato prestazioni significativamente inferiori rispetto agli altri modelli, su tutte le metriche considerate. Si osserva tuttavia una leggera preferenza per la classe positiva, per la quale si registra un miglioramento medio di circa il 6% su precision, recall e F1-score. Questo comportamento è probabilmente riconducibile al lieve sbilanciamento del dataset, in cui la classe positiva risulta

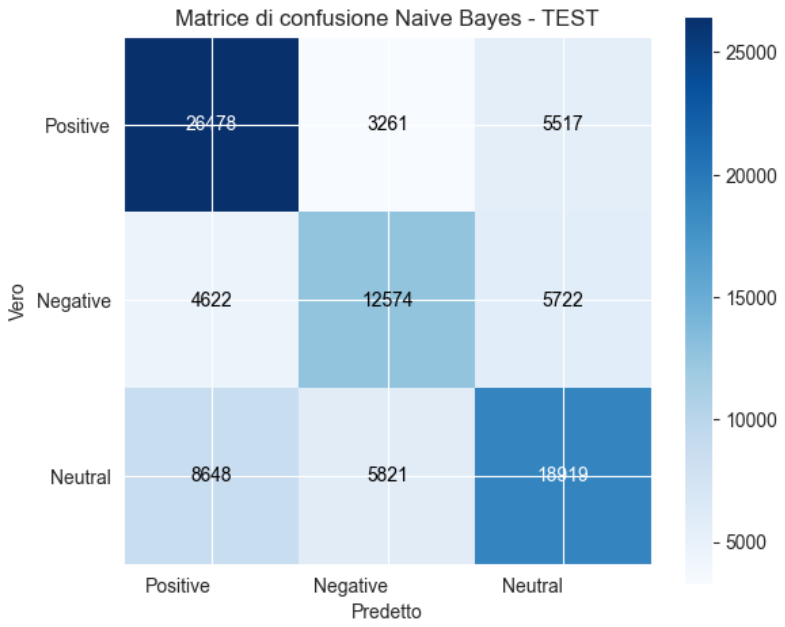

Figure 4.20: *Matrice di confusione per Naive Bayes sul test set.*

leggermente più rappresentata. I risultati completi sono riportati in Tabella 4.5. Ulte-

Table 4.5: *Report di classificazione sul test set per la RNN.*

Classe	Precision	Recall	F1-score	Supporto
0 (Negativo)	0.27	0.31	0.29	22 918
1 (Neutro)	0.37	0.41	0.39	33 388
2 (Positivo)	0.47	0.37	0.42	35 256
Media macro	0.37	0.36	0.36	91 562
Media pesata	0.38	0.37	0.37	91 562

riori evidenze delle difficoltà del modello emergono dall'analisi della matrice di confusione (Figura 4.21): la rete risulta particolarmente inaffidabile nel distinguere tra le classi "negativo" e "neutrale", che tende spesso a confondere. Ciò indica una generale difficoltà della rete nel catturare le sfumature semantiche tra classi con caratteristiche simili. Nonostante la maggiore complessità architetturale della RNN rispetto ai modelli più semplici considerati, le sue prestazioni inferiori potrebbero essere attribuite a un periodo di addestramento insufficiente. Questa ipotesi è supportata dall'analisi della curva di apprendimento (Figura 4.22), che mostra una riduzione costante della funzione di perdita, senza indicazioni di convergenza o stabilizzazione.

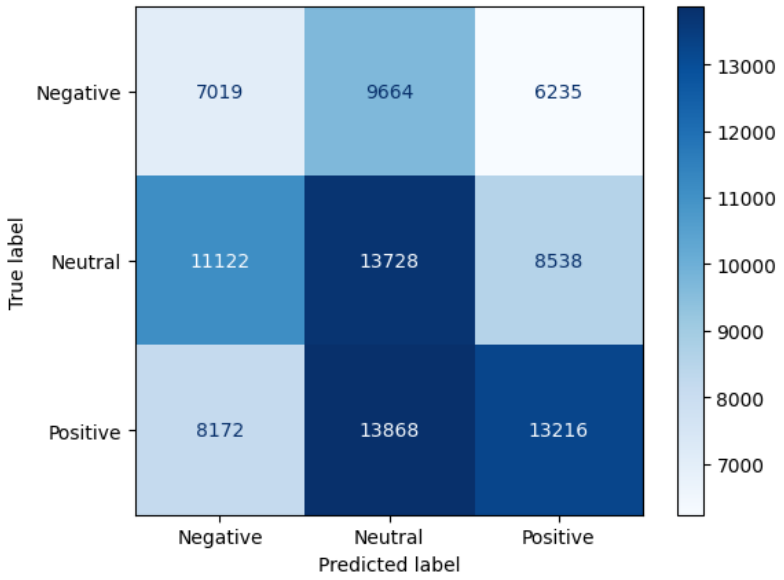

Figure 4.21: *Matrice di confusione per la RNN sul test set.*

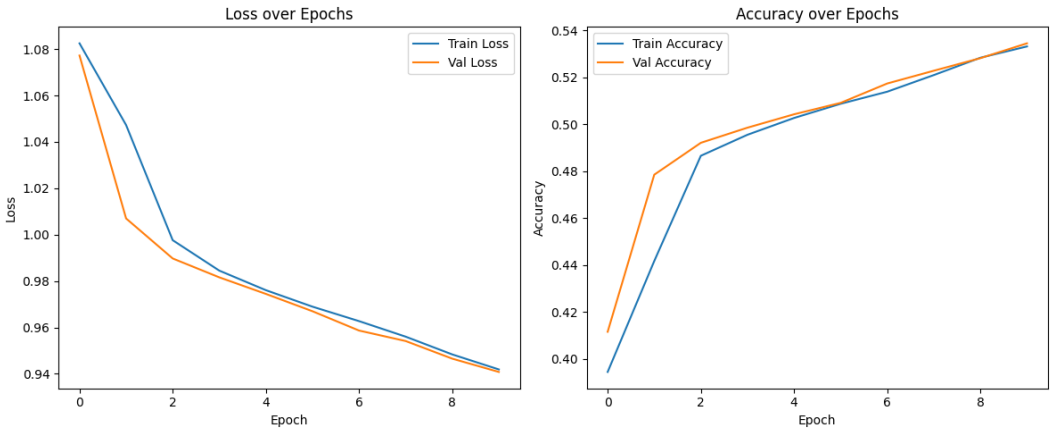

Figure 4.22: *Curve di apprendimento per la RNN durante l'addestramento.*

4.5 Modelli preaddestrati

4.5.1 BERTweet

Il modello BERTweet ha raggiunto un'accuratezza complessiva pari al 0.59 e un F1-score di 0.59. Tali risultati possono essere considerati soddisfacenti se si considera la complessità intrinseca del linguaggio utilizzato nei social media e la natura multidimensionale dei contenuti analizzati. La Tabella 4.6 sintetizza le principali metriche calcolate, suddivise per classe, evidenziando una variabilità nelle metriche di precisione, recall e F1-score tra le diverse categorie sentimentali.

L'esame della matrice di confusione, rappresentata in Figura 4.23, evidenzia alcuni pattern distintivi nel comportamento del modello. La classe associata ai sentimenti positivi presenta la miglior performance complessiva, con un F1-score di 0.66, accom-

Classe	Precision	Recall	F1-score	Supporto
Negative	0.56	0.54	0.55	22.918
Neutral	0.53	0.59	0.56	33.388
Positive	0.69	0.63	0.66	35.256
Macro avg	0.59	0.59	0.59	91.562
Weighted avg	0.60	0.59	0.59	91.562

Table 4.6: *Metriche di classificazione di BERTweet sul test set*

pagnata da una precisione pari a 0.69, mentre la recall è leggermente inferiore a 0.63. In questa categoria, il numero di identificazioni corrette si attesta su circa 24.700 esempi, sebbene siano presenti errori di confusione principalmente con la classe neutra, che genera oltre 15.900 falsi negativi. La categoria neutra, invece, mostra prestazioni intermedie con un F1-score di 0.56, caratterizzata da una recall superiore al 59% e da una precisione inferiore a 0.53, indicando una tendenza ad assorbire errori provenienti dalle classi adiacenti. Infine, la classe negativa presenta difficoltà nell'identificazione, con una recall pari a 0.54, una considerevole quantità di falsi negativi erroneamente classificati come neutri (circa 19.600), e una sovrastima delle istanze positive, con oltre 6.500 falsi positivi.

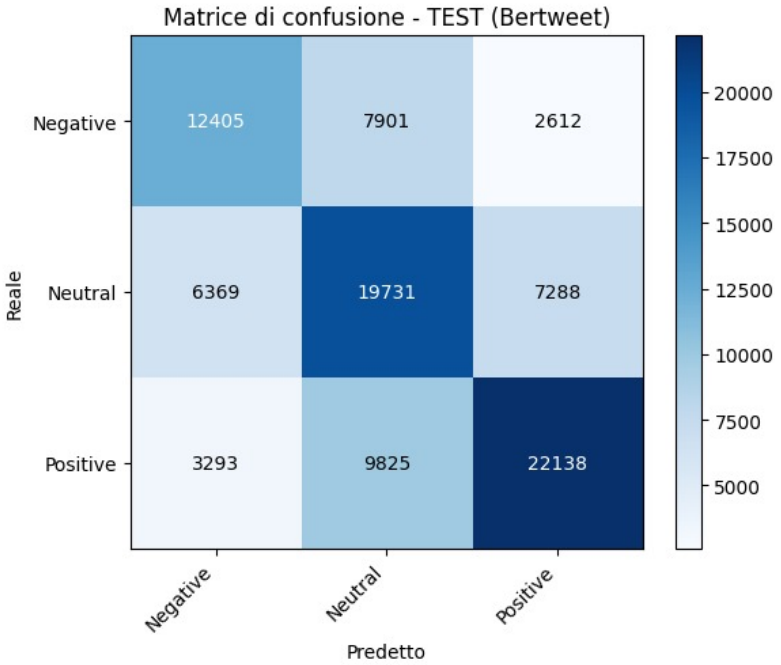

Figure 4.23: *Matrice di confusione per BERTweet sul test set.*

Nel confronto con altri modelli costruiti da zero, BERTweet presenta una diminuzione di accuratezza e di F1-score rispetto ai modelli RandomForest, MLPClassifier e Naive Bayes. La sua superiorità risulta particolarmente marcata se paragonata alla rete neurale ricorrente, dalla quale distacca con un margine del 21%. Inoltre, le prestazioni di BERTweet risultano comparabili a quelle ottenute dal modello RandomForest, con

differenze marginali nell'ordine dello 0.1%. La Tabella 4.8 riporta un confronto generale tra tutti i modelli.

4.5.2 RoBERTa

Le prestazioni ottenute dal modello basato su RoBERTa, riportate in Tabella 4.7, sono pressoché identiche a quelle ottenute da BERTweet, con uno scarto percentuale medio dell' 1% a favore del primo.

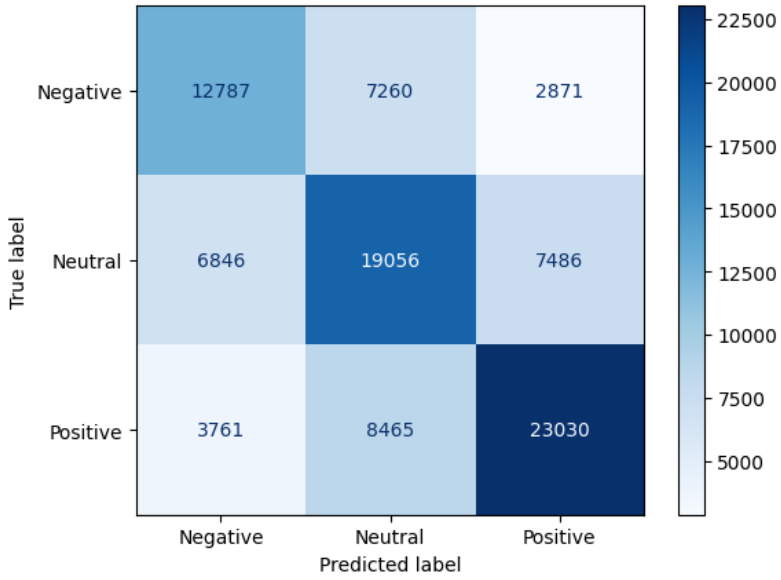

Figure 4.24: *Matrice di confusione per RoBERTa sul test set.*

Analizzando e confrontando le due matrici di confusione (Figura 4.23 e Figura 4.24), si notano delle leggere discrepanze tra i due, come suggerito dai risultati quantitativi. In particolare, RoBERTa sembra essere più propense alla classificazione di testi con sentimento positivo e negativo, mentre BERTweet si comporta meglio nella classificazione di testi con sentimento negativo. Il confronto con gli altri modelli segue gli stessi andamenti di quanto descritto precedentemente per BERTweet. La Tabella 4.8 riporta un confronto generale tra tutti i modelli.

Classe	Precision	Recall	F1-score	Supporto
Negative	0.55	0.56	0.55	22.918
Neutral	0.55	0.57	0.56	33.388
Positive	0.69	0.65	0.67	35.256
Macro avg	0.59	0.59	0.59	91.562
Weighted avg	0.60	0.60	0.60	91.562

Table 4.7: *Metriche di classificazione di RoBERTa sul test set.*

4.6 Confronto Prestazionale

In questa sezione si propone un'analisi comparativa delle performance dei modelli addestrati, con l'obiettivo di evidenziare il compromesso (trade-off) tra accuratezza, complessità del modello, semplicità architetturale e tempo di addestramento. La Tabella 4.8 riassume le metriche principali, includendo anche il numero di parametri e il tempo necessario per l'addestramento.

Prima di procedere con la discussione, è opportuno precisare che, per alcuni modelli, il numero di "parametri" riportato non corrisponde al numero di parametri addestrabili nel senso stretto del termine. In particolare:

- **Random Forest**: il numero indicato si riferisce al totale dei nodi massimi costruiti dagli alberi nella foresta.
- **Naive Bayes**: la quantità riportata rappresenta la dimensione del vocabolario utilizzato durante l'addestramento, ossia il numero di feature (parole uniche) su cui è calcolata la probabilità.

I modelli neurali pre-addestrati (BERTweet e RoBERTa) mostrano buone prestazioni, ma non eccellono rispetto a modelli molto più semplici, come il Naive Bayes o l'MLP. Questo è particolarmente sorprendente considerando il numero di parametri: a fronte di oltre 100 milioni di parametri, entrambi i modelli basati su Transformer non superano un'accuratezza pari a 0.60. Al contrario, il classificatore **MLP**, con appena 1101 parametri e un tempo di addestramento contenuto (2 ore), raggiunge le prestazioni migliori su tutte le metriche. Questo risultato suggerisce che il modello è stato in grado di sfruttare efficacemente le informazioni semantiche veicolate dagli embeddings, integrandole con le feature numeriche aggiuntive, risultando così particolarmente competitivo nonostante la sua semplicità architetturale. Il classificatore Naive Bayes si conferma una soluzione leggera e sorprendentemente competitiva: con un solo minuto di addestramento e meno di un milione di parametri, raggiunge un'accuratezza del 0.60, risultando ideale in scenari a basse risorse computazionali. La RNN, nonostante l'elevata complessità (16 milioni di parametri) e il tempo di addestramento più lungo tra i modelli (15 ore), mostra le prestazioni peggiori. Questo risultato può essere ricondotto a un addestramento non ancora sufficiente, come discusso nella sezione dedicata. Infine, il modello Random Forest rappresenta un buon compromesso tra semplicità e prestazioni, con un numero ridotto di parametri e risultati paragonabili ai modelli Transformer, a fronte però di tempi di addestramento più lunghi rispetto a modelli di pari complessità.

Nel complesso, l'analisi evidenzia che modelli più grandi e complessi non sempre garantiscono un miglioramento prestazionale proporzionale. La scelta del modello più adatto dipende quindi fortemente dai vincoli computazionali e dai requisiti specifici dell'applicazione. In Figura 4.25 è fornita una visualizzazione più diretta della differenza nelle dimensioni e nelle prestazioni del modelli.

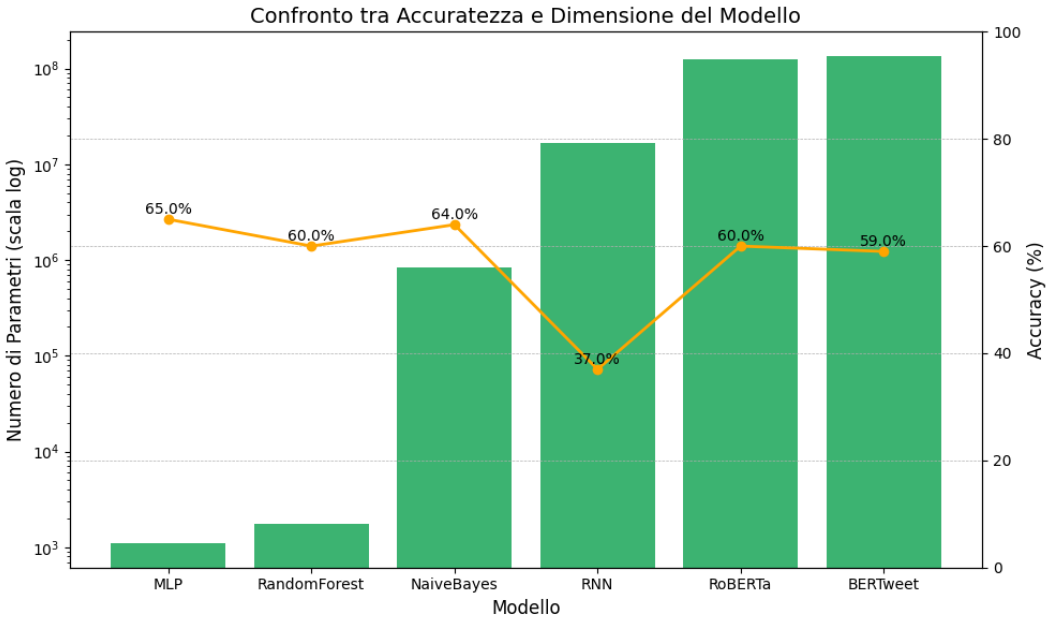

Figure 4.25: *Tradeoff prestazioni vs numero parametri.*

Modello	Accuracy	Precision	Recall	F1-Score	Addestramento	# Parametri
RandomForest	0.60	0.58	0.59	0.60	4 ore	1746
MLPClassifier	**0.65**	**0.65**	**0.66**	**0.65**	2 ore	1101
Naive Bayes	0.63	0.63	0.64	0.63	1 minuto	835K
RNN	0.38	0.38	0.37	0.37	15 ore	16M
BERTweet	0.59	0.60	0.59	0.59	-	134M
RoBERTa	0.59	0.60	0.59	0.59	-	124M

Table 4.8: *Confronto prestazionale tra modelli. In tabella sono riportate le medie pesate per precision, recall e f1-score. I risultati migliori sono evidenziati in grassetto. La scala utilizzata per il numero di parametri è: K = migliaia, M = milioni.*

5 Conclusioni

In questo lavoro è stata proposta un'analisi comparativa di modelli di classificazione caratterizzati da diversa complessità architetturale e computazionale, applicati al task di sentiment analysis su un sottoinsieme di post estratti dal social network Bluesky. In particolare, sono stati addestrati e valutati i modelli Random Forest, Naive Bayes, MLP eRNN, insieme a due modelli linguistici pre-addestrati ampiamente utilizzati nello stato dell'arte: BERTweet [Nguyen et al., 2020] e RoBERTa [Barbieri et al., 2023], entrambi fine-tunati sul task di sentiment-analysis.

I risultati ottenuti sono stati inaspettati: modelli meno complessi, come l'MLP e il Naive Bayes, hanno superato in prestazioni modelli più sofisticati e costosi in termini di addestramento, come la RNN. Tali risultati sottolineano l'importanza di non assumere una correlazione diretta tra dimensione del modello e performance, specialmente in contesti con dati ben strutturati e risorse limitate. L'intera valutazione è stata guidata da una prospettiva pragmatica, evidenziando il trade-off tra accuratezza e complessità modellistica, con l'obiettivo di identificare soluzioni che possano garantire un buon bilanciamento tra efficacia predittiva ed efficienza operativa.

Nei lavori futuri ci proponiamo di:

- Prolungare la fase di addestramento della RNN, al fine di valutare appieno il suo potenziale di generalizzazione, non ancora espresso a causa del training prematuro;
- Sperimentare architetture neurali più avanzate, come LSTM [Hochreiter et al., 1997], GRU o Transformer [Vaswani et al., 2023], per verificare se una modellazione più sofisticata della sequenza testuale possa migliorare le prestazioni;
- Integrare embeddings semantici pre-addestrati nel processo di training del Random RNN, al fine di valutare l'impatto di rappresentazioni linguistiche più ricche sulla qualità della classificazione.
- Valutare l'uso di embedding statici pre-addestrati a livello di parola, come Word2Vec o GloVe, da aggregare a livello di post. Dato che i testi analizzati sono molto brevi, il contesto a disposizione potrebbe essere troppo limitato per sfruttare davvero

i vantaggi degli embedding contestualizzati (come quelli di *all-MiniLM-L6-v2*). In questo scenario, rappresentazioni più semplici e stabili potrebbero risultare più adatte, soprattutto se abbinate a modelli tradizionali come Random Forest o MLP.

Questa pagina è intenzionalmente bianca.

Bibliografia

Alibaba DAMO Academy (2024). *GTE-small: General Text Embeddings (small) model from Alibaba DAMO Academy*. https://huggingface.co/thenlper/gte-small. Accessed: 2025-06-18. Hugging Face.

Barbieri, Francesco et al. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification*. arXiv: 2010.12421 [cs.CL]. URL: https://arxiv.org/abs/2010.12421.

Barbieri, Francesco et al. (2023). *twitter-roberta-base-sentiment-latest*. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. Hugging Face.

Devlin, Jacob et al. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Elman, Jeffrey L. (1990). "Finding Structure in Time". In: *Cognitive Science* 14.2, pp. 179–211. DOI: https://doi.org/10.1207/s15516709cog1402_1. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1.

Failla, Andrea and Giulio Rossetti (Nov. 2024). ""I'm in the Bluesky Tonight": Insights from a year worth of social data". In: *PLOS ONE* 19.11. Ed. by Fabio Saracco, e0310330. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0310330. URL: http://dx.doi.org/10.1371/journal.pone.0310330.

Fellbaum, Christiane, ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Hu, Minqing and Bing Liu (2004). "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177.

Jeong, Ujun et al. (2024). "Descriptor: A Temporal Multi-network Dataset of Social Interactions in Bluesky Social (BlueTempNet)". In: *IEEE Data Descriptions* 1, pp. 71–79. DOI: 10.1109/IEEEDATA.2024.3474640.

Kapur, Keshav and Rajitha Harikrishnan (2022). "Comparative study of sentiment analysis for multi-sourced social media platforms". In: *arXiv preprint arXiv:2212.04688*.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

M, Ali (2023). *Bluesky Posts Dataset*. https://huggingface.co/datasets/withalim/bluesky-posts.

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

Miller, George A. (1995). "WordNet: A Lexical Database for English". In: *Communications of the ACM* 38.11, pp. 39–41.

Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen (2020). "BERTweet: A pre-trained language model for English Tweets". In: *arXiv preprint arXiv:2005.10200*.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162/.

Schuster, M. and K.K. Paliwal (1997). "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11, pp. 2673–2681. DOI: 10.1109/78.650093.

Sentence-Transformers Team (2023). *all-MiniLM-L6-v2: Sentence-Transformers sentence embedding model*. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2025-06-18. DOI: 10.5281/zenodo.8045524. URL: https://doi.org/10.5281/zenodo.8045524.

Sudhir, Prajval and Varun Deshakulkarni Suresh (2021). "Comparative study of various approaches, applications and classifiers for sentiment analysis". In: *Global Transitions Proceedings* 2.2, pp. 205–211.

TabularisAI (2024). *tabularisai/multilingual-sentiment-analysis: DistilBERT-based multilingual sentiment classification (5-class)*. https://huggingface.co/tabularisai/multilingual-sentiment-analysis. Accessed: 2025-06-18.

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Questa pagina è intenzionalmente bianca.