Progetto d'esame

Text Mining

Alessandro Carmellini

INTRODUZIONE

Al giorno d'oggi, la quantità di dati che circola online è in costante crescita e, di conseguenza, anche la loro analisi e gestione diventano sempre più complesse. Se consideriamo che la maggior parte di questi dati è di carattere testuale, la situazione si complica ulteriormente, poiché tali dati possono essere non strutturati, ad esempio, nel linguaggio naturale. È quindi opportuno adottare tecniche specifiche per estrarre al meglio le informazioni dai dati testuali. Esistono diversi contesti in cui la capacità di manipolare i testi può risultare fondamentale. Uno di questi è la sentiment analysis, che permette, a partire da un corpus di documenti, di determinare se l'opinione espressa è positiva o negativa.

In questo progetto, ci proponiamo di analizzare il sentiment delle recensioni della serie televisiva Dr. House. Dr. House è una popolare serie statunitense prodotta dal 2004 al 2012, incentrata sul personaggio del dottor Gregory House, un medico poco convenzionale ma di grande talento, a capo di una squadra di medicina diagnostica presso l'ospedale universitario Princeton-Plainsboro Teaching Hospital, nel New Jersey. La serie è stata acclamata dalla critica e ha registrato alti livelli di ascolto televisivo. Distribuita in 66 paesi, Dr. House è stato il programma televisivo più seguito al mondo nel 2008. La serie ha ricevuto diversi premi, tra cui un Peabody Award, due Golden Globe e tre Emmy Award.

Le recensioni analizzate in questo studio sono state prese da IMDb (Internet Movie Database), un sito web di proprietà di Amazon che cataloga e archivia film, attori, registi, personale di produzione e programmi televisivi. IMDb fornisce un'ampia gamma di informazioni su ciascuna opera, comprese le informazioni essenziali su cast, registi, trame e recensioni.

1.DATASET

Le informazioni necessarie sono state estratte tramite web scraping usando librerie come Selenium e BeautifulSoup per poi salvare il tutto in un file CSV tramite la libreria omonima.

Il dataset è costituito da diverse features:

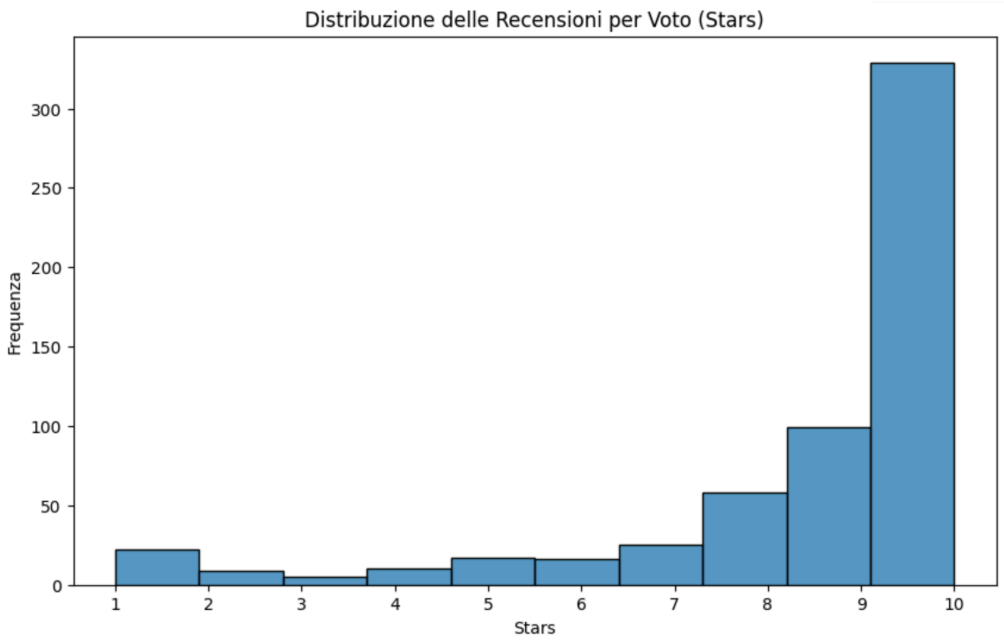
1. User: il nome dell'utente che ha scritto la recensione.
2. Stars: un punteggio facoltativo che va da 1 a 10, assegnato dagli utenti insieme alla recensione testuale.
3. Date: la data in cui la recensione è stata caricata.
4. Text: il testo vero e proprio delle recensioni.

Queste features si distribuiscono su 700 recensioni, fornendo una base per l'analisi del sentiment.

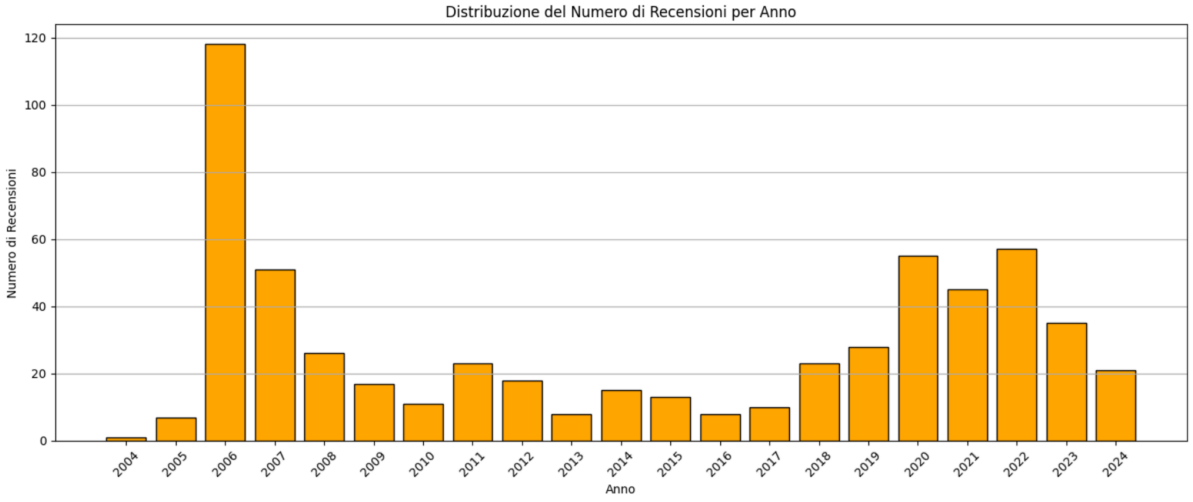
Al fine di realizzare questo obiettivo, è necessario adattare il dataset per l'applicazione di tecniche di classificazione supervisionata. Fortunatamente, la variabile "Stars" rappresenta un utile indicatore per questo scopo.
Dopo una serie di esperimenti volti a determinare la soglia più appropriata per l'etichettatura del sentiment delle recensioni, è emerso che l'approccio più adatto per i nostri scopi è l'etichettatura binaria. In particolare, in un primo momento si era tentato un'etichettatura multiclasse, associando valori da 1 a 3 a recensioni negative, da 4 a 6 a recensioni neutre e da 7 a 10 a recensioni positive. Tuttavia, si è constatato che più del 90% delle recensioni erano state classificate come positive, mentre la restante percentuale veniva distribuita tra le classi negative e neutre. Questo ha portato alla luce il problema di un dataset fortemente sbilanciato. Al fine di affrontare questo problema, si è deciso di aggregare le due classi minoritarie in un'unica categoria denominata "negativa". Sebbene ciò non risolva completamente il problema dello sbilanciamento del dataset, ha contribuito a mitigare le difficoltà durante la fase di addestramento dei modelli.

2.STATISTICA DESCRITTIVA

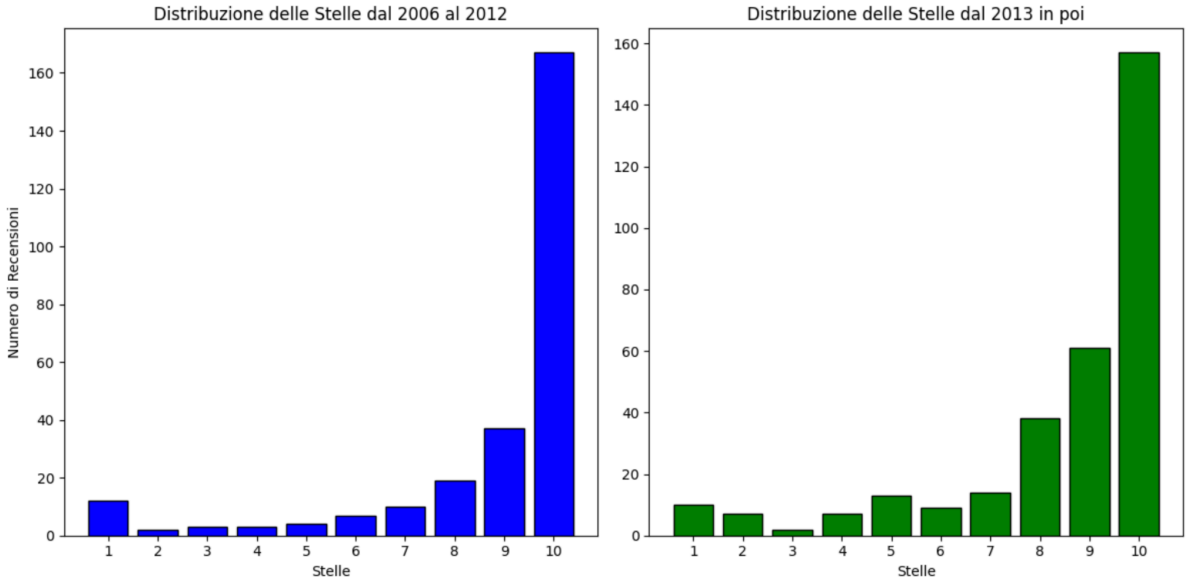
Nell'ambito dell'analisi del dataset, una delle prime osservazioni fondamentali riguarda la variabile "Stars", che rappresenta un punteggio numerico assegnato alle recensioni, variabile chiave per comprendere il sentiment complessivo degli utenti. L'analisi della distribuzione di questa variabile lungo l'intero dataset rivela una predominanza di valutazioni elevate, con la maggior parte delle recensioni assegnate a un punteggio di 10 stelle, seguite da 9 e 8 stelle. Tuttavia, dal settimo punteggio in poi, si osservano poche occorrenze, suggerendo una prevalenza di sentiment molto positivi.



Un'ulteriore analisi temporale delle recensioni rivela due distribuzioni ben definite, coincidenti con le date antecedenti e successive alla conclusione della serie televisiva, rispettivamente dal 2004 al 2012 e dal 2013 in poi.

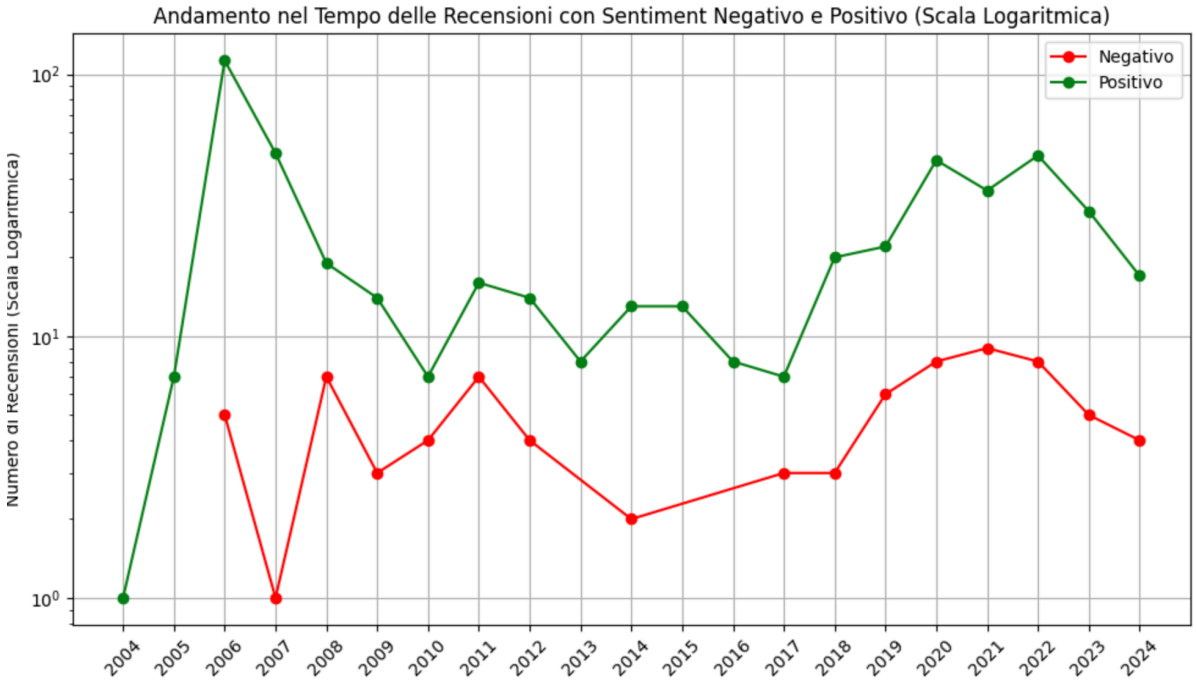


Questa divisione temporale ci porta a indagare sulle variazioni nei punteggi assegnati nel corso del tempo, al fine di valutare eventuali cambiamenti nell'apprezzamento della serie. L'analisi rivela che, nonostante una predominanza di punteggi elevati rimanga costante in entrambe le epoche, nel periodo successivo al 2012 si osserva un aumento delle recensioni con valutazioni inferiori, indicando una maggiore presenza di opinioni negative rispetto al periodo precedente.

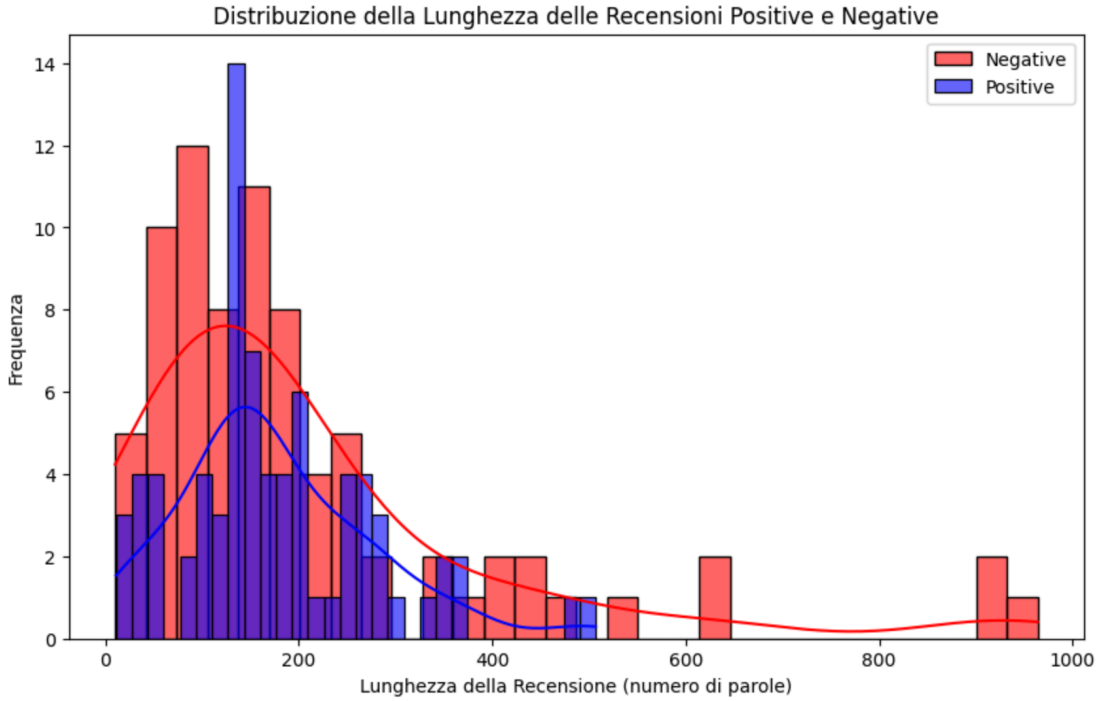


Per comprendere i momenti specifici in cui la serie ha avuto meno successo rispetto a quando è stata maggiormente apprezzata, è fondamentale analizzare l'andamento delle recensioni positive e negative nel corso degli anni. Data la forte prevalenza di recensioni positive, è stato necessario eseguire una trasformazione logaritmica dei valori registrati ogni anno in maniera da analizzare al meglio anche l'andamento della classe minoritaria.

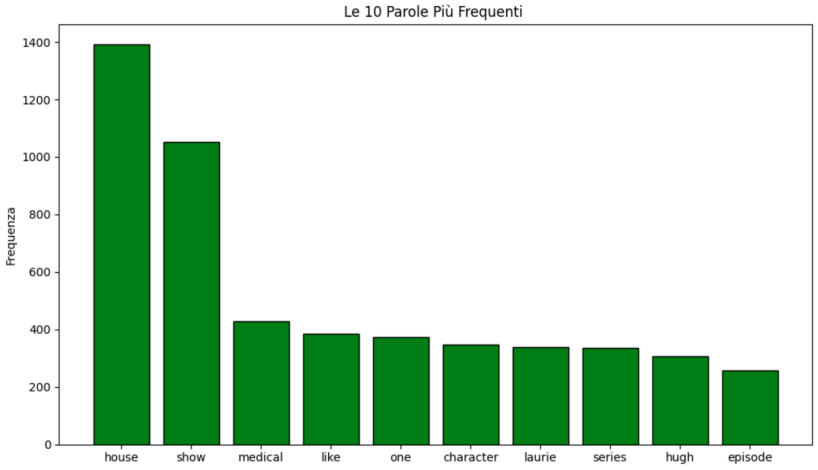
Come mostrato nei grafici precedenti, nel periodo successivo alla conclusione della serie e in particolare dal 2018 al 2022, si è verificato un aumento delle recensioni negative. Altri due picchi di recensioni negative si sono registrati negli anni 2008 e 2011, corrispondenti rispettivamente alle stagioni 4 e 7.



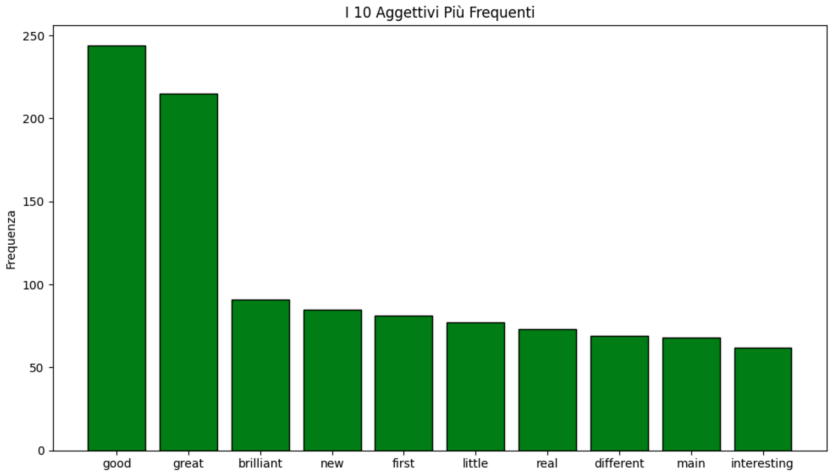
Per quanto riguarda le recensioni positive, il punto di massimo è stato raggiunto nel 2006, anno che coincide con la fine della prima stagione della serie. Questo suggerisce che la prima stagione abbia ricevuto un'ottima accoglienza, mentre le stagioni successive hanno incontrato un'accoglienza più critica in determinati periodi.
Continuando l'analisi, ci concentriamo sulla relazione tra la lunghezza del testo delle recensioni, rappresentata dalla variabile "Text", e il sentiment espresso. L'indagine sulle lunghezze dei testi evidenzia una distribuzione decentrata, con una media di 174 parole, una minima di 9 e una massima di 1295. Tuttavia, non sembra emergere una differenza significativa nella lunghezza tra recensioni positive e negative. Le recensioni più lunghe potrebbero essere associate a sentimenti negativi, ma è necessario approfondire ulteriormente questa ipotesi attraverso un'analisi più dettagliata.

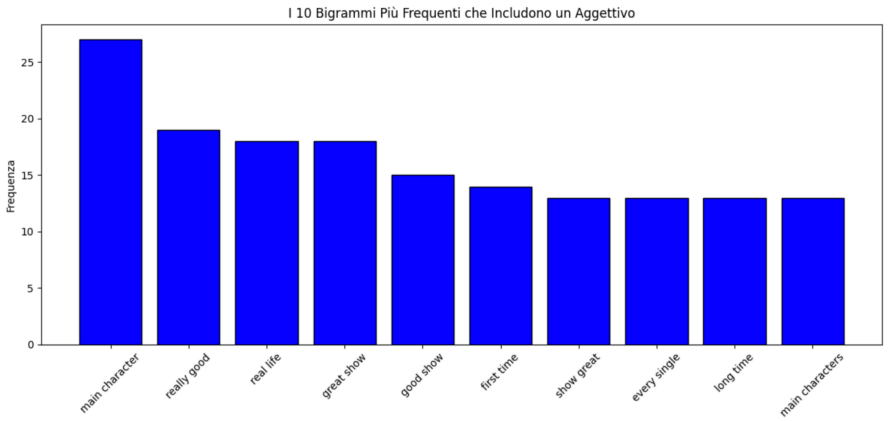


Entrando nel dettaglio dell'analisi dei testi, abbiamo condotto un'indagine sulla frequenza delle parole all'interno delle recensioni. Come evidenziato nel grafico, le parole più comuni sono quelle generalmente attese nel contesto della serie.

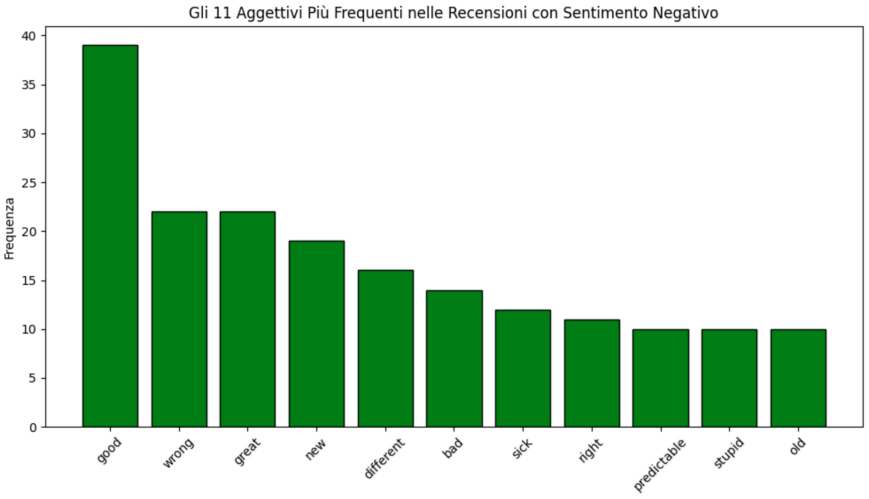


Successivamente, abbiamo esaminato una selezione di parole specifiche ai fini della valutazione del sentimento. Attraverso questa selezione, abbiamo identificato i primi dieci aggettivi e i primi dieci bigrammi contenenti aggettivi.





Risultati coerenti con le valutazioni degli utenti sono emersi sia dagli aggettivi singoli che dalle coppie di parole, indicando un tono generalmente positivo nelle recensioni.
Inoltre, focalizzandoci esclusivamente sulle recensioni negative, abbiamo individuato nuovi aggettivi con connotazioni negative, coerenti con le valutazioni meno positive degli utenti.



In conclusione, possiamo affermare che la serie ha ottenuto e mantenuto un ottimo riscontro da parte del pubblico, specialmente nei primi anni di uscita. Tuttavia, nonostante la grande accoglienza iniziale, sembra che negli anni successivi alla fine della serie si siano manifestati più dissensi da parte degli spettatori. Questo potrebbe essere dovuto a diversi fattori, come l'aumento di show a tema ospedaliero o la presenza di personaggi simili al protagonista di Dr. House con cui fare paragoni. Nel 2004, quando uscì la prima stagione, il concept era una novità, il che potrebbe spiegare l'entusiasmo iniziale rispetto al periodo post-finale della serie.

3.BAG OF WORDS

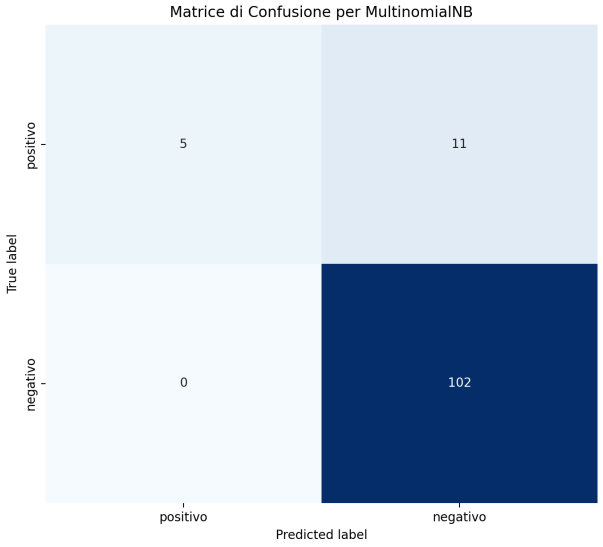
A questo punto, iniziamo con il primo approccio verso l'analisi del sentiment delle nostre recensioni. Questo primo step rappresenta anche quello più semplice e intuitivo dal punto di vista tecnico. Il metodo Bag of Words, infatti, permette di creare dei vettori basati sulla frequenza delle parole all'interno di un corpus di documenti. La sua semplicità presenta tuttavia alcuni difetti considerevoli: non cattura il significato semantico delle parole né le loro relazioni contestuali, ignora completamente l'ordine delle parole, per cui frasi diverse con le stesse parole avranno la stessa rappresentazione, e può generare vettori molto grandi e sparsi se il corpus contiene molte parole uniche. Prima di applicare questo metodo, è necessario effettuare alcune operazioni di pre-processing, quali l'eliminazione di stop words, la conversione dei caratteri in minuscolo e un processo di tokenizzazione. Queste funzioni sono fornite principalmente da due librerie fondamentali per tutti i modelli e approcci di cui discuteremo: scikit-learn e NLTK. Quest'ultima, in particolare, è specifica per problemi di classificazione del testo e linguaggio naturale.

Gli algoritmi implementati per questo progetto includono Decision Tree, Naive Bayes e SVM. Come metrica di valutazione, l'F1 score risulta la più performante, in quanto tiene

conto del dataset sbilanciato. Parlando di punteggi, nelle tabelle sottostanti sono riportati gli score dei singoli modelli e le relative matrici di confusione.

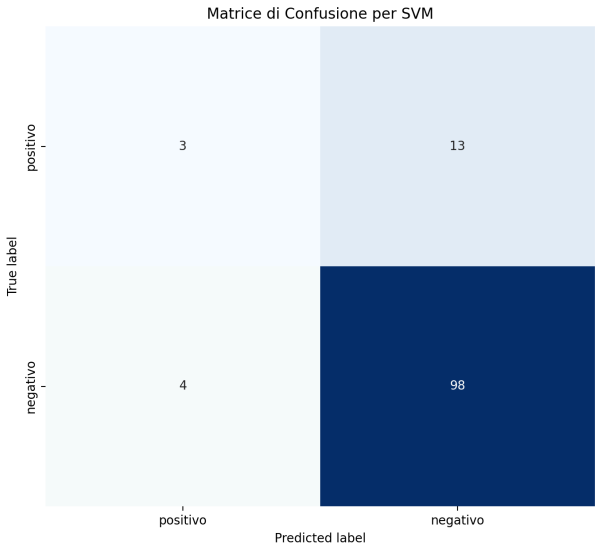
MultinomialNB
F1-score totale: 0.88

	Precision	Recall	F1-Score
Negativo	1.00	0.31	0.48
Positivo	0.90	1.00	0.95



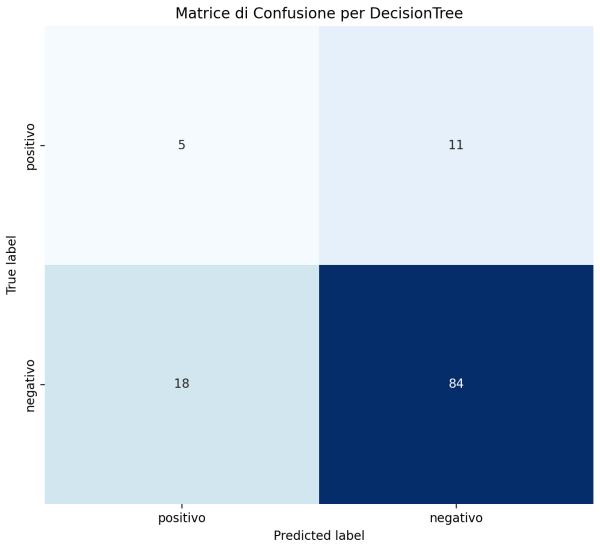
Support Vector Machine (SVM)
F1-score totale: 0.83

	Precision	Recall	F1-Score
Negativo	0.43	0.19	0.26
Positivo	0.88	0.96	0.92



Decision Tree
F1-score totale: 0.77

	Precision	Recall	F1-Score
Negativo	0.22	0.31	0.26
Positivo	0.88	0.82	0.85



Il risultato è piuttosto soddisfacente in tutti e tre i casi. L'algoritmo Naive Bayes, tuttavia, è quello che performa meglio, con un punteggio totale dell'88% e con una percentuale di accuratezza per la classe negativa del 48%.

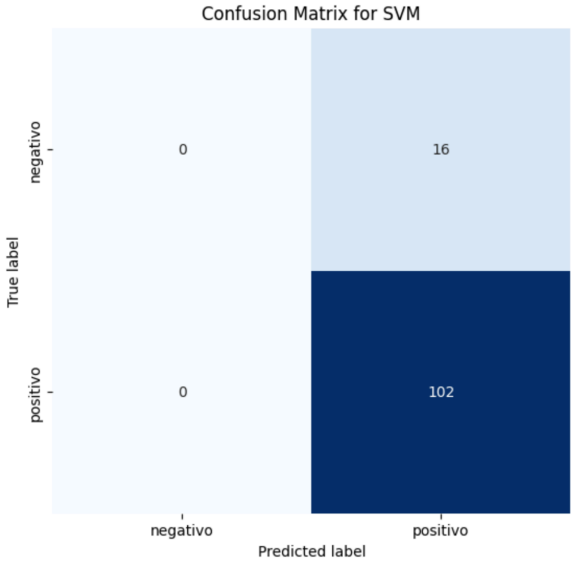
4.WORD2VEC

A differenza del Bag of Words, Word2Vec cattura le relazioni semantiche tra le parole, fornendo rappresentazioni vettoriali (embedding) che collocano parole con significati simili vicine nello spazio vettoriale. Word2Vec funziona attraverso due architetture principali: il Continuous Bag of Words (CBOW) e il Skip-Gram. CBOW predice una parola dato il contesto delle parole circostanti, mentre Skip-Gram fa il contrario, predice il contesto circostante data una parola centrale. Il risultato dell'addestramento è un insieme di vettori di parole dove la distanza e la direzione del vettore spaziale riflettono le relazioni semantiche e sintattiche tra le parole. Word2Vec ha molte applicazioni pratiche, tra cui la traduzione automatica e la sintesi del linguaggio naturale. Inoltre, i vettori di Word2Vec possono essere utilizzati come caratteristiche di input per altri modelli di machine learning. Nonostante i suoi vantaggi, ha alcune limitazioni. Richiede grandi quantità di dati per addestrare modelli accurati e la qualità delle rappresentazioni può dipendere fortemente dalla dimensione e dalla qualità del corpus di addestramento. La sua implementazione è stata possibile grazie alla

libreria open-source Gensim, la quale è stata progettata per il processamento e l'analisi di dati testuali non strutturati. In questa circostanza, si sono implementati gli stessi modelli usati in precedenza, aggiungendo RandomForest.

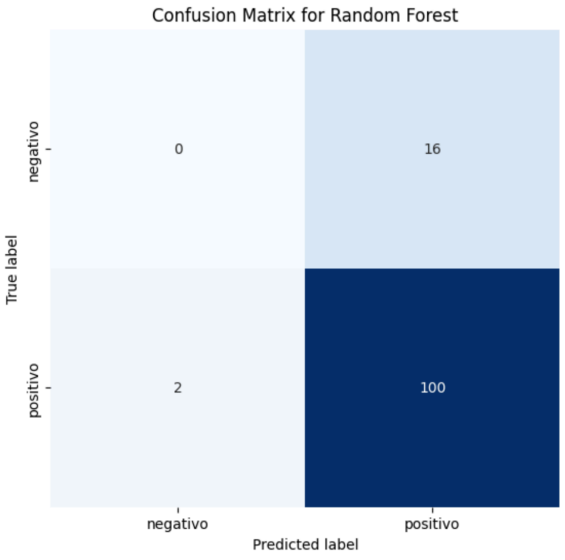
Support Vector Machine (SVM)
F1-score complessivo: 0.80

	Precision	Recall	F1-Score
Negativo	0	0	0
Positivo	0.86	1	0.93



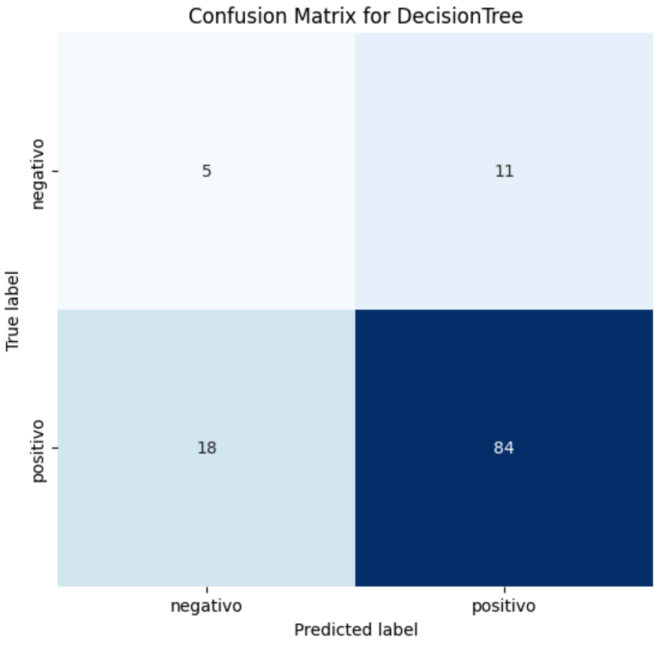
Random Forest
F1-score complessivo: 0.79

	Precision	Recall	F1-Score
Negativo	0	0	0
Positivo	0.86	0.98	0.92



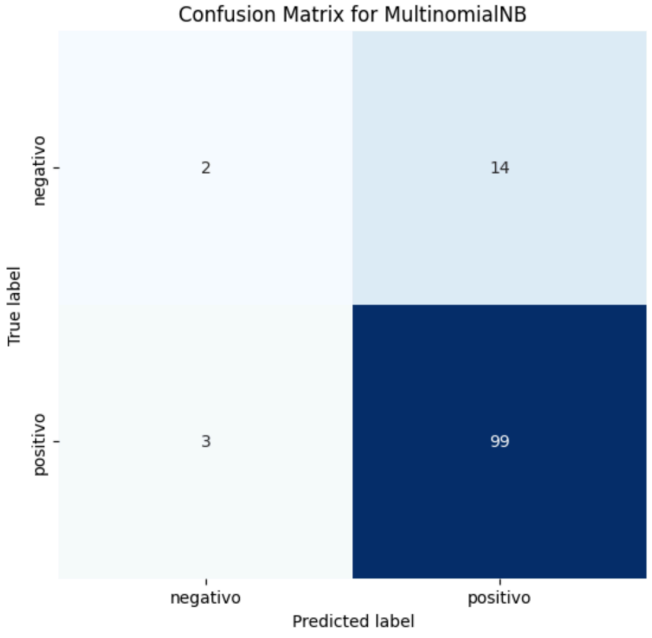
Decision Tree
Fi-score complessivo: 0.77

	Precision	Recall	F1-Score
Negativo	0.22	0.31	0.26
Positivo	0.88	0.82	0.85



MultinomialNB
F1-score complessivo: 0.82

	Precision	Recall	F1-Score
Negativo	0.40	0.12	0.19
Positivo	0.88	0.92	0.92



I risultati, anche in questo caso, in termini di F1 sono stati abbastanza soddisfacenti, con tutti i punteggi generali intorno all'80%, con Naive Bayes che toccano l'82%. Nonostante ciò, risulta evidente come questi valori vengano trainati principalmente dalle prestazioni sulla classe maggioritaria, evidenziando una carenza importante nella predizione delle occorrenze negative. DecisionTree su questo aspetto risulta il più performante, ottenendo un punteggio del 26%, il quale però rappresenta un valore decisamente troppo basso per i nostri scopi.

5.TF-IDF

Una delle tecniche fondamentali nell'ambito della sentiment analysis è il TF-IDF, che ci aiuta a valutare l'importanza delle parole nei documenti rispetto all'intero insieme di documenti disponibili. Questo processo coinvolge due aspetti principali: la frequenza delle parole all'interno di un documento (TF) e la loro importanza rispetto all'intero corpus (IDF).
A questa tecnica abbiniamo un processo di tokenizzazione diverso dai precedenti basato su subwords, il quale è stato fornito dall'implementazione della libreria open source Sentencepiece. Nonostante l'applicazione di modelli di machine learning associati a queste tecniche, i risultati ottenuti non sembrano essere soddisfacenti. Questo potrebbe essere attribuito alla lunghezza delle recensioni, che potrebbero essere costituite da periodi particolarmente estesi. Questa caratteristica delle recensioni potrebbe influenzare le prestazioni dei modelli, portando a una F1-score massima del 64%, che non può considerarsi ottimale per le nostre esigenze di analisi del sentiment.

MultinomialNB
F1-score complessivo: 0.46

	Precision	Recall	F1-Score
Negativo	0	0	0
Positivo	0.86	1	0.92

Support Vector Machine
F1-score complessivo: 0.46

	Precision	Recall	F1-Score
Negativo	0	0	0
Positivo	0.88	0.92	0.92

Decision Tree
F1-score complessivo: 0.48

	Precision	Recall	F1-Score
Negativo	0.39	0.41	0.40
Positivo	0.90	0.89	0.90

6.TRANSFORMERS

L'architettura Transformer ha rivoluzionato la NLP grazie al suo meccanismo di attenzione. Questo meccanismo permette al modello di cogliere relazioni a lungo raggio tra parole all'interno di una sequenza, eliminando la necessità di ricorrenze esplicite come LSTM o RNN. Di conseguenza, i Transformer possono elaborare sequenze di input in parallelo, ottenendo un notevole incremento di prestazioni ed efficienza.
Tra i modelli linguistici di grandi dimensioni (LLM) basati su Transformer, BERT si distingue per le sue elevate prestazioni in diverse attività NLP, tra cui la comprensione del linguaggio naturale, la generazione di testo e la classificazione del testo.
Per implementare l'architettura Transformer, è stato fondamentale individuare un modello pre-addestrato idoneo. In questo contesto, la piattaforma Hugging Face ha svolto un ruolo cruciale, offrendo un'ampia raccolta di modelli linguistici avanzati. Tra questi, BERT si è distinto come uno degli LLM (Large Language Model) più potenti per il text mining, risultando adatto al compito di analisi del sentiment.
La selezione finale è ricaduta sul modello "cardiffnlp/twitter-roberta-base-sentiment", classificandosi tra i primi 10 modelli più diffusi su Hugging Face per i task di text classification. Con oltre 3 milioni di download solo nell'ultimo mese, questo modello vanta un dataset di training supervisionato di 124 milioni di tweet, annotati con tre etichette di sentiment: positivo, negativo e neutro.

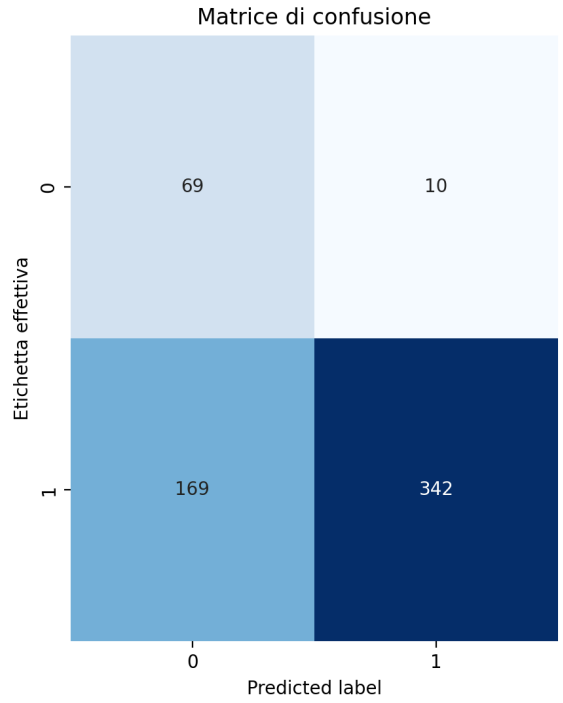
6.1 Implementazione del modello in due fasi

- **Fase 1: Utilizzo Diretto del Modello**
  Nella prima fase, il modello pre-addestrato è stato applicato direttamente ai dati estratti da IMDb. Questa fase ha permesso di ottenere una valutazione preliminare delle prestazioni del modello e di identificare eventuali aree di miglioramento.

Il primo tentativo di applicazione del modello BERT ha prodotto un F1-score generale del 74%, con una performance sulla classe minoritaria del 44%. Tuttavia, esaminando la matrice di confusione, emergono le reali performance del modello. In particolare, possiamo dire che il modello è stato molto esaustivo riguardo la classe minoritaria, sbagliando molto di più invece sulla classe dei positivi. Questo errore era in parte prevedibile, dato che le etichette del modello BERT differiscono da quelle binarie applicate nel nostro dataset. Probabilmente, l'assenza della categoria "neutro" ha contribuito a generare l'errore osservato.

F1-score complessivo: 0.74

	Precision	Recall	F1-Score
Negativo	0.29	0.87	0.44
Positivo	0.97	0.67	0.79



Note: label 0= negativo, label 1= positivo

➢ **Fase 2: Fine-Tuning del Modello**

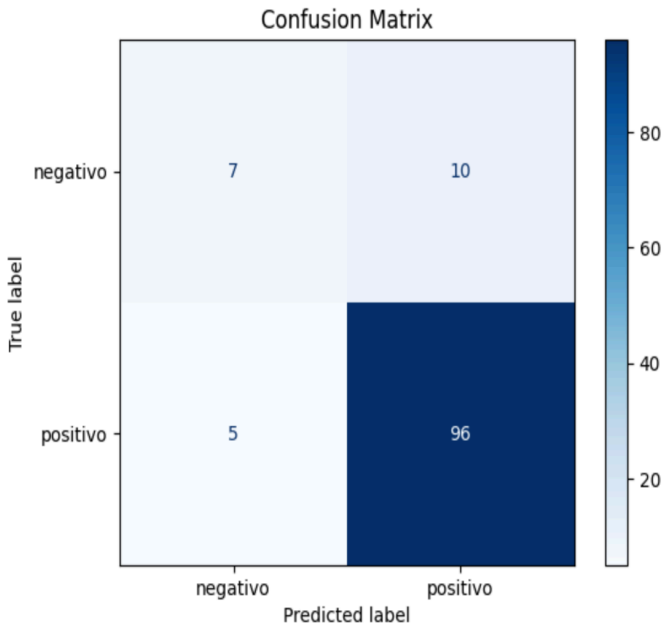
Per ottimizzare ulteriormente le prestazioni del modello, è stato eseguito un processo di fine-tuning. Questa tecnica consiste nel riaddestrare il modello pre-addestrato su un dataset più piccolo e specifico, in questo caso le recensioni di film di IMDb. Il fine-tuning ha permesso di adattare il modello al dominio specifico del dataset, migliorando significativamente la sua accuratezza nell'analisi del sentiment.

6.2 FINE TUNING

Dopo aver constatato l'inadeguatezza del modello BERT nel fare inferenza sul nostro dataset, si è reso necessario un processo di fine tuning del modello. Sono stati effettuati diversi step, con parametri scelti di volta in volta.

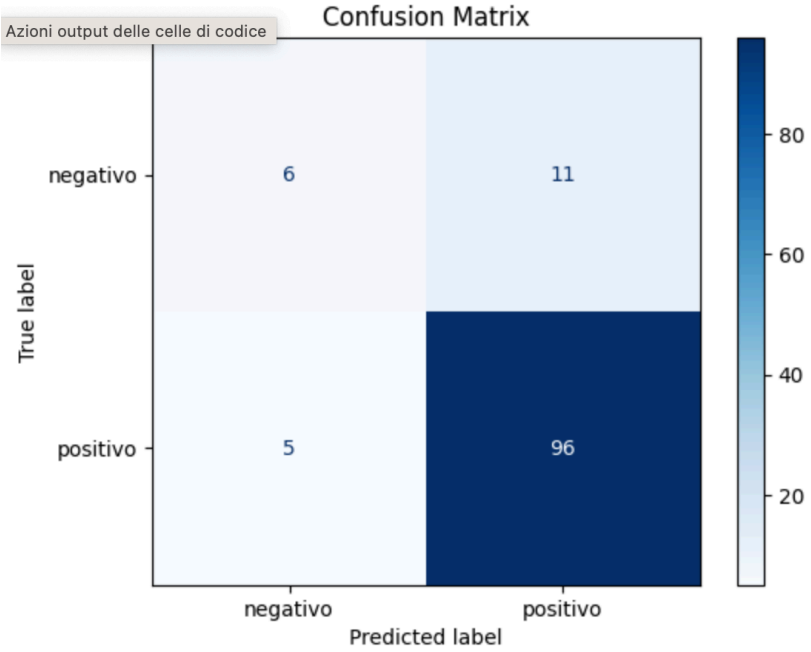
6.2.1. Primo Tentativo di Fine Tuning

Nel primo tentativo di fine tuning, sono state effettuate 3 epoche di addestramento. Sebbene questo numero sia relativamente basso rispetto agli standard, i risultati sono migliorati, come evidenziato dalla matrice di confusione. La classe dei positivi è stata individuata molto bene, mentre quella dei negativi ha presentato un maggior numero di falsi positivi rispetto ai negativi predetti correttamente. Considerando lo sbilanciamento del dataset, questo errore è comprensibile ma richiede ulteriori miglioramenti.



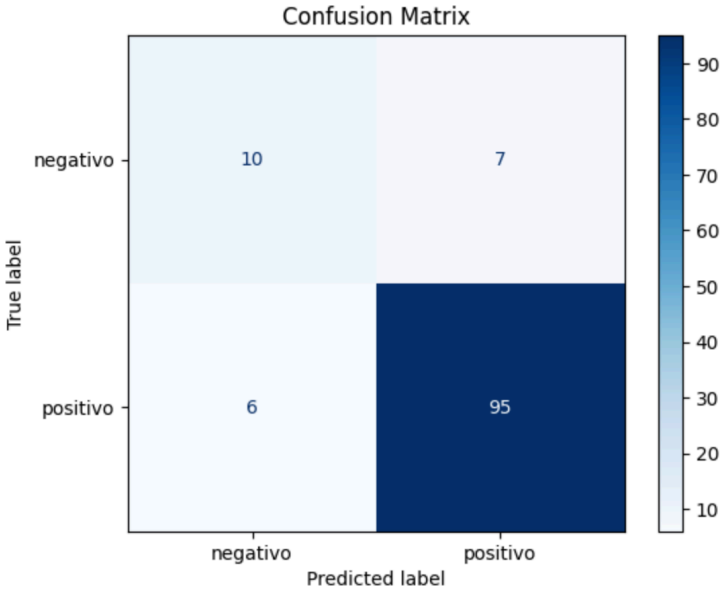
6.2.2. Secondo Tentativo di Fine Tuning

Nel secondo tentativo, sono state effettuate 10 epoche di addestramento. I risultati ottenuti non mostrano differenze significative rispetto al primo tentativo, anzi si osserva un leggero peggioramento nei falsi positivi.



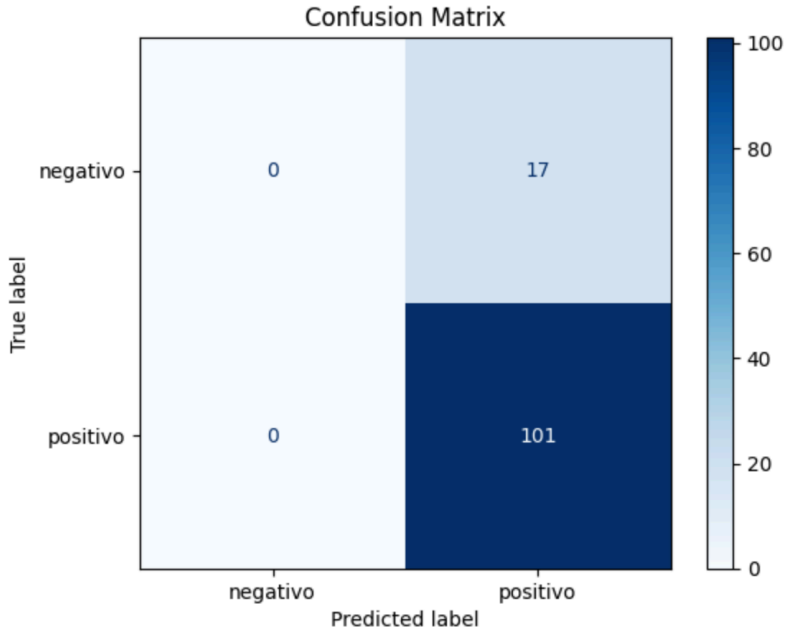
6.2.3. Terzo Tentativo di Fine Tuning

Il terzo tentativo ha comportato un addestramento di 15 epoche. In questo caso, si è osservato un miglioramento effettivo nella predizione della classe minoritaria, con un aumento delle predizioni corrette rispetto ai falsi positivi.



6.2.4. Quarto Tentativo di Fine Tuning

Nell'ultimo tentativo, il numero di epoche è stato incrementato a 20. In questa fase, il modello ha catturato molto bene la classe maggioritaria, ma ha performato male con la classe minoritaria, classificando tutte le occorrenze come positive.



7.WEIGHTS & BIASES

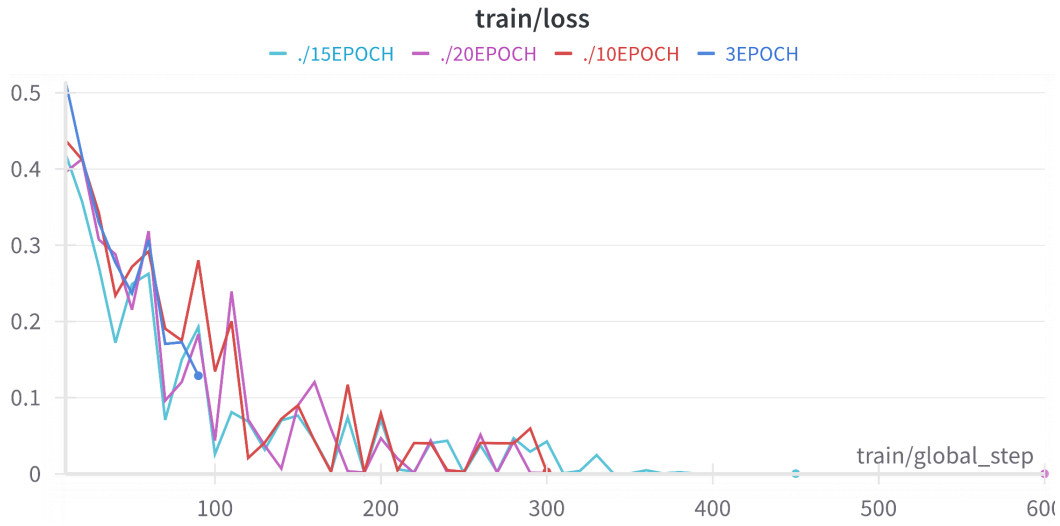
Per valutare graficamente le prestazioni dei modelli, abbiamo utilizzato uno strumento supplementare: Weight & Biases. Questa piattaforma è ampiamente utilizzata nella valutazione e rappresentazione dei modelli di machine learning. Nel nostro caso, Weight & Biases offre un supporto grafico per i parametri specifici presenti nel modello utilizzato, basato su RoBERTa. L'andamento delle funzioni di cui discuteremo viene rappresentato con le epoche, le quali sono espresse in termini di Steps. Nello specifico un'epoca corrisponde all'incirca a 30 Steps, quindi i valori dell'ascissa andranno da 0 a 600 Steps, corrispondenti a 20 epoche.

7.1.Funzione di Perdita (Loss)

Uno dei parametri visualizzati è l'andamento della funzione di perdita, gestita automaticamente dal modello e corrispondente, nel nostro caso, alla Cross-Entropy Loss. Questa funzione misura la differenza tra le distribuzioni di probabilità predette dal modello e le distribuzioni delle etichette reali.

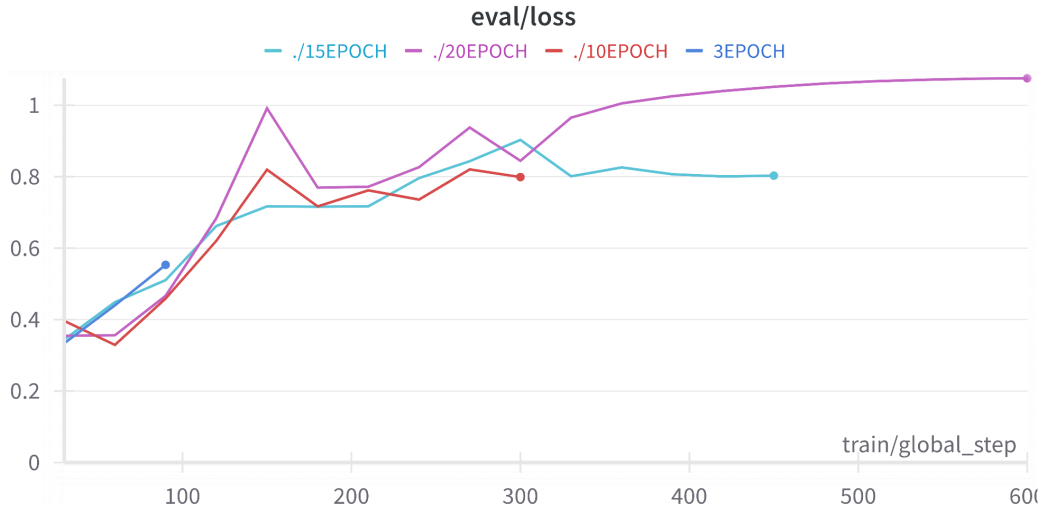
7.1.1.Fase di Training

Durante la fase di training, si osserva che il valore della loss diminuisce con l'aumentare delle epoche per tutti e quattro i modelli. In particolare, i modelli a 10, 15 e 20 epoche si stabilizzano intorno alla decima epoca (300 step), mantenendo poi un valore costante prossimo allo zero. Questo andamento decrescente rende tutti i modelli ottimali, almeno in questa fase.



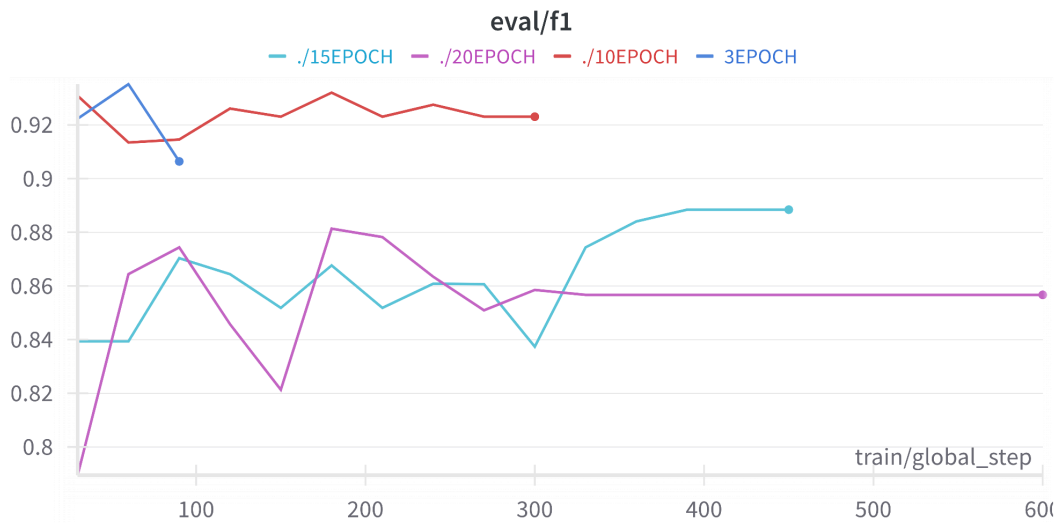
7.1.2.Fase di Valutazione

Nella fase di valutazione, si nota che tutti e quattro i modelli mostrano un andamento crescente della loss. Il modello a 20 epoche, in particolare, mostra un incremento più spiccato rispetto agli altri modelli a partire dalla decima epoca (300 step), stabilizzandosi successivamente a un valore poco superiore a 1. Nel complesso quindi, in termini di Loss tutti i modelli hanno un andamento rassicurante.



7.2.Andamento della F1-Score

L'ultimo grafico mostra l'andamento della F1-score durante le epoche. Anche se i modelli con 15 e 20 epoche mantengono un punteggio valido, quello a 10 epoche risulta nettamente migliore, con un punteggio di 0.92. Questo conferma ulteriormente la preferibilità del modello a 10 epoche, che offre un equilibrio ottimale tra capacità di apprendimento e generalizzazione.



8.CONCLUSIONI

In generale, le recensioni della serie "Dr. House" risultano poco polarizzate, mostrando uno sbilanciamento netto verso il sentiment positivo. Dal punto di vista delle prestazioni dei modelli impiegati per la stima del sentiment, possiamo affermare che, sia in termini di F1 score che nei risultati evidenziati dalle matrici di confusione, quasi tutti i modelli hanno mantenuto un livello di prestazioni costante, sia per la classe dei sentiment positivi che per quella dei negativi.

Considerando solo le prestazioni sul dataset da noi analizzato, il modello basato su Bag of Words ha mantenuto un livello di accuratezza paragonabile a quello dei modelli più moderni, come i transformers. Tuttavia, se prendiamo in considerazione la capacità di fare inferenza su dati non supervisionati, il modello basato su BERT, addestrato per 10 epoche, ha dimostrato di essere particolarmente efficace nel riconoscere anche le osservazioni della classe minoritaria.

In generale, quindi, il miglior modello risulta essere quello basato su BERT, anche se è evidente che vi siano margini di miglioramento. I fattori principali che hanno limitato le prestazioni del modello sono stati principalmente due:

1. Dataset sbilanciato: Come già evidenziato, la presenza di un dataset molto sbilanciato ha reso comprensibilmente più difficile prevedere correttamente le recensioni negative.
2. Tipo di modello BERT utilizzato: Il pre-addestramento di BERT con dati provenienti da tweet probabilmente non ha fornito una base di partenza ottimale, a causa delle differenze tra i tweet e le recensioni. Una differenza rilevante è la lunghezza del testo. Va tuttavia considerato che, sebbene i tweet siano generalmente più brevi delle recensioni, condividono con queste ultime la caratteristica del testo non strutturato tipico del linguaggio naturale, il che ha probabilmente compensato parzialmente la capacità del modello di fare classificazione.

In conclusione, nonostante le limitazioni, il modello BERT si è dimostrato il più efficace per il compito di classificazione del sentiment delle recensioni della serie "Dr. House", pur lasciando spazio a futuri miglioramenti.

The Data Cards Playbook

RECENSIONI_BINARY	Il dataset contiene le recensioni della serie "Dr. House" raccolte da IMDb. Include recensioni testuali degli utenti, utilizzate per un task di sentiment analysis. Le recensioni riflettono una gamma di opinioni su vari aspetti della serie.
DATASET LINK https://www.imdb.com/title/tt0412142/reviews?ref_=tt_urv	DATA CARD AUTHOR(S) **IMDb:** Owner **Alessandro Carmellini:** Contributor

Authorship

Publishers

PUBLISHING ORGANIZATION(S)	INDUSTRY TYPE(S)	CONTACT DETAIL(S)
Libera Università Maria Santissima Assunta (LUMSA)	Academic - Tech	**Website:** https://www.imdb.com/title/tt0412142/reviews?ref_=tt_urv

Dataset Owners

TEAM(S)	CONTACT DETAIL(S)	AUTHOR(S)
IMDb	**Dataset Owner:** IMDb **Group Email:** a.carmellini@lumsastud.it	IMDb Alessandro Carmellini

Funding Sources

INSTITUTION(S)	FUNDING OR GRANT SUMMARY(IES)
Libera Univeristà Maria Santissima Assunta (LUMSA)	

Dataset Overview

DATA SUBJECT(S)

Non-Sensitive Data about people

Data about products

DATASET SNAPSHOT

Size of Dataset	721 KB
Number of Instances	700
Number of Fields	4

CONTENT DESCRIPTION

Il dataset contiene le recensioni della serie "Dr. House" raccolte da IMDb. Include recensioni testuali degli utenti, utilizzate per un task di sentiment analysis. Le recensioni riflettono una gamma di opinioni su vari aspetti della serie.

DESCRIPTIVE STATISTICS

Statistic	User	Stars	Date	Text
count	700	584	700	700
mean		9		175
min		1		9
max		10		1295

Additional Notes: Le unità considerate per la variabile Text sono le singole parole non univoche

Dataset Version and Maintenance

MAINTENANCE STATUS

Limited Maintenance

(The data will not be updated, but any technical issues will be addressed.)

VERSION DETAILS

Current Version: 1.0

Last Updated: 05/2024

Release Date: 05/2024

MAINTENANCE PLAN

Versioning: No

Updates: Il dataset non verrà aggiornato

Feedback: No

Additional Notes: No

Example of Data Points

PRIMARY DATA MODALITY	SAMPLING OF DATA POINTS	DATA FIELDS
Text Data	No sampling (full data)	

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Retrieved from repository website	**Source:** https://www.imdb.com/title/tt0412142/reviews?ref_=tt_urv **Platform:** IMDb Website **Is this source considered sensitive or high-risk?** No **Dates of Collection:** [MAR 2024 - MAR 2024] **Primary modality of collected data:** Text Data	

Motivations & Intentions

Motivations

PURPOSE(S)	DOMAIN(S) OF APPLICATION	MOTIVATING FACTOR(S)
Mining Training Testing Validation	Sentiment analysis	*Analizzare il testo delle recensioni per estrarre il relativo sentiment*

Intended Use

DATASET USE(S)	SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
Safe for production use Safe for research use	**Suitable Use Case** : Scopo didattico **Suitable Use Case**: Scopo commerciale **Suitable Use Case**: Marketing	**Unsuitable Use Case** : Valutazione dell'Accuratezza Medica

Access, Retention, & Wipeout

Access

ACCESS TYPE	DOCUMENTATION LINK(S)	PREREQUISITE(S)
External - Open Access	https://www.imdb.com/title/tt0412142/reviews?ref_=tt_urv	

Provenance

Collection

METHOD(S) USED	METHODOLOGY DETAIL(S)	SOURCE DESCRIPTION(S)
Scraping web	**Platform:** IMDb **Is this source considered sensitive or high-risk?** No **Dates of Collection:** MAR- 2024- MAR- 2024 **Primary modality of collected data:** Text Data **Additional Notes:** <Add here>	

COLLECTION CADENCE	DATA INTEGRATION	DATA PROCESSING
Static (Data was collected once from single or multiple sources.)	**<Source>** **Included Fields** (Data fields that were collected and are included in the dataset.)	**<Collection Method or Source>** **Description:** Scraping web **Tools or libraries:** Pandas, Selenium, Beautifulsoup

Field Name	Description
User	La colonna specifica il nome dell'utente che ha scritto la recensione
Stars	Variabile numerica di interi da 1 a 10, opzionale, che accompagna la recensione testuale
Date	La data in cui è stata caricata la recensione sul sito
Text	Il testo vero e proprio della recensione

Version and Maintenance

	FIRST VERSION	NOTE(S) AND CAVEAT(S)
	Release date: 05/2024 **Status:** Limited Maintenance **Size of Dataset:** 711 KB **Number of Instances:** 700	
CADENCE	LAST AND NEXT UPDATE(S)	CHANGES ON UPDATE(S)
Static	**Date of last update:** 30/05/2024	

Extended Use

Use with Other Data

SAFETY LEVEL	KNOWN SAFE DATASET(S) OR DATA TYPE(S)	BEST PRACTICES
Safe to use with other data Conditionally safe to use with other data		

Forking & Sampling

SAFETY LEVEL	ACCEPTABLE SAMPLING METHOD(S)	BEST PRACTICE(S)
Safe to form and/or sample	Cluster Sampling Multi-stage Sampling Random Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unsampled	

Use in ML or AI Systems

DATASET USE(S)	NOTABLE FEATURE(S)	USAGE GUIDELINE(S)
Training Testing Validation Fine Tuning		

Reflections on Data

Recensioni Serie	I dati presenti nel dataset sono stati estratti dal sito IMDb al solo scopo didattico

---