UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea Magistrale in Ingegneria Informatica e dell'Automazione

PROGETTO DI DATA SCIENCE

BERT

Docenti

Ursino Domenico,
Bonifazi Gianluca

A cura di

Visi Andrea,
De Grazia Davide,
Piergallini Enrico,

ANNO ACCADEMICO 2024-2025

Indice

1 IL DATASET SCELTO **1**
1.1 Descrizione del dataset 1

2 Introduzione a BERT **3**
2.1 Elaborazione del Linguaggio Naturale (NLP) 3
2.2 Transformer 3
2.3 Introduzione a BERT 3

3 SVILUPPO **5**
3.1 Analisi preliminare del dataset 5
3.2 Preprocessing e tokenizzazione del testo 6
3.2.1 Preprocessing del testo 6
3.2.2 Tokenizzazione 7
3.2.3 Distribuzione della lunghezza dei token 7
3.3 Training 8
3.3.1 Valutazione dei modelli 9

Elenco delle figure

3.1 Distribuzione iniziale del sentiment nel dataset. 5
3.2 Distribuzione del sentiment dopo il bilanciamento. 6
3.3 Esempio di tokenizzazione. 7
3.4 Visualizzazione del DataFrame durante le diverse fasi di preprocessing e tokenizzazione. 7
3.5 Distribuzione della lunghezza delle frasi in token. 8
3.6 Andamento della Training Loss e Validation Loss per il modello BERT. 9
3.7 Andamento della Training Loss e Validation Loss per il modello RoBERTa. . . 9
3.8 Andamento della Training Loss e Validation Loss per il modello DistilBERT. . 9
3.9 Confusion Matrix per i modelli BERT, RoBERTa e DistilBERT sul dataset di test. 10
3.10 BERT 10
3.11 RoBERTa 10
3.12 DistilBERT 10

CAPITOLO 1

IL DATASET SCELTO

In questo capitolo verrà fornita una descrizione sintetica ma esaustiva del dataset selezionato per condurre un'analisi del sentiment utilizzando BERT. Il dataset, focalizzato sul contesto finanziario delle notizie, è stato scelto su Kaggle ed è reperibile al seguente link: `https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis`.

1.1 Descrizione del dataset

Il dataset `data.csv` è stato creato per supportare la ricerca nell'analisi del sentiment finanziario, combinando due importanti risorse: il *FiQA* e il *Financial PhraseBank*. Questo dataset unificato fornisce una rappresentazione semplice e immediata, in formato CSV, contenente frasi finanziarie etichettate con il relativo sentiment. È particolarmente utile per sviluppare e testare modelli di analisi del sentiment in contesti finanziari.

Il dataset è composto da un totale di 5322 righe, ciascuna rappresentante una singola frase estratta da notizie o dichiarazioni finanziarie, accompagnata dall'etichetta del sentiment corrispondente. Le colonne principali sono:

- `Sentence`: contiene il testo della frase relativa al contesto finanziario. Ogni riga del dataset include una singola frase estratta da articoli, comunicati o report economici.
- `Sentiment`: rappresenta un'etichetta categoriale che indica il sentiment associato alla frase. Sono disponibili tre categorie principali:
  - `negative`: indica che la frase ha un sentiment negativo. Questo tipo di frasi suggerisce, in genere, un potenziale impatto negativo sugli indicatori finanziari, come una diminuzione dei prezzi delle azioni o un peggioramento delle prospettive economiche.
  - `neutral`: indica che la frase è neutrale. Tali frasi rappresentano informazioni che non hanno un impatto significativo o diretto sul sentiment generale, limitandosi spesso a riportare dati o fatti oggettivi.
  - `positive`: indica che la frase ha un sentiment positivo. In contesti finanziari, ciò suggerisce una potenziale crescita o miglioramento, come un aumento dei prezzi delle azioni o una prospettiva economica favorevole.

Di seguito è riportato un esempio rappresentativo di come sono strutturati i dati nel dataset:

Sentence	**Sentiment**
"The company posted a strong quarterly revenue growth."	positive
"Investors remain cautious amid growing global uncertainties."	negative
"The market closed flat after a day of mixed trading activity."	neutral
"Strong job reports boost confidence in the economic recovery."	positive
"Concerns over inflation continue to weigh on market sentiment."	negative

Tabella 1.1: Esempio di dati nel dataset `data.csv`.

Il dataset presenta un mix di frasi con sentiment positivo, negativo e neutrale, offrendo una base ideale per sviluppare e valutare modelli di *machine learning* e *deep learning* per l'analisi del sentiment finanziario. La sua dimensione contenuta, pari a circa 750KB, garantisce una gestione agevole durante la fase di elaborazione e training, senza compromettere la diversità e la qualità delle informazioni contenute.

Grazie alla combinazione di due dataset consolidati e affidabili, il `data.csv` rappresenta una risorsa preziosa per esplorare le relazioni tra le dichiarazioni finanziarie e il sentiment che esse generano, fornendo informazioni utili per analisi approfondite o applicazioni pratiche nel settore finanziario.

CAPITOLO 2

Introduzione a BERT

Questo capitolo offre una panoramica sulle basi dell'NLP, sull'architettura Transformer e sull'importanza di BERT nel rivoluzionare le applicazioni di analisi e comprensione del linguaggio naturale.

2.1 Elaborazione del Linguaggio Naturale (NLP)

L'Elaborazione del Linguaggio Naturale (NLP) è un campo interdisciplinare che combina informatica, intelligenza artificiale e linguistica computazionale per analizzare, rappresentare e comprendere il linguaggio umano. L'obiettivo principale dell'NLP è sviluppare sistemi in grado di processare testi o discorsi in modo simile a come lo farebbe un essere umano. Applicazioni comuni includono la traduzione automatica, i chatbot, l'analisi del sentiment e i sistemi di risposta alle domande. Negli ultimi anni, l'integrazione di tecniche di *deep learning* ha migliorato significativamente le prestazioni nei compiti di NLP, portando alla nascita di modelli come BERT.

2.2 Transformer

Il modello Transformer, introdotto da Vaswani et al. nel 2017, ha rivoluzionato il campo del deep learning per l'NLP grazie alla sua capacità di elaborare intere sequenze di testo simultaneamente anziché in modo sequenziale. Il cuore del Transformer è il meccanismo di *attenzione multi-testa* (*multi-head attention*), che consente al modello di concentrarsi su diverse parti di una frase per comprenderne meglio il significato contestuale. Questa architettura, priva di vincoli sequenziali, permette l'elaborazione parallela dei dati e ha reso possibile l'addestramento su dataset di dimensioni enormi.

2.3 Introduzione a BERT

BERT (Bidirectional Encoder Representations from Transformers) è un modello di apprendimento automatico pre-addestrato che si basa sul modello Trasformer ed è progettato per affrontare una vasta gamma di compiti legati all'elaborazione del linguaggio naturale (NLP). Introdotto da Google nel 2018, BERT è rapidamente diventato uno dei modelli di NLP

più popolari e influenti, grazie alla sua capacità di produrre rappresentazioni contestuali di parole e frasi estremamente efficaci.

Prima dell'introduzione di BERT, i modelli di NLP erano limitati nella loro capacità di comprendere il contesto globale di una frase, poiché elaboravano le parole in modo unidirezionale (da sinistra a destra o viceversa). Questo approccio rendeva difficile catturare pienamente il significato complessivo di una frase o di un testo. BERT supera queste limitazioni utilizzando un pre-addestramento bidirezionale che analizza simultaneamente il contesto sia precedente che successivo di ogni parola, permettendo una comprensione più profonda e accurata del significato del testo.

Grazie al suo approccio bidirezionale, BERT riesce a catturare le sottigliezze semantiche. Una parola infatti, può avere più significati a seconda del constesto in cui questa viene utilizzata.

Il successo di BERT ha portato allo sviluppo di numerosi modelli derivati, ottimizzati per compiti e domini specifici. Tra i più noti vi sono:

- **BERT-large**: una versione più grande e potente di BERT-base, addestrata su un corpus di circa 7,5 miliardi di parole.
- **RoBERTa**: una variante sviluppata da Facebook AI Research (FAIR), addestrata su un corpus di testo ancora più ampio, pari a circa 160 GB.
- **DistilBERT**: una versione compressa di BERT che mantiene alte prestazioni con meno parametri, riducendo così i requisiti computazionali.
- **BioBERT**: un modello pre-addestrato ottimizzato per il text mining biomedico.

BERT è stato addestrato su due task principali:

- **Masked Language Modeling (MLM)**: durante l'addestramento, alcune parole di una frase vengono mascherate e il modello deve prevederle basandosi sul contesto circostante. Questa tecnica consente al modello di apprendere rappresentazioni linguistiche più ricche rispetto ai metodi precedenti.
- **Next Sentence Prediction (NSP)**: il modello riceve in ingresso una coppia di frasi e deve determinare se la seconda frase segue logicamente la prima. Questo task migliora la comprensione del contesto e della coerenza nel linguaggio naturale.

Per riassumere, l'introduzione di BERT ha rappresentato un importante passo avanti nell'elaborazione del linguaggio naturale, rendendolo uno strumento essenziale per molte applicazioni, tra cui la classificazione di testi, l'estrazione di informazioni, la risposta automatica alle domande e molte altre. Nel corso del prossimo capitolo verranno dettagliati tutti i passaggi che abbiamo compiuto al fine di applicare il modello BERT al nostro dataset, fornendo infine delle metriche quantitative per valutare la bontà del modello che abbiamo ottenuto.

CAPITOLO 3

SVILUPPO

3.1 Analisi preliminare del dataset

Il dataset scelto, come già descritto in precedenza, contiene 5322 righe con due colonne principali:

- `Sentence`, che rappresenta il testo di una frase in ambito finanziario.
- `Sentiment`, che può assumere tre valori: `negative`, `neutral` o `positive`.

Per iniziare l'analisi, è stata verificata la distribuzione iniziale delle etichette di sentiment al fine di valutare se il dataset fosse bilanciato. Di seguito, in Figura 3.1, è mostrato il grafico della distribuzione del sentiment nel dataset originale.

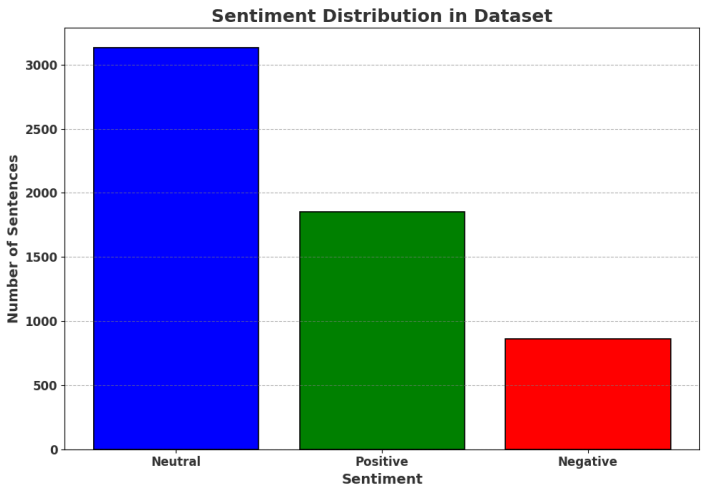


Figura 3.1: Distribuzione iniziale del sentiment nel dataset.

Come si può osservare dal grafico, il dataset risulta sbilanciato, con la seguente distribuzione delle etichette:

- `neutral`: 3124 campioni.
- `positive`: 1852 campioni.

- `negative`: 860 campioni.

Per procedere con l'addestramento di modelli di machine learning e deep learning, è importante che il dataset sia bilanciato, ovvero che ciascuna classe contenga un numero simile di campioni. Pertanto, è stato applicato un processo di bilanciamento che garantisce una distribuzione uniforme tra le classi.

Il bilanciamento è stato realizzato tramite la funzione `create_balanced_dataset()`, che consente di ottenere un numero uniforme di frasi per ciascuna classe (1666 campioni per classe), applicando il campionamento con sostituzione per le classi con un numero insufficiente di dati.

Il risultato del bilanciamento è mostrato in Figura 3.2, con una identica distribuzione delle etichette tra le classi.

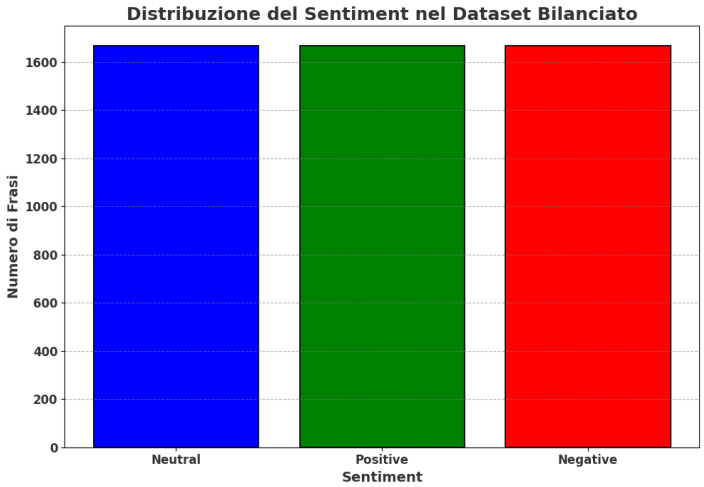


Figura 3.2: Distribuzione del sentiment dopo il bilanciamento.

Il dataset risulta ora bilanciato, consentendo un'analisi affidabile nelle fasi successive del progetto.

3.2 Preprocessing e tokenizzazione del testo

Una volta bilanciato il dataset, il passo successivo consiste nel preparare il testo delle frasi contenute nella colonna `Sentence` per l'analisi. Questa fase si divide in due parti: preprocessing e tokenizzazione.

3.2.1 Preprocessing del testo

In primo luogo il testo delle frasi viene ripulito applicando una serie di trasformazioni, con l'obiettivo di rimuovere elementi superflui e garantire una rappresentazione uniforme dei dati. Le operazioni principali includono:

- Conversione del testo in minuscolo per eliminare le differenze dovute alla capitalizzazione (`convert_to_lowercase(df)`).
- Rimozione di caratteri speciali, come simboli di punteggiatura e numeri, che non sono utili per l'analisi (`remove_special_characters(df)`).
- Eliminazione di URL (`remove_urls(df)`).

- Rimozione di spazi bianchi extra (`remove_extra_whitespace(df)`).

Il risultato di queste operazioni è una versione pulita del testo, pronta per la fase di tokenizzazione.

3.2.2 Tokenizzazione

In questa fase il testo viene suddiviso in token utilizzando il metodo `word_tokenize` della libreria `nltk`. I token rappresentano le unità linguistiche fondamentali (ad esempio, parole o simboli) in cui il testo è segmentato. In Figura 3.3 è possibile vedere un esempio di tokenizzazione di una frase presa casualmente dal dataset, con i relativi ID.

Tokens	Token IDs
interest	3037
heats	18559
up	2039
for	2005
yahoo	20643
y	1061
##ho	6806
##o	2080
the	1996
wall	2813
street	2395
journal	3485

Figura 3.3: Esempio di tokenizzazione.

Un'illustrazione completa del DataFrame con tutte le fasi di preprocessing e tokenizzazione può essere vista nella Figura 3.4.

	text	sentiment	text_lower	text_clean	text_no_urls	text_stripped	tokens
0	The GeoSolutions technology will leverage Bene...	positive	the geosolutions technology will leverage bene...	the geosolutions technology will leverage bene...	the geosolutions technology will leverage bene...	the geosolutions technology will leverage bene...	[the, geosolutions, technology, will, leverage...
1	$ESI on lows, down $1.50 to $2.50 BK a real po...	negative	$esi on lows, down $1.50 to $2.50 bk a real po...	esi on lows down to bk a real possibility	esi on lows down to bk a real possibility	esi on lows down to bk a real possibility	[esi, on, lows, down, to, bk, a, real, possibi...
2	For the last quarter of 2010 , Componenta 's n...	positive	for the last quarter of 2010 , componenta 's n...	for the last quarter of componenta s net sal...	for the last quarter of componenta s net sal...	for the last quarter of componenta s net sales...	[for, the, last, quarter, of, componenta, s, n...
3	According to the Finnish-Russian Chamber of Co...	neutral	according to the finnish-russian chamber of co...	according to the finnishrussian chamber of com...	according to the finnishrussian chamber of com...	according to the finnishrussian chamber of com...	[according, to, the, finnishrussian, chamber, ...
4	The Swedish buyout firm has sold its remaining...	neutral	the swedish buyout firm has sold its remaining...	the swedish buyout firm has sold its remaining...	the swedish buyout firm has sold its remaining...	the swedish buyout firm has sold its remaining...	[the, swedish, buyout, firm, has, sold, its, r...

Figura 3.4: Visualizzazione del DataFrame durante le diverse fasi di preprocessing e tokenizzazione.

3.2.3 Distribuzione della lunghezza dei token

Dopo il processo di tokenizzazione, è stato analizzato il numero di token generati per ciascuna frase presente nel dataset. Questa analisi permette di comprendere meglio la struttura delle frasi, fornendo informazioni utili per impostare correttamente i parametri per l'addestramento del modello, come la lunghezza massima dei token (`MAX_LEN`).

La distribuzione della lunghezza dei token è mostrata nella Figura 3.5. L'asse orizzontale rappresenta la lunghezza delle frasi in termini di numero di token, mentre l'asse verticale indica la frequenza con cui si verifica ciascuna lunghezza nel dataset.

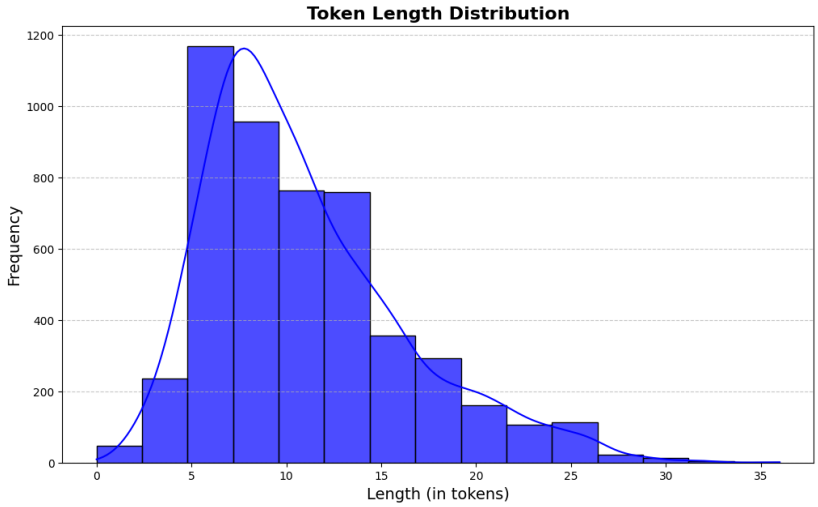


Figura 3.5: Distribuzione della lunghezza delle frasi in token.

Come si può osservare dal grafico, la maggior parte delle frasi ha una lunghezza compresa tra 5 e 15 token, con un picco intorno ai 7-8 token. Inoltre, dall'analisi delle lunghezze dei token, sono stati ottenuti i seguenti risultati:

- Lunghezza massima: 36 token.
- Lunghezza media: 10.89 token.

3.3 Training

Per l'analisi e il confronto delle performance, sono stati selezionati tre modelli pre-addestrati: BERT, RoBERTa e DistilBERT (gli ultimi due sono modelli ottimizzati di BERT). Il training è stato effettuato, in seguito a diverse prove, utilizzando i seguenti parametri:

- Numero di epoche (`num_epochs`): 2
- Dimensione del batch (`batch_size`): 16
- Learning rate (`learning_rate`): 2×10^{-5}

L'obiettivo è valutare come ciascun modello performa sul dataset bilanciato e identificare eventuali differenze nei risultati ottenuti. Durante il training, sono state monitorate due metriche principali: il Training Loss e il Validation Loss. I grafici riportati di seguito illustrano l'andamento di queste due metriche per ciascun modello durante il processo di addestramento.

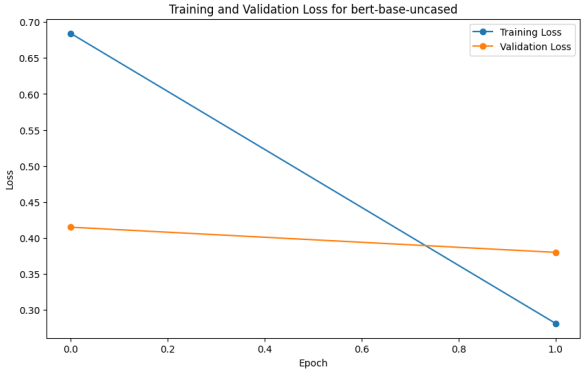


Figura 3.6: Andamento della Training Loss e Validation Loss per il modello BERT.

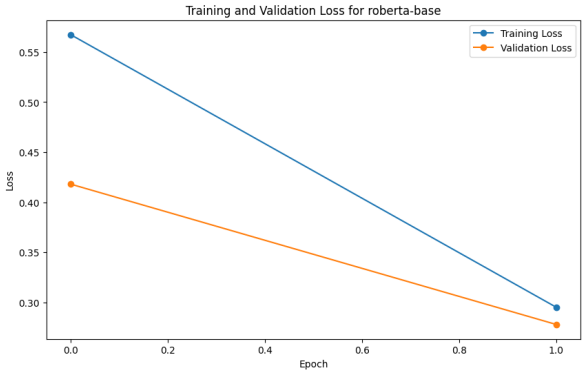


Figura 3.7: Andamento della Training Loss e Validation Loss per il modello RoBERTa.

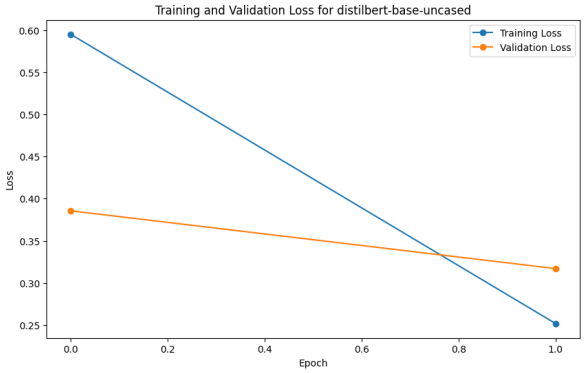
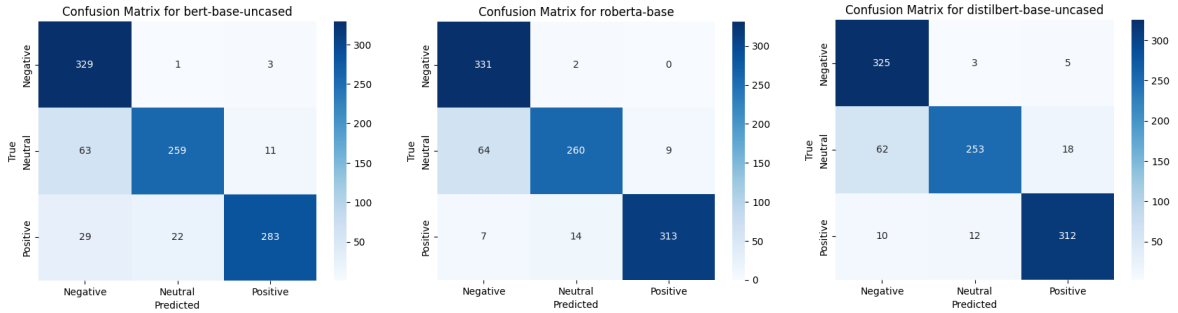


Figura 3.8: Andamento della Training Loss e Validation Loss per il modello DistilBERT.

Dall'osservazione dei grafici, si osserva che RoBERTa potrebbe avere una migliore capacità di equilibrio tra training e validazione rispetto agli altri due modelli.

3.3.1 Valutazione dei modelli

Per valutare le performance dei modelli, sono state generate le confusion matrix basate sui risultati del dataset di test (separato inizialmente da quello di train/validation, secondo un rapporto 80-20). Le confusion matrix forniscono una rappresentazione visiva delle predizioni corrette e degli errori commessi dai modelli per ciascuna classe (`negative`, `neutral`, `positive`). Le confusion matrix per i tre modelli sono riportate nella Figura 3.9.



BERT RoBERTa DistilBERT

Figura 3.9: Confusion Matrix per i modelli BERT, RoBERTa e DistilBERT sul dataset di test.

Di seguito vengono riportate le principali metriche quantitative ottenute durante il test del modello: **precision**, **recall** e **f1-score** di ogni singola classe e **accuracy** del modello con relativa media pesata e media globaleme.

	precision	recall	f1-score
Negative	0.78	0.99	0.87
Neutral	0.92	0.78	0.84
Positive	0.95	0.85	0.90
accuracy			0.87
macro avg	0.88	0.87	0.87
weighted avg	0.88	0.87	0.87

Figura 3.10: BERT

	precision	recall	f1-score
Negative	0.82	0.99	0.90
Neutral	0.94	0.78	0.85
Positive	0.97	0.94	0.95
accuracy			0.90
macro avg	0.91	0.90	0.90
weighted avg	0.91	0.90	0.90

Figura 3.11: RoBERTa

	precision	recall	f1-score
Negative	0.82	0.98	0.89
Neutral	0.94	0.76	0.84
Positive	0.93	0.93	0.93
accuracy			0.89
macro avg	0.90	0.89	0.89
weighted avg	0.90	0.89	0.89

Figura 3.12: DistilBERT