Università Politecnica delle Marche

Facoltà di Ingegneria

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

---

Applicazione tecniche di NLP su un dataset contentente Tweet e commenti di Facebook, Instagram e Twitter

Docenti

Prof. Ursino Domenico
Dott. Marchetti Michele
Dott. Buratti Christopher

Componenti del gruppo

Bellante Luca
Coccia Giansimone
Ferretti Laura

Anno Accademico 2024-2025

Contents

1 Introduzione al Natural Language Processing (NLP)

Il *Natural Language Processing (NLP)* è un campo dell'intelligenza artificiale che si occupa dell'interazione tra computer e linguaggio umano. L'obiettivo del NLP è permettere ai computer di comprendere, analizzare, generare e rispondere al linguaggio naturale in modo significativo e utile.

Componenti fondamentali del NLP sono:

1. **Comprensione del Linguaggio Naturale (NLU - Natural Language Understanding):**
   - Si concentra sul comprendere il significato del testo scritto o parlato.
   - Include attività come il riconoscimento delle entità (NER - Named Entity Recognition), l'analisi del sentiment e la comprensione semantica.
2. **Generazione del Linguaggio Naturale (NLG - Natural Language Generation):**
   - Consente ai computer di creare testi comprensibili e coerenti.
   - Esempi di applicazione sono i sistemi di risposta automatica, i riassunti automatici e la traduzione.
3. **Elaborazione del Linguaggio Naturale (NLP - core processing):**
   - Riguarda il trattamento dei dati testuali grezzi per estrarre informazioni utili.
   - Include compiti come la tokenizzazione, la stemming/lemmatizzazione, il parsing sintattico e la rimozione di stop-word.

Tali ambiti possono essere raggiunti attraverso tecniche specifiche ed efficaci, come ad esempio:

- **Metodi Statistici:** Utilizzano modelli probabilistici per analizzare i testi (ad esempio, modelli di Markov nascosti o bag-of-words).
- **Apprendimento Automatico (Machine Learning):** Modelli supervisionati o non supervisionati per compiti specifici, come classificazione o clustering dei testi.
- **Reti Neurali e Deep Learning:** Tecniche avanzate che includono reti neurali ricorrenti (RNN), LSTM, GRU e i modelli basati su trasformer come BERT e GPT.

Al giorno d'oggi l'NLP ha trovato un largo margine di applicazione, basti pensare alle seguenti categorie di tecnologie introdotte:

- **Chatbot e Assistenti Virtuali:** Sistemi come Alexa, Siri e Google Assistant.
- **Motori di Ricerca:** Ottimizzazione delle query e generazione di risultati rilevanti.
- **Traduzione Automatica:** Strumenti come Google Translate.
- **Analisi del Sentiment:** Utilizzata nel marketing e nel monitoraggio dei social media.
- **Riconoscimento Vocale:** Trascrizione e interpretazione di comandi vocali.

2 Introduzione al progetto

Negli ultimi anni, l'analisi del linguaggio naturale (NLP) ha assunto un ruolo sempre più centrale nello studio dei dati testuali provenienti dai social media. Le piattaforme come Facebook, Instagram e Twitter generano quotidianamente enormi quantità di testo, offrendo una preziosa fonte di informazioni per comprendere opinioni, tendenze e comportamenti degli utenti.

L'analisi di questi dati consente di ottenere insight significativi, sia a livello individuale che collettivo. Queste informazioni possono essere impiegate in numerosi ambiti, dalla ricerca di mercato all'analisi dell'opinione pubblica, fino alla prevenzione della disinformazione.

2.1 NLP: Dataset utilizzato

Il dataset utilizzato per il nostro lavoro è reperibile al seguente link: Dataset Social Media.

Questo dataset, disponibile su Kaggle, è un insieme di dati utilizzato per analisi di sentiment sui social media. Contiene una raccolta di post di piattaforme di social media etichettati in base al sentimento espresso (positive, negative, neutral, joy, fear, ecc...). Questo dataset è particolarmente utile per addestrare e testare modelli di Natural Language Processing (NLP) per attività di analisi del sentiment.

Di seguito è riportata una tabella con le feature incluse nel dataset e una breve descrizione del loro significato:

Feature	**Descrizione**
`Text`	Contenuti generati dagli utenti.
`Sentiment`	Tipo di sentimento che esprime il Text.
`Timestamp`	Informazioni data e ora dei contenuti.
`User`	Utente che ha generato il contenuto.
`Platform`	Piattaforma social sulla quale è stato pubblicato il contenuto.
`Hashtags`	Identifica gli argomenti e i temi di tendenza
`Likes`	Numero di like espressi.
`Retweets`	Riflette la popolarità del contenuto.
`Country`	Paese di origine di ciascun post.
`Year`	Anno della pubblicazione.
`Month`	Mese della pubblicazione.
`Day`	Giorno della pubblicazione.
`Hour`	Ora della pubblicazione.

Table 1: Descrizione delle feature del dataset di analisi

2.2 Pre-processing e analisi descrittiva

Prima di procedere con lo sviluppo delle analisi, sono stati apportati dei lavori di modifica al dataset. In particolare sono state eliminate due colonne *"Unnamed"* riportanti degli indici non significativi, e sostituite con un'unica colonna *ID* impostata come indice del dataframe.

Inoltre, è stata eseguita una pulizia approfondita dei dati, rimuovendo spazi superflui e risolvendo problemi di duplicazione di alcune feature, causati dalla presenza di spazi aggiuntivi nei valori testuali.

Successivamente, è stata condotta un'analisi descrittiva del dataset per esaminare la distribuzione e la tipologia dei dati disponibili, fornendo così una visione più chiara della loro struttura e delle caratteristiche principali.

Innanzitutto, abbiamo effettuato un'analisi esplorativa visualizzando la distribuzione della tipologia di sentimenti presenti all'interno del nostro dataset (vedi Figura 1). Per motivi di spazio il plotting è stato ridotto ai soli primi quaranta elementi in ordine di distribuzione.

Dall'analisi emerge che i sentiment predominanti sono quelli positivi, tra cui gioia, eccitazione e appagamento. Questo suggerisce che la maggior parte dei contenuti a disposizione tende a esprimere emozioni favorevoli, il che potrebbe riflettere una propensione generale degli utenti social a condividere esperienze positive. Tuttavia, è importante considerare che questa distribuzione potrebbe essere influenzata dalle piattaforme considerate e dal comportamento degli utenti su di esse.

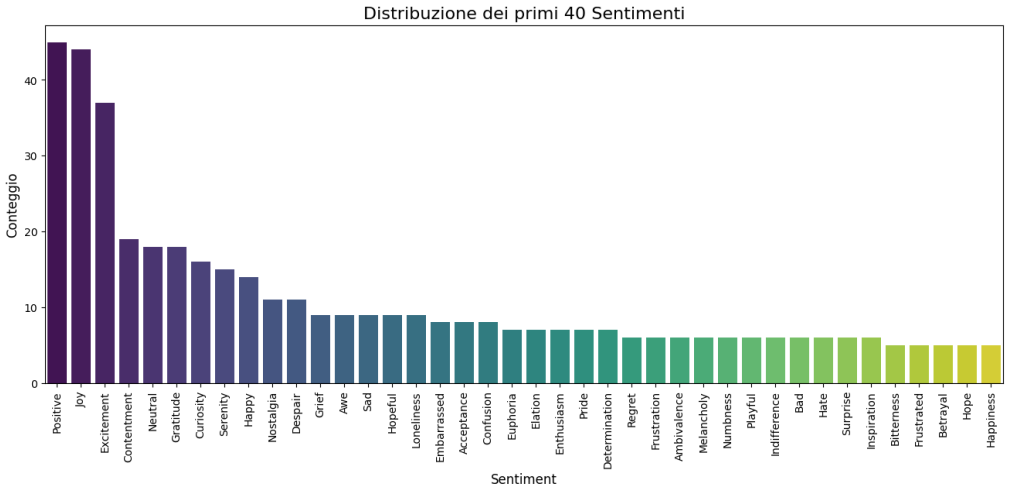

Figure 1: Distribuzione dei primi 40 sentimenti presenti nel dataset.

In secondo luogo è stata fatta un'analisi per piattaforma social, andando a vedere il numero di contenuti presenti su ciascuna delle piattaforme social proposte. Come possiamo osservare in Figura 2 i contenuti sono presenti in quantità piuttosto bilanciate;

il social media che vanta il maggior numero di contenuti è Instagram, seguito da Twitter e poi Facebook.

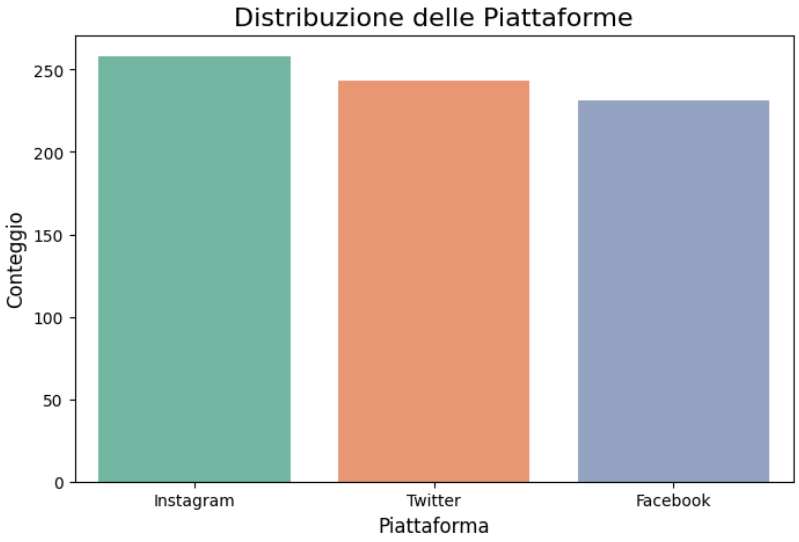

Figure 2: Distribuzione dei contenuti sulle tre piattaforme Facebook, Twitter e Instagram

In seguito a quest'ultima analisi sono stati effettuati ulteriori approfondimenti in merito al numero di contenuti caricati ad intervalli di ore (vedi Figura 3), nel quale risulta un boom di interazioni intorno alle ore quattordici, ed in base al numero di contenuti postati in base al paese di origine.

Questa distribuzione, riportata in Figura 4, può essere influenzata da diversi fattori, tra cui la diffusione dei social media nei vari paesi, il numero di abitanti ma, soprattutto, dalla lingua del dataset. Poiché il dataset analizzato contiene esclusivamente post in lingua inglese, è naturale che i paesi con il maggior numero di contenuti siano quelli anglofoni, come Stati Uniti, Regno Unito, Canada, Australia e India. Questo aspetto potrebbe aver limitato la rappresentatività di altre nazioni, dove l'uso dell'inglese nei social media è meno prevalente.

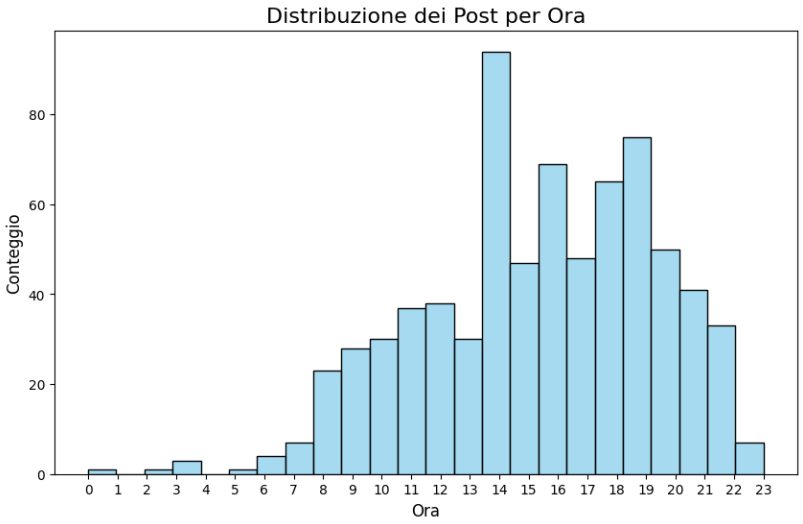

Figure 3: Numero di contenuti caricati per ora

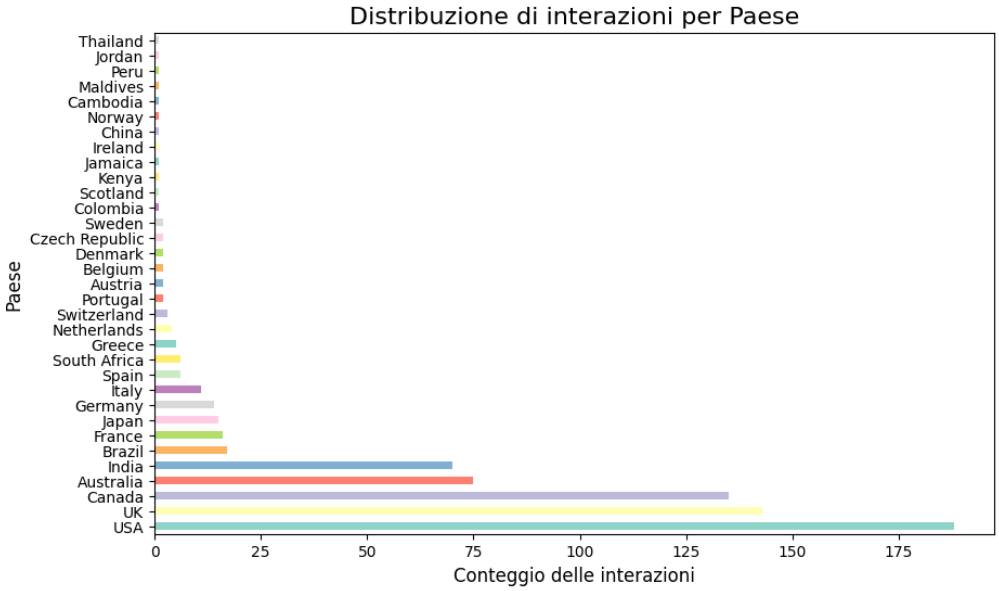

Figure 4: Numero contenuti caricati per paese di provenienza

Infine, come ultima analisi descrittiva, abbiamo studiato il trend dei top dieci sentimenti dal 2010 ad oggi (vedi Figura 5). Si può notare una certa omogeneità ed oscillazione piuttosto controllata dei vari sentimenti, tranne per quelli *Positive*, che presentano un picco ben visibile nell'anno 2023. Questo incremento potrebbe essere spiegato dalla ripresa globale dopo gli anni della pandemia da COVID-19. L'attenzione mediatica e il miglioramento delle condizioni socio-economiche potrebbero aver portato a un au-

mento della condivisione di esperienze positive sui social media, contribuendo a questo picco.

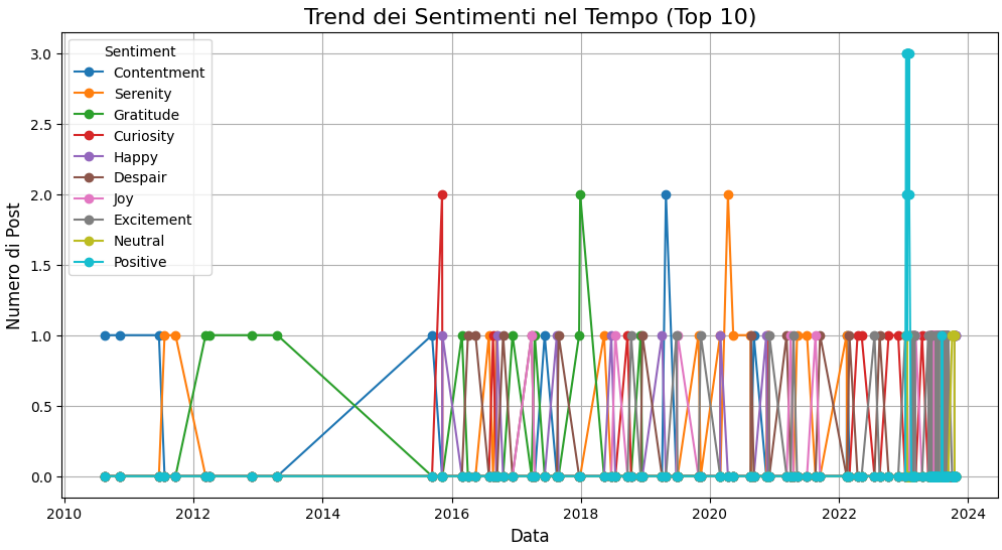

Figure 5: Trend dei top 10 sentimenti nel tempo

L'analisi esplorativa ha fornito una panoramica dettagliata sulla distribuzione dei dati all'interno del dataset, evidenziando tendenze interessanti legate alla tipologia dei sentiment, alla piattaforma social, alla provenienza geografica e alla variabilità temporale.

Questi risultati offrono una base solida per l'applicazione delle tecniche di NLP che verranno approfondite nel prossimo capitolo. Attraverso il processamento avanzato del linguaggio naturale, sarà possibile estrarre informazioni ancora più dettagliate dai contenuti postati sulle diverse piattaforme e comprendere meglio le differenze e le peculiarità di ciascuna di esse.

3 Applicazione tecniche di NLP

3.1 WordCloud

Una *WordCloud* (o "nuvola di parole") è una rappresentazione visiva di parole, dove la dimensione di ciascuna parola è proporzionale alla sua frequenza o rilevanza in un determinato insieme di testi. Le parole più grandi appaiono più frequentemente o sono più significative, mentre quelle più piccole appaiono meno o sono meno rilevanti.

L'analisi delle WordCloud si è rivelata particolarmente utile per individuare le parole più rilevanti all'interno degli hashtag, nelle interazioni totali e nelle diverse piattaforme social. Questo approccio ha permesso di evidenziare le differenze nel linguaggio e nei temi trattati sui vari social network, offrendo un confronto visivo immediato tra le piattaforme.

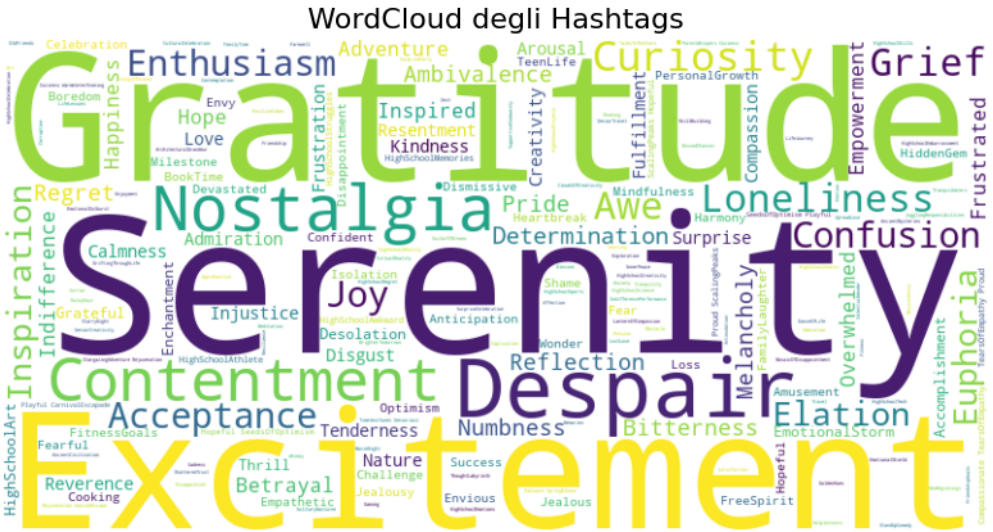

Figure 6: WordCloud degli hashtags presenti nel dataset

La WordCloud riportata in Figura 6 evidenzia le parole più frequenti negli hashtag presenti all'interno del dataset analizzato. Tra le parole più grandi e quindi maggiormente ricorrenti troviamo termini come Gratitude, Serenity, Excitement, Nostalgia e Contentment, che riflettono una prevalenza di sentimenti positivi ed emozioni legate a stati d'animo di appagamento e gioia.

Questa distribuzione suggerisce che gli utenti tendono a utilizzare hashtag che esprimono emozioni personali o stati d'animo legati a esperienze gratificanti, probabilmente per enfatizzare o condividere momenti significativi.

Tuttavia, la presenza di emozioni meno positive, come Despair, Loneliness e Grief, pur se rappresentate da parole di dimensioni più piccole, offre uno spunto interessante.

Sebbene meno frequenti, queste parole indicano che anche esperienze emotive difficili vengono condivise, seppur con minore intensità. Questo potrebbe suggerire una minore visibilità o un livello inferiore di espressione pubblica dei sentimenti negativi, rispetto a quelli positivi, sui social media.

Questa analisi evidenzia come i sentimenti tendano a dominare la comunicazione tramite hashtag.

Figure 7: WordCloud generale dei contenuti

In Figura 7, invece, è possibile osservare quelle che sono le parole maggiormente utilizzate all'interno dei contenuti generali presenti nel nostro dataset, e quindi presenti sulle tre diverse piattaforme social.

Da questa visualizzazione emerge una differenza chiara rispetto agli hashtag. Mentre gli hashtag tendono a focalizzarsi su emozioni e stati d'animo, i post stessi sembrano concentrarsi maggiormente su eventi, azioni e oggetti. Parole come new, day, life, moment e simili, che rappresentano attività quotidiane o situazioni, sono significativamente più frequenti. Questa differenza potrebbe essere attribuita al fatto che gli hashtag fungono da "tag" concisi per esprimere emozioni, mentre i post descrivono più dettagliatamente l'esperienza in corso, con un'attenzione maggiore al contesto e ai dettagli.

La combinazione di queste osservazioni suggerisce che gli utenti utilizzano gli hashtag per esprimere sinteticamente il loro stato d'animo, mentre nei post tendono ad offrire una visione più complessa e contestualizzata della loro esperienza. Questi due aspetti della comunicazione sui social media sono quindi complementari, ma mostrano come gli utenti possano scegliere diverse modalità per esprimere emozioni e condividere la propria vita online.

3.1.1 Tecnica WordCloud applicata ai singoli social

La tecnica del WordCloud è stata da noi impiegata per evidenziare le parole più significative che si utilizzano sulle tre principali piattaforme social: Facebook, Twitter ed Instagram. La scelta di approfondire l'analisi per ognuno dei social mira a studiare l'eventuale differenza nelle tipologie di messaggi che pubblicano.

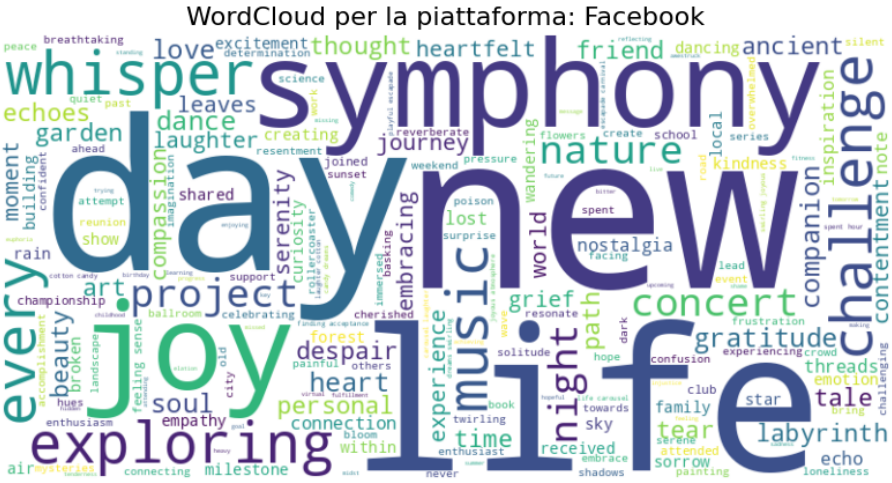

Figure 8: WordCloud dei contenuti caricati sul social Facebook

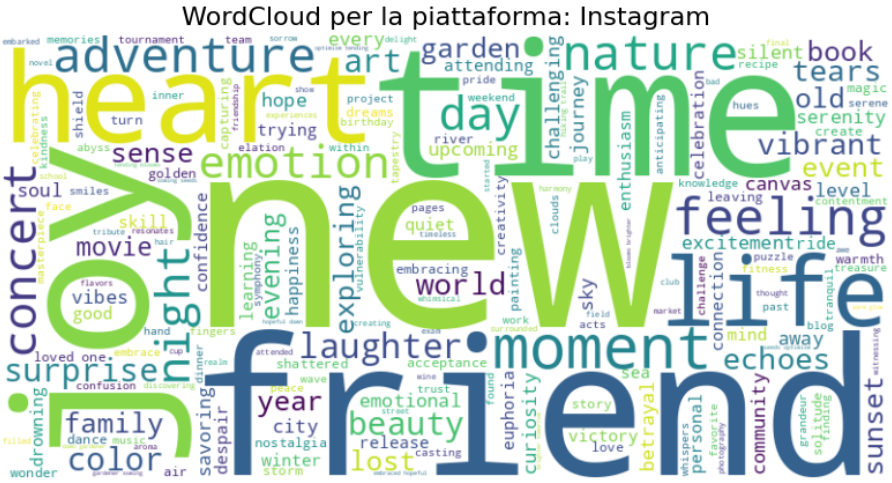

Figure 9: WordCloud dei contenuti caricati sul social Instagram

Come possiamo notare dalle Figure 8 e 9, nelle quali sono riportati i WordCloud relative ai social media Facebook ed Instagram, le parole maggiormente utilizzate in questi due casi sono pressoché simili e riconducibili al caso generale visto precedente-

mente (vedi Figura 7). Questo riflette la natura di queste piattaforme, che sono spesso utilizzate per mantenere connessioni personali, condividere momenti significativi e esprimere emozioni in modo chiaro agli amici o ai follower. Le somiglianze tra Facebook e Instagram possono essere dovute alla forte interazione visiva che entrambe le piattaforme promuovono, sia attraverso foto che attraverso post brevi, favorendo una condivisione di contenuti positivi, di esperienze e di eventi personali.

In Twitter, invece, emerge un quadro piuttosto diverso. La piattaforma è nota per la sua brevità e immediatezza, ma anche per l'enfasi sulle notizie, opinioni e discussioni su eventi globali. I WordCloud di Twitter mostrano una maggiore presenza di parole di ampio significato, come dream, journey, world, life, e termini relativi a emozioni, come feeling o lost. Questo suggerisce che gli utenti di Twitter sono più propensi a condividere riflessioni, opinioni e osservazioni su temi di rilevanza globale, che vanno ben oltre la mera espressione di stati d'animo e condivisione di vita personale. Twitter infatti è una piattaforma in cui la comunicazione è più focalizzata sulla discussione pubblica, sulla diffusione di idee e sulla riflessione su eventi esterni, rispetto a Facebook e Instagram, che si concentrano maggiormente sulla sfera privata e personale.

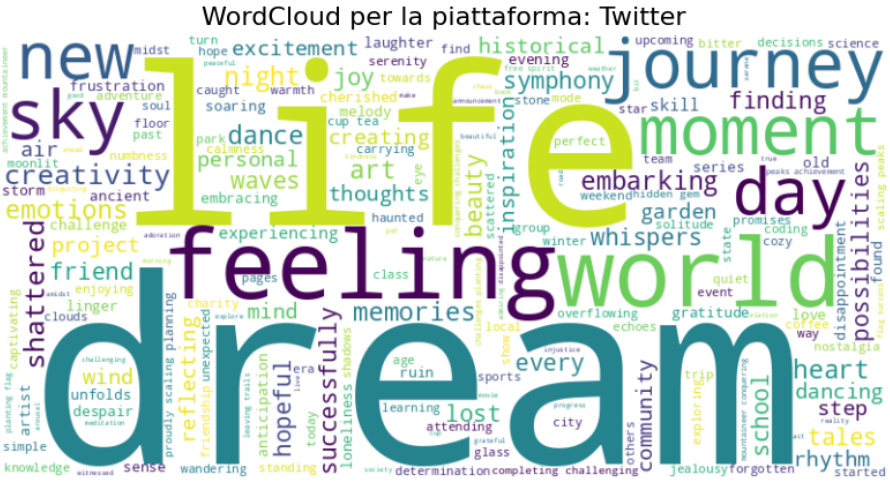

Figure 10: WordCloud dei contenuti caricati sul social Twitter

3.2 Sentiment Analysis

Dopo aver esplorato le parole più significative attraverso le WordCloud, è stato naturale proseguire con una Sentiment Analysis per approfondire ulteriormente la comprensione delle emozioni espresse dagli utenti sui social media. Le WordCloud ci forniscono una panoramica delle parole più frequentemente utilizzate, rivelando così le tematiche dominanti e gli argomenti di discussione. Tuttavia, un'analisi qualitativa delle parole da sola

non è sufficiente per capire il tono emotivo e il sentimento complessivo espresso in un post o in un hashtag.

La Sentiment Analysis, invece, ci permette di attribuire un valore numerico o qualitativo al sentimento espresso, classificandolo come positivo, negativo o neutro. Questa analisi va ad arricchire le considerazioni tratte dalle WordCloud, poiché non solo ci dice quali parole vengono utilizzate, ma anche come vengono percepite e interpretate dal pubblico. Ad esempio, se la WordCloud evidenzia la presenza di parole come "happiness" o "sadness", la Sentiment Analysis ci aiuterà a capire se il tono complessivo di questi contenuti è effettivamente positivo o negativo.

L'approccio combinato delle due tecniche ci consente di ottenere una visione più completa: le WordCloud ci mostrano cosa gli utenti condividono (quali parole e argomenti), mentre la Sentiment Analysis ci fornisce un'indicazione di come questi argomenti vengono percepiti. Insieme, queste due metodologie ci permettono di esplorare i contenuti social non solo in termini di frequenza e visibilità, ma anche di tono emotivo, contribuendo a una comprensione più profonda del comportamento e delle emozioni degli utenti sui social media.

3.2.1 Cos'è la sentiment analysis

La *Sentiment Analysis* (analisi del sentiment) è un campo dell'intelligenza artificiale e della linguistica computazionale che si occupa di identificare e classificare le emozioni e opinioni espresse nei testi. L'obiettivo principale è determinare se un testo esprime un sentimento positivo, negativo o neutro.

Questa tecnica viene comunemente utilizzata per analizzare il tono e l'umore dei contenuti, come recensioni online, commenti sui social media, articoli di giornale, e altro ancora.

La Sentiment Analysis si basa su diversi approcci, tra cui:

1. **Classificazione di polarità**: Determina se un testo è positivo, negativo o neutro. Ad esempio, una recensione di un prodotto potrebbe essere etichettata come "positiva" se l'utente esprime soddisfazione, "negativa" se esprime insoddisfazione, o "neutra" se è più informativa senza un forte giudizio emotivo.

2. **Analisi di intensità o emozioni**: In alcuni casi, oltre alla polarità, la sentiment analysis può cercare di identificare l'intensità del sentimento o le emozioni specifiche, come felicità, rabbia, tristezza, sorpresa, ecc.

3. **Soggettività e oggettività**: Un'altra dimensione della sentiment analysis è distinguere tra affermazioni soggettive (opinioni, sentimenti) e oggettive (fatti, descrizioni neutrali).

3.2.2 Approccio utilizzato

La nostra analisi è stata portata avanti utilizzando *TextBlob*, una libreria Python utilizzata per l'analisi del testo, con particolare attenzione all'elaborazione del linguaggio naturale (NLP). Fornisce una semplice API per compiti comuni di NLP, come l'analisi del sentiment, la traduzione, l'analisi grammaticale, l'estrazione di frasi e parole chiave, e altre operazioni.

Per quanto riguarda i risultati dell'analisi, la distribuzione dei sentimenti, riportata in Figura 11, mostra una netta predominanza di sentimenti di tipo "Neutral". È interessante notare che solo un numero limitato di post (poco più di 100) risulta essere associato a sentimenti negativi, indicando che la maggior parte dei contenuti analizzati tende a mantenere un tono relativamente neutro o positivo.

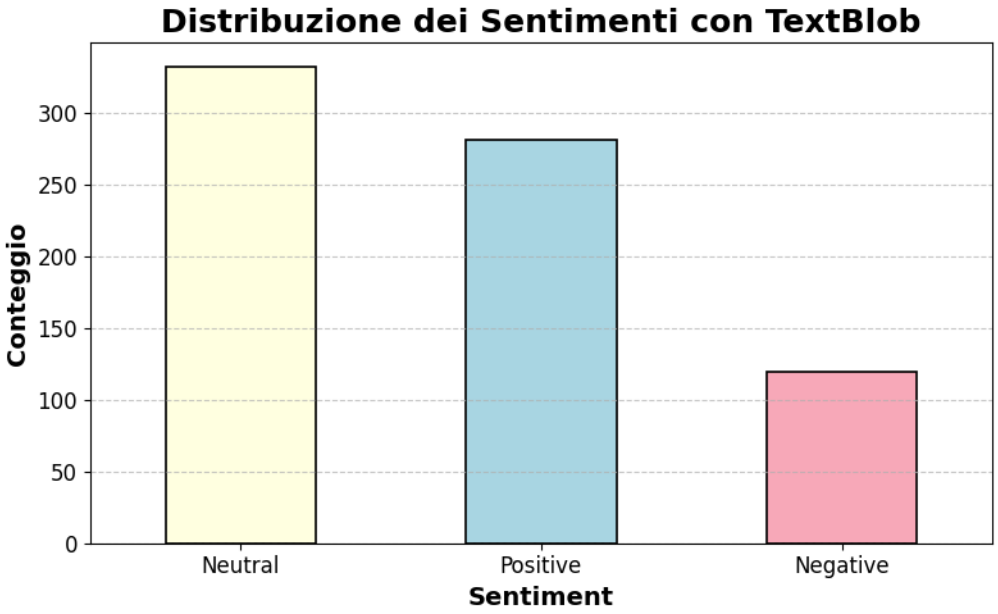

Figure 11: Distribuzione dei sentimenti ottenuti con textBlob

Un altro aspetto importante che si tiene conto quando si intraprende uno studio di NLP sulla Sentiment Analysis è quello riguardante la soggettività. TextBlob fornisce un'analisi della soggettività esprimendo un numero all'interno dell'intervallo compreso tra zero e uno; nello specifico:

- **0**: indica che il testo è oggettivo (cioè non contiene opinioni o emozioni personali, ma solo fatti).
- **1**: indica che il testo è completamente soggettivo (cioè contiene opinioni, emozioni o valutazioni personali).

In Figura 12 abbiamo riportato la distribuzione della soggettività riguardante i contenuti del nostro dataset. In particolare, la stragrande maggioranza dei contenuti proposti sono oggettivi, non riguardano cioè informazioni personali o sentimenti, mentre in numero inferiori sono quelli soggettivi.

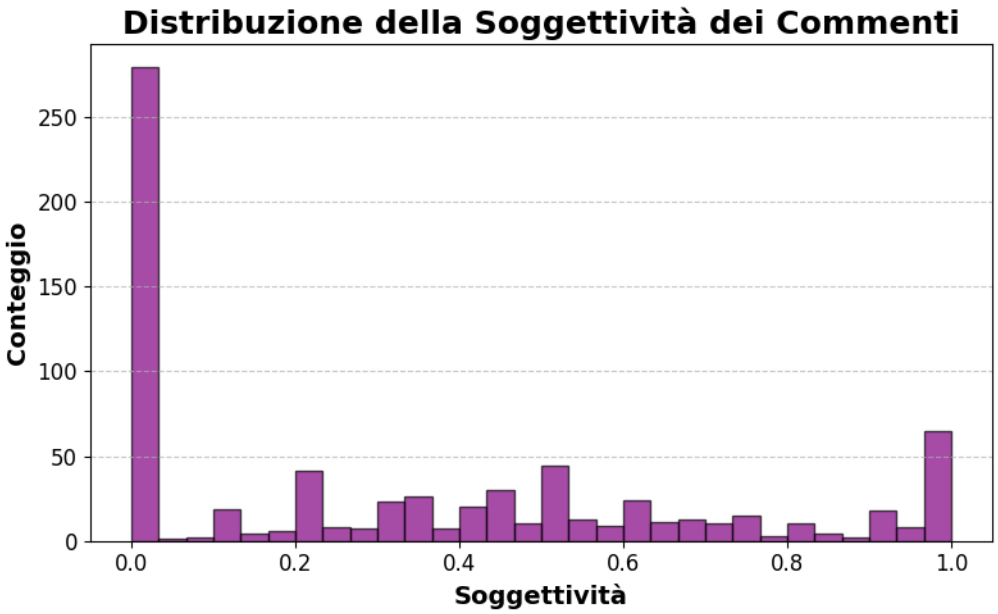

Figure 12: Distribuzione della soggettività

Altra analisi importante è quella riguardante il numero di contenuti positivi, negativi o neutri per le diverse piattaforme social e per i diversi paesi di origine.

I risultati di queste ultime due analisi sono riportati nelle Figure 13 e 14. Possiamo notare, infatti, come i contenuti "neutri" siano pressoché costanti ed omogenei sulle diverse piattaforme social, mentre diversamente accade per i contenuti negativi e positivi. Quest'ultimi, infatti, sono maggiormente presenti sul social media instagram. Questa caratteristica potrebbe essere influenzata dalla natura intrinsecamente visuale della piattaforma. Instagram è spesso utilizzato per condividere momenti felici e immagini che trasmettono emozioni positive, suggerendo che la piattaforma favorisca contenuti più ottimisti e "vissuti", rispetto a Twitter, che ha un tono più riflessivo.

In merito al paese di origine, invece, i risultati sono stati riportati in Figura 14 e mostrano alcuni fattori interessanti. Il Canada, ad esempio, "a parità" di numero contenuti con il Regno Unito, presenta un numero di post negativi più alto, ed anche il suo rapporto è decisamente più alto. Infatti, dato il numero totale di contenuti provenienti da quel paese, la percentuale negativa è sostanzialmente più alta e impatta maggiormente rispetto ad altri paesi, come ad esempio gli Stati uniti in cui la stragrande maggioranza sono contenuti positivi.

Le differenze di sentimenti nei paesi suggeriscono come fattori culturali e sociali possano giocare un ruolo significativo nel modo in cui gli utenti esprimono emozioni sui social. La predominanza di contenuti positivi negli Stati Uniti, ad esempio, potrebbe riflettere la cultura prevalentemente ottimista e motivazionale che caratterizza molte interazioni online.

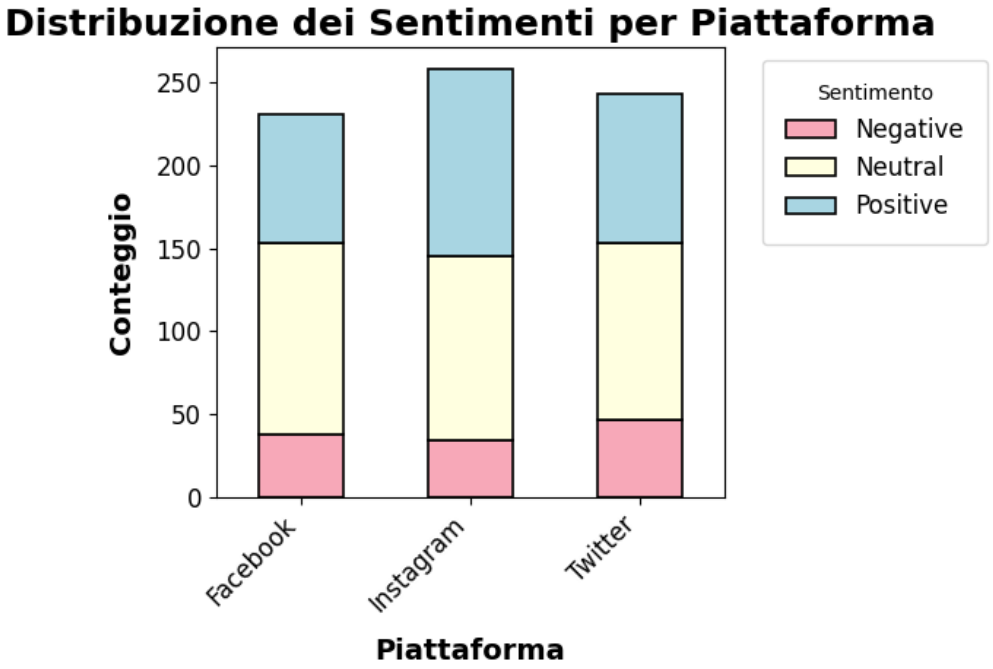

Figure 13: Distribuzione dei sentimenti per piattaforma social

Ancora una volta, dunque, capiamo la potenza del NLP in campo sociale e, soprattutto legato ai media, dove tutt'oggi si ha la stragrande maggioranza di condivisione di informazione rispetto ad altre periferiche.

3.3 Text classification con FastText

Dopo aver esaminato le parole più significative attraverso le WordCloud, analizzato i sentimenti e valutato la soggettività dei contenuti, la fase successiva del nostro studio si è concentrata sulla Text Classification, utilizzando la libreria FastText. Mentre le analisi precedenti ci hanno fornito una panoramica qualitativa sui contenuti, identificando i temi più ricorrenti, le emozioni prevalenti e il grado di soggettività, la Text Classification ci permette di andare oltre, attribuendo categorie specifiche ai contenuti in base a criteri predefiniti.

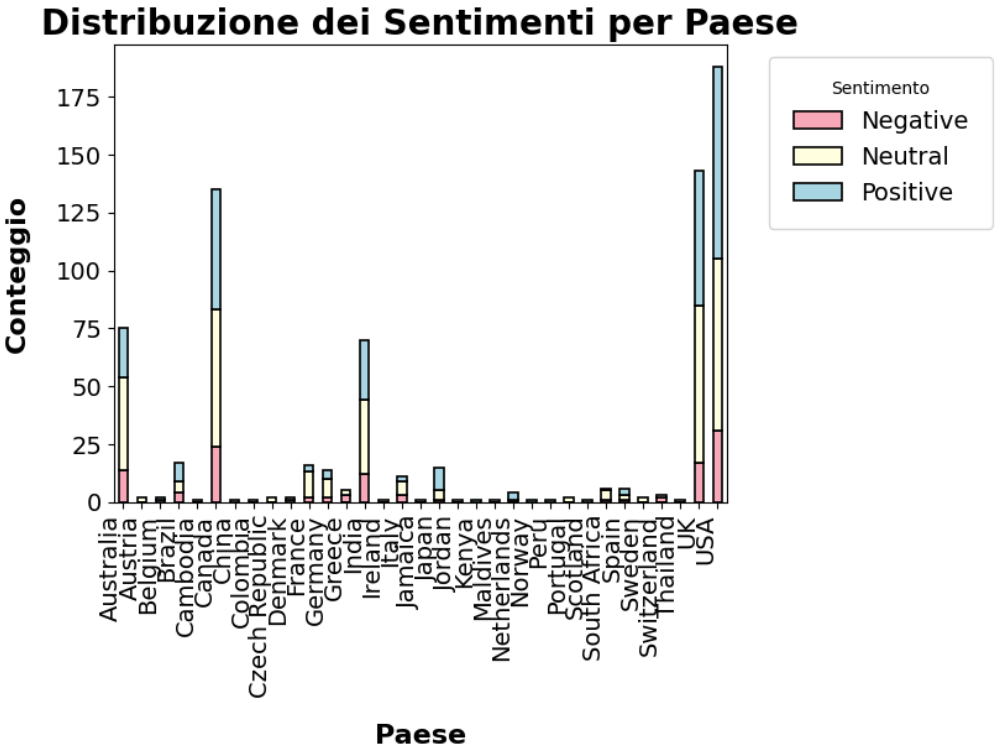

Figure 14: Distribuzione dei sentimenti per paese di origine del contenuto

3.3.1 Cos'è la text classification

La *text classification* (classificazione del testo) è un compito di apprendimento automatico in cui si assegna una categoria o etichetta a un dato testo in base al suo contenuto.

I principali tipi di text classification sono:

- **Classificazione binaria**: Il testo viene assegnato a una di due classi (ad esempio, "positivo" o "negativo").
- **Classificazione multi-classe**: Il testo viene assegnato a una delle tante possibili classi.
- **Classificazione multilabel**: Un testo può essere associato a più di una classe contemporaneamente.

3.3.2 Approccio utilizzato

Per raggiungere il nostro obiettivo abbiamo utilizzato *FastText*. FastText è una libreria sviluppata da Facebook AI Research (FAIR) che facilita l'addestramento di modelli di

classificazione del testo (e non solo). È progettata per essere veloce, efficiente e facile da usare, ed è particolarmente adatta per operare con grandi dataset.

Di seguito elenchiamo quelli che sono i principali passaggi per utilizzare questo tool.

1. **Prepara i dati**: I dati devono essere nel formato giusto per essere utilizzati con FastText. Per la classificazione del testo, ogni esempio nel dataset dovrebbe essere rappresentato come una riga nel file di testo, dove il testo è preceduto da un'etichetta.

2. **Addestramento del modello**: È possibile addestrare un modello FastText specificando il percorso al file di addestramento.

3. **Valutazione del modello**: Una volta addestrato, si può valutare il modello sui dati di test utilizzando il metodo di valutazione di FastText.

Per effettuare questa classificazione abbiamo effettuato una rielaborazione del dataset, modificando le etichette in nostro possesso, come ad esempio "joy, love, ecc" nella corrispettiva etichetta *positive*, mentre sentimenti negativi come "fear" sono stati tradotti nell'etichetta generale *negative*, mentre quelli intermedi sono stati etichettati come *neutral*.

Dopo aver fatto ciò, le classi sono risultate ovviamente sbilanciate, quindi abbiamo optato per una SMOTE[1] per il bilanciamento delle classi. In particolare, siamo partiti da una situazione in cui le quantità di dati per ciascuna classe erano le seguenti:

- *Training:*
  - positive: 206
  - negative: 111
  - neutral: 268
- *Test:*
  - positive: 52
  - negative: 28
  - neutral: 67

[1]SMOTE (Synthetic Minority Over-sampling Technique) è una tecnica di bilanciamento del dataset utilizzata per affrontare il problema delle classi sbilanciate nel contesto dell'apprendimento automatico. Quando un dataset contiene un numero significativamente maggiore di esempi appartenenti a una classe (di solito la classe dominante) rispetto a un'altra (di solito la classe minoritaria), il modello tende a essere più incline a predire la classe dominante, ignorando o sovrastimando la classe minoritaria. SMOTE interviene in questo scenario creando nuove istanze sintetiche per la classe minoritaria, al fine di bilanciare le distribuzioni delle classi. Questo aiuta a migliorare le prestazioni del modello, in particolare nella classificazione delle classi minoritarie.

Dopo aver effettuato la SMOTE siamo riusciti ad avere i seguenti numeri per le tre classi a disposizione:

- *Training:*
  - positive: 268
  - negative: 268
  - neutral: 268

In questo modo, dopo aver trasformato i file in un formato comprensibile da FastText, abbiamo poi potuto procedere con la classificazione dei nostri contenuti in una delle tre possibili classi appena descritte.

Il comando utilizzato per lanciare l'addestramento è *fasttext.train_supervised*, in cui abbiamo deciso di passare i seguenti parametri:

- **input**: Specifica il file di addestramento da utilizzare. In questo caso, è stato passato `fasttext_train_file`, che rappresenta il percorso al file di dati di addestramento.
- **epoch**: Numero di epoche (iterazioni) per cui il modello viene addestrato sui dati. È stato impostato a `25`, il che significa che l'algoritmo passerà attraverso il dataset 25 volte.
- **lr** (learning rate): Il tasso di apprendimento per l'algoritmo. È stato impostato a `0.5`, indicando un tasso di apprendimento relativamente alto, che può accelerare l'apprendimento ma con il rischio di sovradattamento se troppo alto.
- **wordNgrams**: Questo parametro definisce la dimensione massima dei n-grammi da utilizzare. È stato impostato a `3`, quindi FastText considererà n-grammi fino a 3 parole consecutive (trigrammi) per catturare meglio le dipendenze contestuali tra le parole.
- **verbose**: Questo parametro controlla il livello di verbosità durante l'addestramento, cioè quante informazioni vengono mostrate durante il processo. È stato impostato a `2`, che significa un livello di verbosità medio, mostrando informazioni utili sul progresso dell'addestramento.
- **minCount**: Imposta il numero minimo di occorrenze di una parola per essere considerata durante l'addestramento. È stato impostato a `1`, il che significa che tutte le parole, anche quelle che compaiono una sola volta, saranno incluse nel modello.

Dopo aver concluso l'addestramento, che impiegherà qualche secondo vista l'efficienza di FastText e vista la poca numerosità di occorrenza all'interno del nostro dataset, abbiamo effettuato una validazione sui nostri dati di test, attraverso il comando seguente: *model.test(fasttext_test_file)*.

I parametri di precision[2] e recall[3] ottenuti in fase di test sono i seguenti:

- precision: 0.6938775510204082
- recall: 0.6938775510204082

Ovviamente sono dei parametri non ottimali, e questo può essere ricondotto al basso numero di dati a disposizione che avevamo (meno di mille). Nonostante il processo di augmentation messo in pratica con SMOTE, per effettuare un addestramento con valori in fase di test eccellenti, bisognerebbe avere molti più dati a disposizione.

[2]La precisione misura l'accuratezza delle previsioni positive fatte dal modello. In altre parole, indica quanti degli esempi che il modello ha classificato come positivi sono effettivamente positivi.

[3]Il recall misura la capacità del modello di identificare correttamente tutti i veri positivi. In altre parole, indica quanti degli esempi positivi effettivi sono stati identificati correttamente dal modello.

List of Figures