SENTAI

Progetto FIA 2024/2025

Sentiment analysis di commenti social

Giovanni Paolo Chierchia
Mat. [0512117783]

Introduzione

SENTAI

- Problema: Identificare e classificare il sentiment dei commenti online come positivo o negativo. Problema di **classificazione binario**.
- Utilità: Classificazione del sentiment viene **sfruttata dalle aziende** o da **enti governativi** per monitorare la propria reputazione e condurre analisi di mercato.
- Obiettivo: Sviluppare un modello in grado di analizzare testi in inglese brevi e generici e **determinarne il sentiment** (positivo o negativo). Verrà scelto un classificatore tra **Naive Bayes** e **Logistic Regression**.
- Metodologia: Modello **CRISP-DM** per gestire il ciclo di vita dell'analisi dei dati.

Specifica P.E.A.S

SENTAI

- Performance: Metriche **accuracy, precision** e **recall** e **analisi della matrice di confusione**. Tempo di risposta massimo: **5 secondi** dall'interfaccia grafica.

- Environment: Frasi o post dei social media.

- Actuators: Assegnazione di un’etichetta di sentiment (positivo/negativo).

- Sensors: Percezione dell’ambiente tramite il testo in input

Business Understanding

SENTAI

- Soglia minima delle metriche di valutazione: Accuracy, precision e recall non inferiori a 0.65.
- Dataset utilizzato: Dataset contenente 1.600.000 tweet: **Sentiment140**
- Problematiche:
  - Testi social spesso **molto rumorosi** e con presenza di sarcasmo e ironia.
  - Il datataset Sentiment140 risale al 2009, problemi con evoluzione del linguaggio social.
  - Lunghezza limitata dei tweet su cui è stato addestrato (massimo 140 caratteri).
  - Alcuni errori di scraping.

Data Understanding

SENTAI

- Dataset originale: 1.600.000 tweet con distribuzione bilanciata (50% per ciascuna classe) e assenza di valori nulli o stringhe vuote.

- Caratteristiche chiave:
  - Target: Etichetta di sentiment del tweet (0 = negativo, 4 = positivo).
  - Text: Testo del tweet.
- Feature secondarie: id, date, flag e user.

- Riduzione del dataset a 100.000 tweet mantenendo la distribuzione originale tramite campionamento stratificato

Analisi del dataset ridotto

- Nel dataset sono presenti 296 testi duplicati.

- Analisi della lunghezza dei tweet:

  - I tweet con lunghezza **superiore a 140 caratteri** sono il risultato di **errori di scaping**:

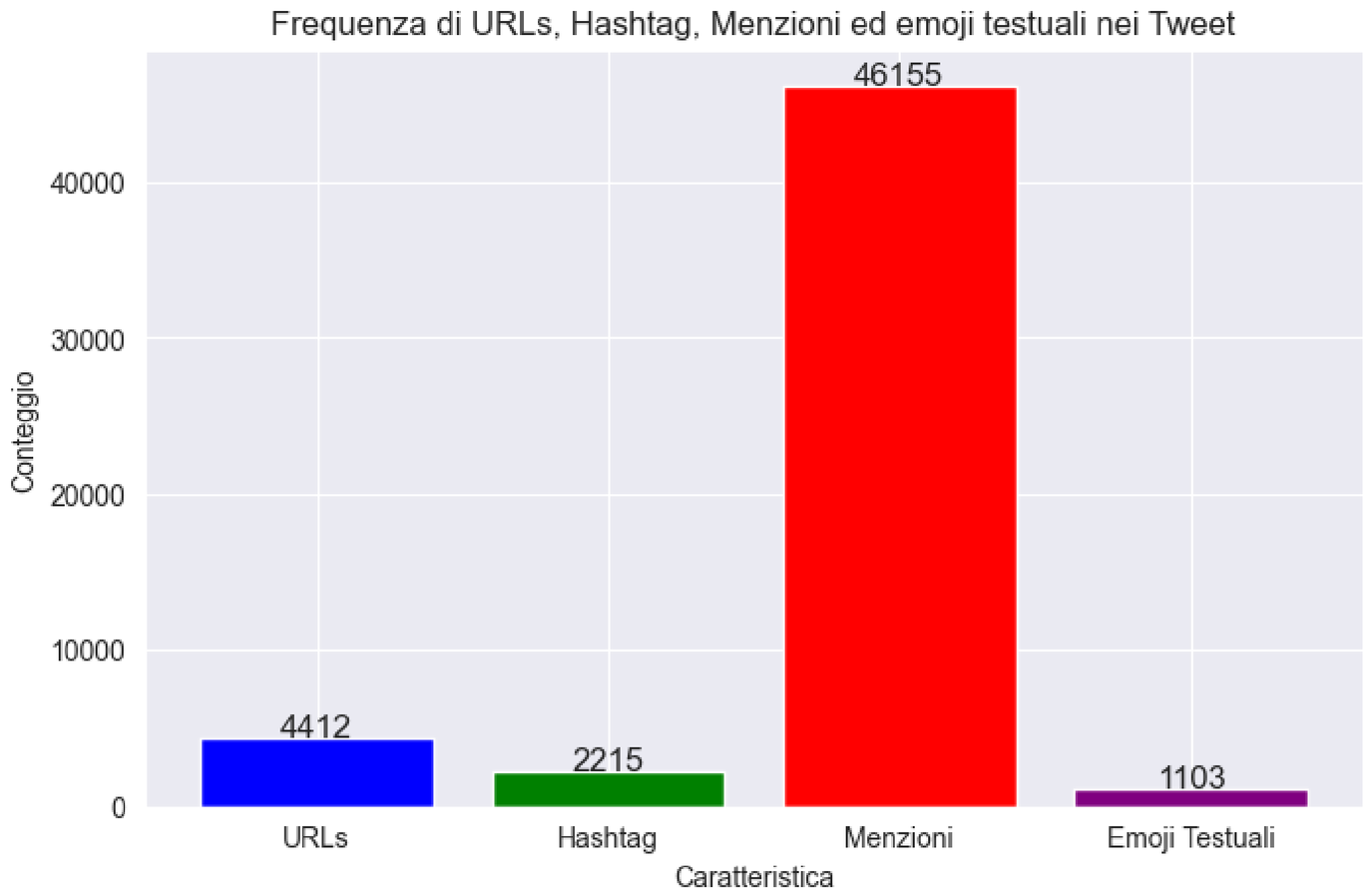
```
Things â00 Ñ?Ñ0Ð¾ Ð0ÐµÑ0Ñ0!  Ð¡Ñ0Ð¾ÐµÐ» Ñ0Ð°Ð·Ð³Ñ0ÐµÑ?Ñ0Ð¸ Ñ0Ð°Ð°Ð¾Ð¹ Ð·Ð°Ð²Ð°Ð» Ð² ⸮
⸮Ð¿Ð¾Ð´Ð³Ð¾Ñ0Ð¾Ð²Ð°Ðµ Ð¿Ñ0ÐµÐ´Ñ?Ñ0Ð¾Ñ?Ñ0Ð¸Ñ0 Ð¿Ñ0Ð¾ÐµÐ°Ñ0Ð¾Ð², Ñ0Ñ0Ð¾ Ñ?ÐµÑ0Ð´Ñ0Ðµ ⸮
⸮Ñ0Ð°Ð´Ñ0ÐµÑ0Ñ?Ñ?! Ð¢ÐµÐ¿ÐµÑ0Ñ0 Ð² Ð´Ñ0Ñ0 Ð¸ Ñ0Ñ0Ð¸Ñ0Ñ0Ñ?Ñ?, Ñ0Ñ0Ð¸Ñ0Ñ0Ñ?Ñ?...
```

  - Alcuni tweet con **meno di 10 caratteri** non apportano informazioni rilevanti.

Analisi del dataset ridotto(2)

SENTAI

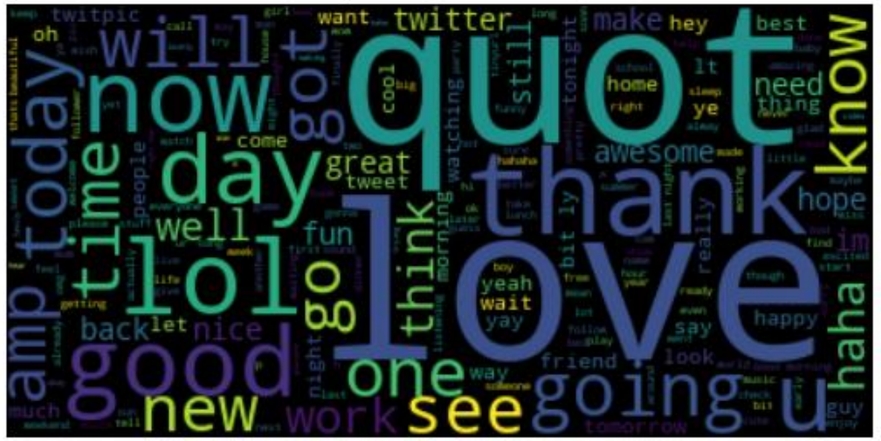
Numero elevato di tweet contenenti **URL, menzioni, hashtag** e **emoji testuali:**



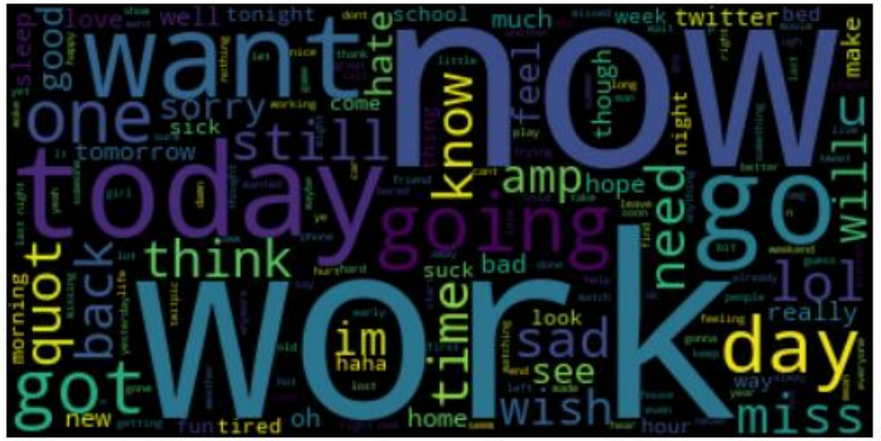
Analisi wordcloud

Da entrombi i wordcloud emergono **parole dominanti** che riflettono chiaramente un contesto emotivo.
È evidente anche la **presenza di artefatti testuali** non processati correttamente come «amp» e «quot»

Parole Frequenti - Sentiment Positivo



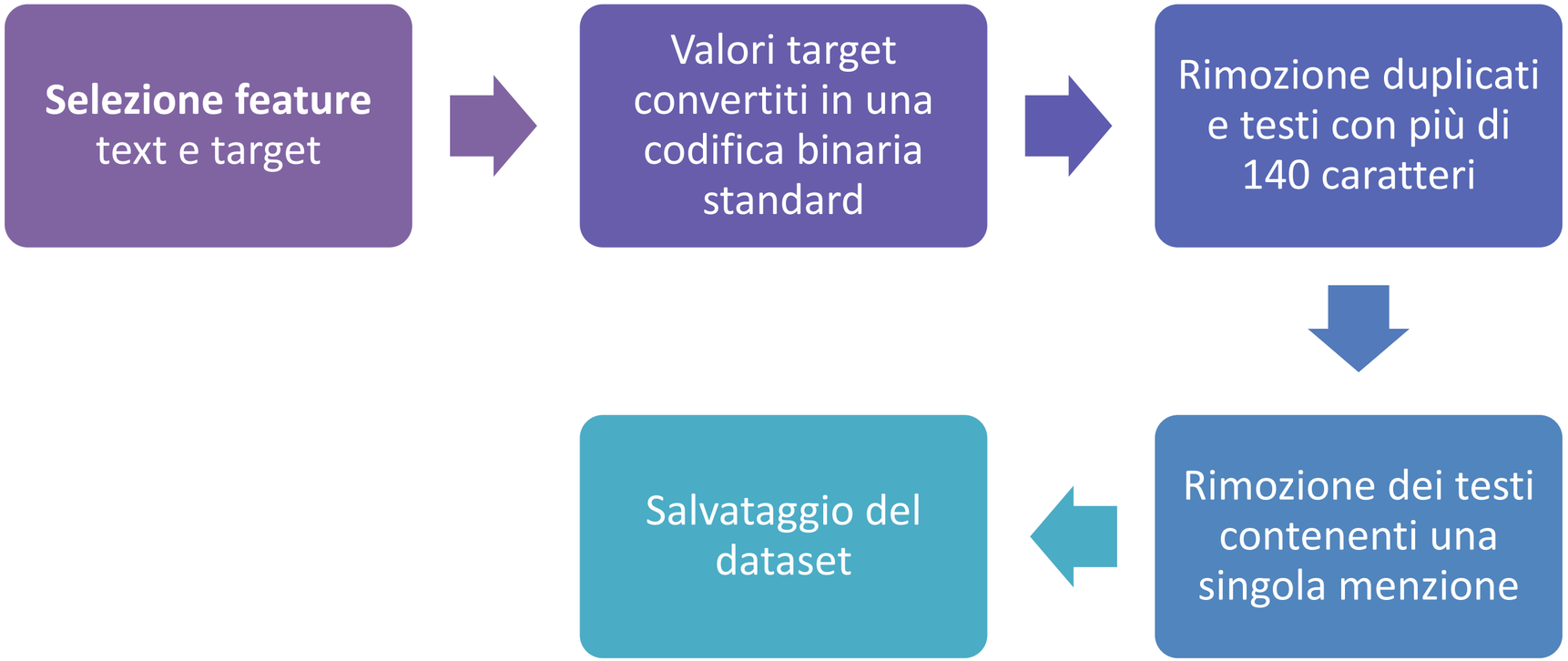
Parole Frequenti - Sentiment Negativo



Data Preparation: Pulizia preliminare

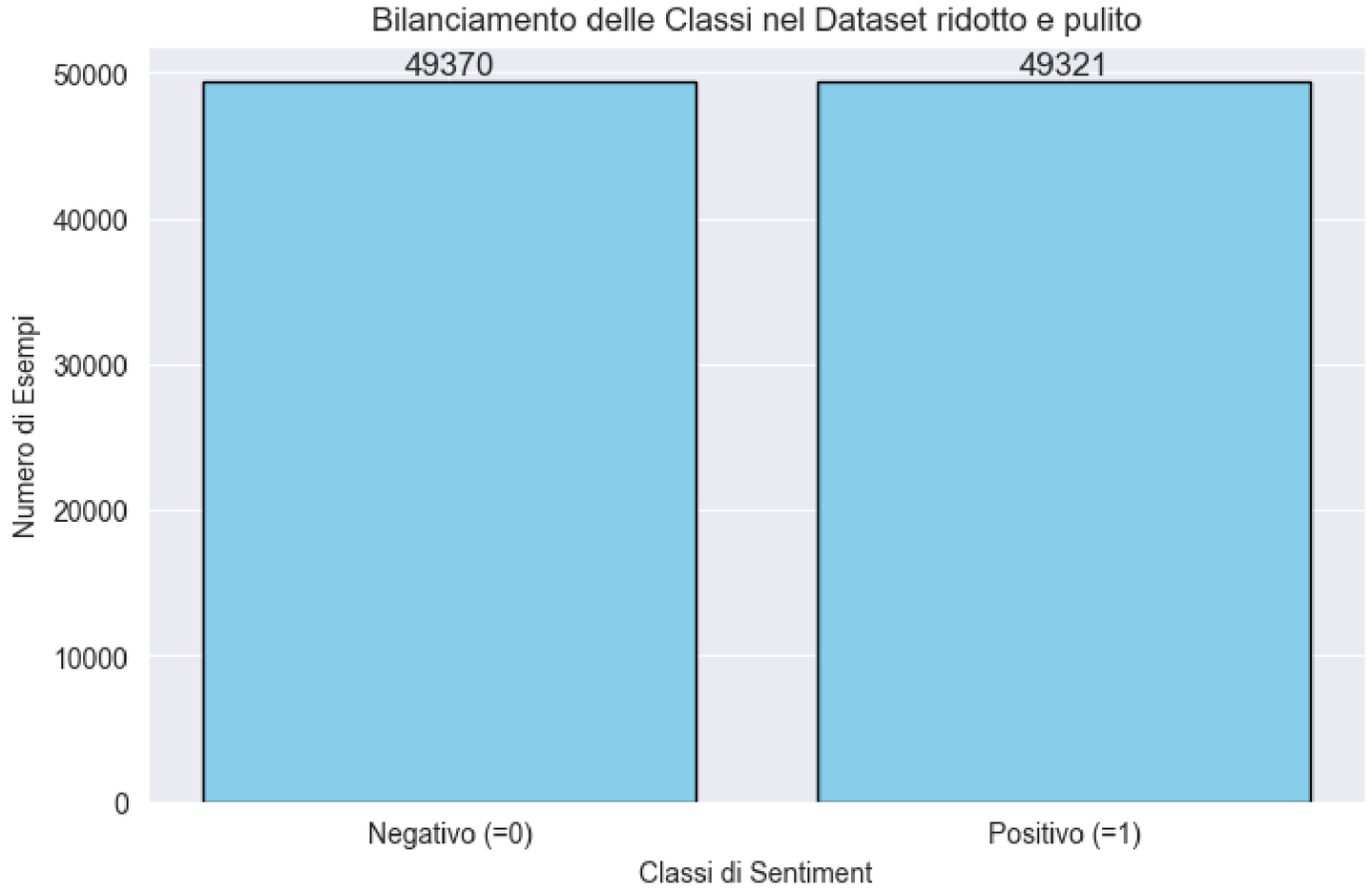
SENTAI

Pulizia preliminare per garantire un dataset più pulito e per prepararlo alla fase di modeling e al NLP.



Data Preparation: pulizia preliminare(2)

Dopo la pulizia preliminare il dataset **rimane comunque bilanciato**:



Data Preparation: NLP

Sono stati individuati vari passaggi per **rimuovere il rumore dal testo** ed **estrarre le informazioni significative** utili ai modelli di classificazione, i passaggi verranno anche inclusi all'interno della pipeline in fase di data modeling:

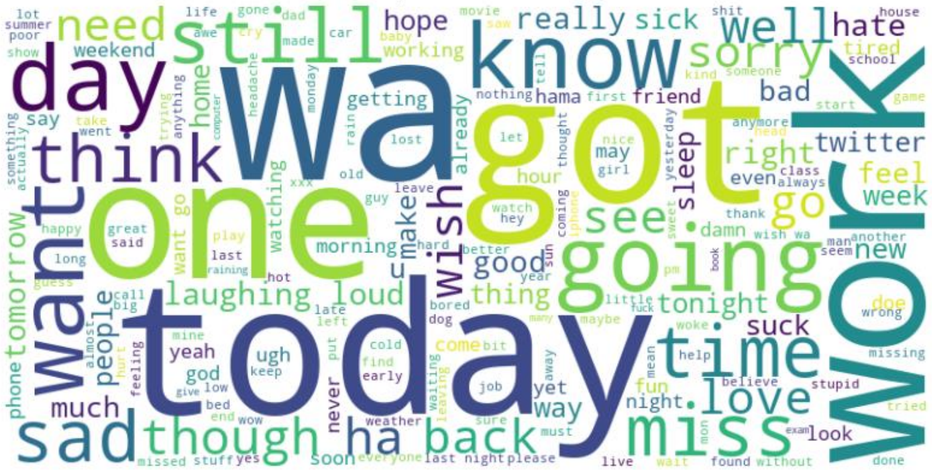
- Conversione del testo in minuscolo
- Rimozione degli URL e delle menzioni
- Rimozione/Sostituzione delle entità HTML
  - *& " < e >* vengono rimosse. *'* invece viene convertita in un apostrofo
- Sostituzione delle emoji testuali
- Rimozione degli hashtag
- Espansione di abbreviazioni e contrazioni
- Rimozione punteggiatura e caratteri speciali
- Correzione Ortografica
- Normalizzazione degli spazi
- Tokenizzazione, Lemmatizzazione e rimozione delle Stopword
- Ricostruzione del testo

Data Preparation: NLP(2)

Dopo aver fatto la pulizia, molte parole che creavano rumore non sono più presenti.

È emersa la **parola «wa»** che compare sia nei tweet positivi che negativi. Potrebbe essere un'abbreviazione ma in generale non aggiunge informazioni, per questo motivo è stata rimossa.

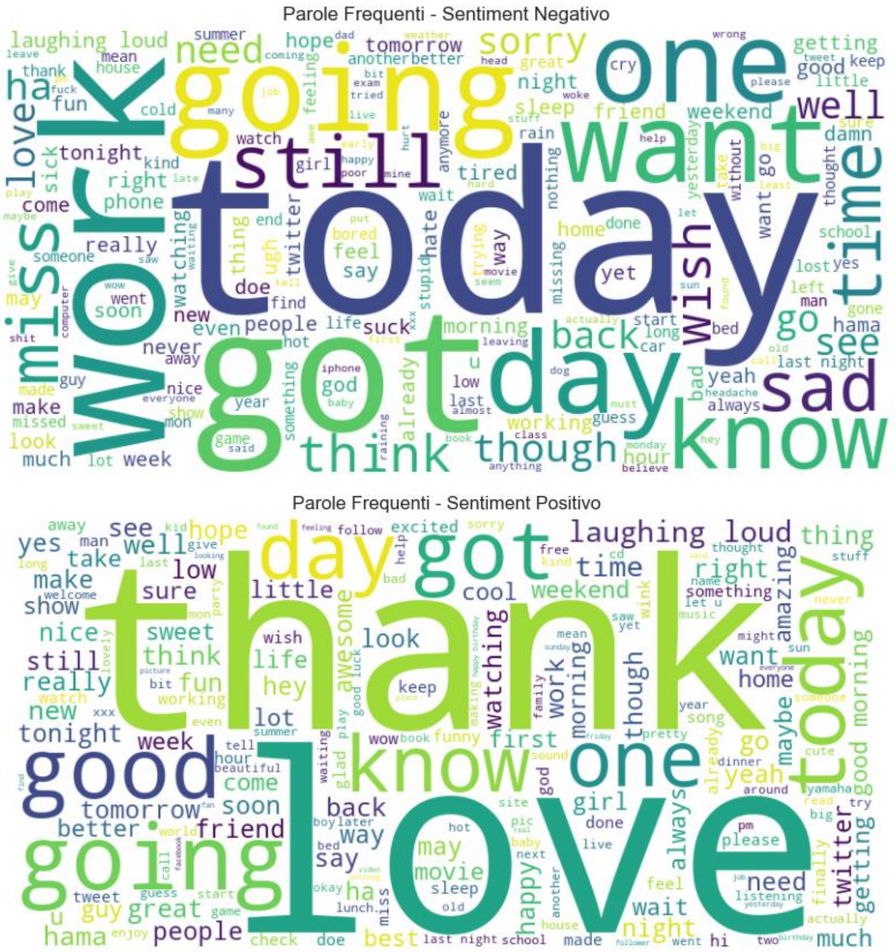
Parole Frequenti - Sentiment Negativo



Parole Frequenti - Sentiment Positivo

Data Preparation: NLP(3)

La parola «wa» è stata rimossa e anche questo passaggio è stato aggiunto alle operazioni di NLP.



Data Preparation: Feature Extraction

SENTAI

Sono state scelte due tecniche di rappresentazione testuale: Bag of Words e TF-IDF. Queste due tecniche permettono di trasformare il testo in vettori numerici:

Bag of Words: Ogni frase viene convertita in un vettore numerico in cui ogni posizione rappresenta il numero di volte che una determinata parola compare nella frase.

TF-IDF: Anch'esso converte in un vettore allo stesso modo di BoW ma i conteggi vengono **trasformati in pesi** mediante il calcolo del TF-IDF. In questo modo il peso di ogni parola aumenta se il termine è frequente in una specifica frase, ma diminuisce se è comune in tutto il dataset.

Non è detto che TF-IDF sia generalmente migliore: Alcune parole molto frequenti possono comunque essere utili per capire il sentiment, e in quindi il semplice conteggio BoW in alcuni casi può rivelarsi più efficace.

Data Modeling

SENTAI

Saranno addestrate 4 configurazioni diverse:

- Naive Bayes con BoW
- Naive Bayes con TF-IDF
- Logistic Regression con BoW
- Logistic Regression con TF-IDF

Valutazione:

- **Stratified k-Fold** (k=10)
- Calcolo di **accuracy**, **precision** e **recall**
- Selezione della configurazione finale in base alla **migliore media delle metriche nei fold**, confermata da **performance ottimali su un dataset sintetico** di 1200 tweet

Data Modeling: Pipeline

Pipeline comune alle 4 configurazioni:

Durante la fase di fit (**addestramento**):

1. Il testo viene preprocessato.
2. Viene costruita la rappresentazione numerica e viene creato il **dizionario**.
3. Il classificatore viene addestrato.

Durante la fase di predict (**predizione**):

1. Il testo viene preprocessato.
2. I testi puliti vengono trasformati utilizzando il dizionario già appreso durante il fit.
3. Il classificatore **utilizza ciò che ha imparato durante la fase di fit** per fare la predizione.

Data Modeling: Confronto performance

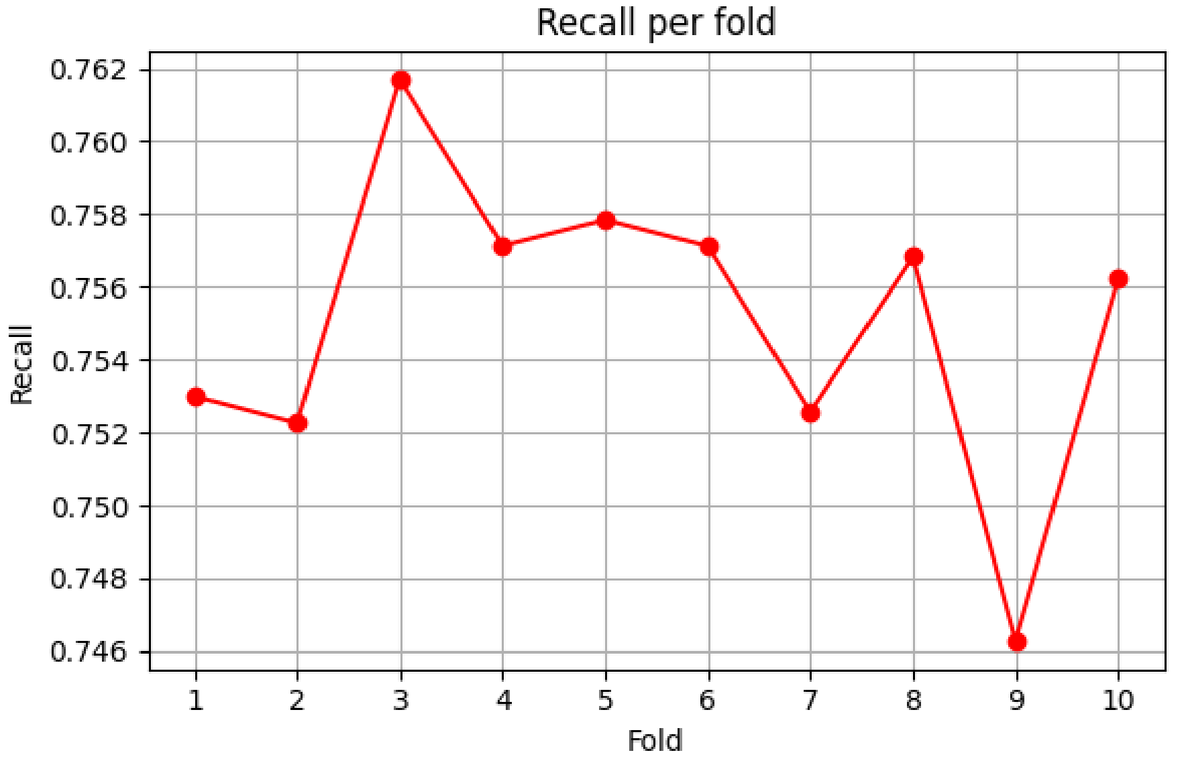
SENTAI

Performance ottenute per le 4 configurazioni:

1. **Naive Bayes con Bow**:
   - Sui fold: Ac(0.748) Pr(0.748) Rec(0.748)
   - Su dataset sintetico: Ac(0.938) Pr(0.941) Rec(0.938)
2. **Naive Bayes con TF-IDF**:
   - Sui fold: Ac(0.745) Pr(0.745) Rec(0.745)
   - Su dataset sintetico: Ac(0.941) Pr(0.943) Rec(0.941)
3. **Logistic Regression con Bow**:
   - Sui fold: Ac(0.751) Pr(0.752) Rec(0.751)
   - Su dataset sintetico: Ac(0.946) Pr(0.947) Rec(0.946)
4. **Logistic Regression con TF-IDF**:
   - Sui fold: Ac(0.755) Pr(0.755) Rec(0.755)
   - Su dataset sintetico: Ac(0.953) Pr(0.954) Rec(0.953)

Data Modeling: Scelta

È stata quindi scelta la configurazione **Logistic Regression con TF-IDF** in quanto ha le migliori prestazioni:



TN	

FP
561

Evaluation

Il modello Logistic Regression con TF-IDF viene ulteriormente testato per verificare che i valori minimi definiti negli obiettivi di business siano rispettati. Il test viene eseguito sulla **parte rimanente del dataset Sentiment140**, non utilizzata per il training. Questi sono i risultati:

Accuracy	Precision	Recall
0.756	0.756	0.756

TN	FP	FN	TP
548255	194324	165491	575753

Evaluation(2)

SENTAI

- I risultati ottenuti sono **superiori agli obiettivi** fissati inizialmente.

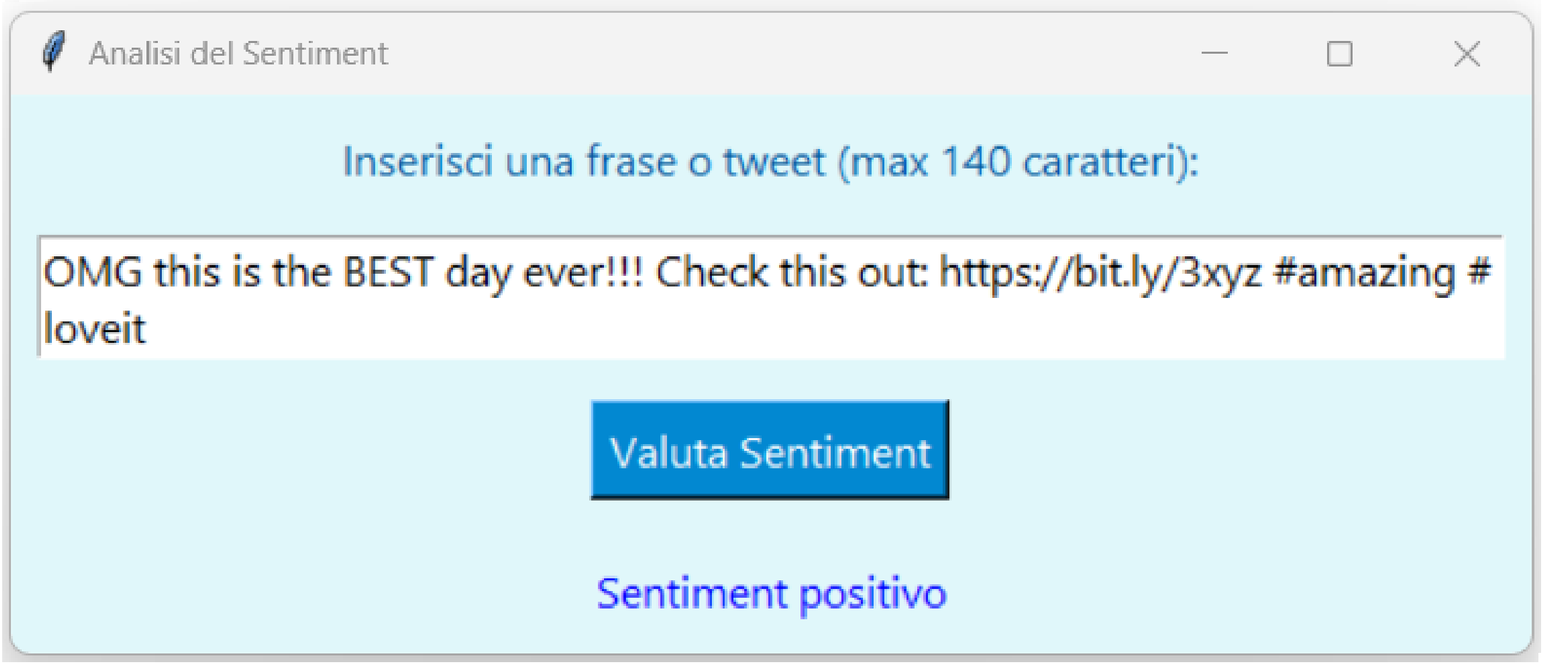
- Il modello presenta **basso** overfitting:
  - Performance costante su dataset diversi.
  - Minima varianza tra i fold.
  - Aumento significativo delle prestazioni su dataset sintetico.

- Sono però presenti **alcuni segni di** underfitting:
  - Difficoltà a cogliere la complessità semantica
  - Errori nel riconoscimento di sfumature linguistiche

Deployment

Per il deployment è stata realizzata una **semplice interfaccia con tkinter** che permette all'utente di inserire una frase di massimo 140 caratteri.

Tramite un pulsante viene fatta la classificazione. In caso di un basso livello di confidenza viene stampato un messaggio apposito.



Conclusione

SENTAI

Possibili miglioramenti:

- Configurazione più approfondita dei parametri del modello
- Utilizzo di metodi più avanzati
- Aggiunta di un ulteriore classe per il sentiment neutro.
- Ensemble tra Naive Bayes e Logistic Regression.

Il progetto mi ha permesso di scoprire numerosi aspetti che non erano stati approfonditi durante le lezioni, offrendo l'opportunità di mettere in pratica le conoscenze teoriche in un contesto applicativo.

Ho sperimentato l'ottimizzazione dei modelli e la gestione dei dati, arricchendo le mie competenze.