UNIVERSITÀ POLITECNICA DELLE MARCHE

FACOLTÀ DI INGEGNERIA

Corso di Laurea Magistrale in
Ingegneria Informatica e dell'Automazione

Corso di Data Science

Realizzazione di un classificatore per un dataset di recensioni di film tramite architettura BERT

Studenti:

Angelone Mattia Domenico
Romanelli Marco
Giannelli Edoardo
Ali Waqar Badar

Anno Accademico 2022-2023

Bert

Sommario

1. Introduzione

1.1 Il Natural Language Processing

Il Natural Language Processing (NLP) è un ramo della data science che si concentra sulla capacità dei calcolatori di elaborare e comprendere il linguaggio naturale, al fine di interpretare, comprendere e comunicare l'input testuale in modo efficace.
Il NLP utilizza tecniche di Intelligenza Artificiale, Machine Learning e Deep Learning per elaborare i dati, che possono essere forniti in forma di testo, suoni, immagini e video.

Il NLP ha diverse applicazioni, tra cui la traduzione automatica di testi, il riassunto automatico di testi, l'identificazione di mail spam, la creazione di chatbot e question answering e la sentiment analysis.
Grazie al NLP, è possibile automatizzare molti processi che altrimenti richiederebbero una grande quantità di tempo e risorse umane.

La traduzione automatica di testi è una delle applicazioni più conosciute del NLP. Con l'aiuto di tecniche di elaborazione del linguaggio naturale, i sistemi di traduzione automatica possono tradurre testi in diverse lingue in modo rapido ed efficace. Inoltre, il riassunto automatico di testi può essere utile per estrarre le informazioni più importanti da documenti o articoli di lunghezza considerevole, aiutando gli utenti a risparmiare tempo nella lettura.

Può anche essere utilizzato per identificare mail spam, cioè messaggi indesiderati che intasano le caselle di posta elettronica.
Attraverso l'utilizzo di tecniche di classificazione, il NLP può identificare i messaggi spam e spostarli in una cartella separata, riducendo cos`ı il tempo necessario per gestire la posta elettronica.

Un'altra importante applicazione del NLP è la creazione di chatbot e question answering, che permettono agli utenti di interagire con i sistemi informatici in modo naturale, utilizzando il linguaggio naturale.

I chatbot sono sempre più utilizzati in diversi contesti, come ad esempio nel settore del customer service, dove possono aiutare a rispondere alle domande dei clienti in modo efficace e veloce.

Infine, può essere utilizzato per analizzare il sentiment, cioè la valutazione emotiva associata a un particolare testo.
L'analisi del sentiment può essere utile in molti contesti, come ad esempio per monitorare l'opinione pubblica su un prodotto o un'azienda, o per analizzare i dati dei social media al fine di individuare trend e comportamenti degli utenti.

1.2 Bert

BERT, acronimo di Bidirectional Encoder Representations from Transformers, è uno dei modelli più avanzati di deep learning basati sulla tecnologia dei transformer. Questo modello si differenzia dai precedenti poichè ogni elemento di output è collegato a ogni elemento di input, e i pesi tra di essi sono calcolati dinamicamente sulla base della loro connessione. Questa caratteristica rende BERT

un modello molto potente e adatto per diverse applicazioni nel campo del Natural Language Processing (NLP).

Uno degli aspetti più interessanti di BERT è la sua capacità di leggere il contesto in entrambe le direzioni contemporaneamente. Questa funzionalità permette al modello di cogliere il significato di una frase o di un intero documento in modo più preciso e completo rispetto ai modelli precedenti.

BERT è stato pre-addestrato su due task di NLP diversi, ma correlati, ovvero il Masked Language Modeling e il Next Sentence Prediction.

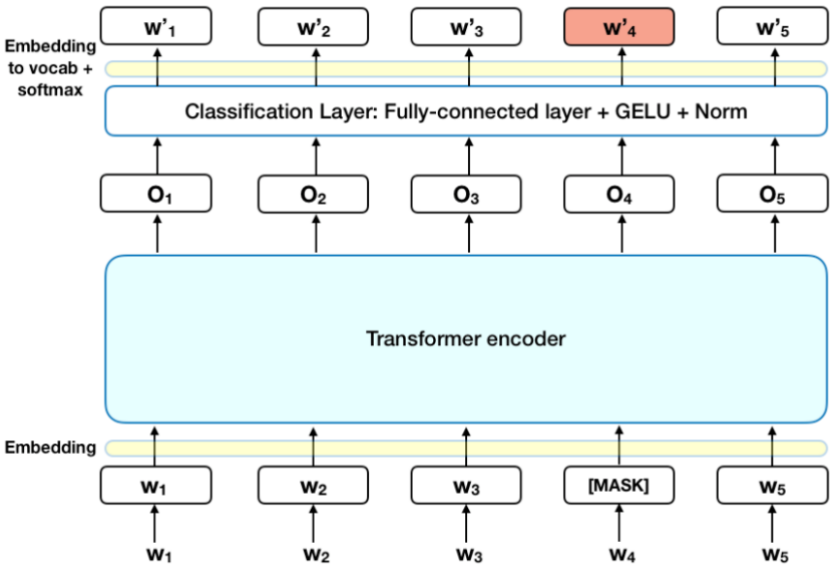

Figura 1 BERT model

Il primo obiettivo del Masked Language Model è quello di nascondere una parola all'interno di una frase e far s`ı che il programma preveda quale parola è stata nascosta (mascherata) sulla base del contesto circostante. Questo processo aiuta il modello a comprendere la relazione tra le parole all'interno di una frase e a costruire una rappresentazione semantica piu` accurata.

Il secondo obiettivo di BERT è il Next Sentence Prediction. In questo caso, il programma deve prevedere se due frasi date hanno una connessione logica e sequenziale, oppure se la loro relazione è casuale. Questa funzionalità aiuta il modello a comprendere la struttura del testo e la relazione tra le frasi, un aspetto importante per molte applicazioni di NLP.

2 Descrizione ed elaborazione del dataset

Il dataset utilizzato contiene 50.000 recensioni di film per l'elaborazione del linguaggio naturale o l'analisi del testo ed è reperibile al seguente link: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews .

Il file è un csv che presenta due colonne:

- Review: la prima colonna che contiene, riga per riga, tutte le recensioni;
- Sentiment: la colonna che associa a ciascuna review un sentiment indicato come 'positive' o 'negative'.

Dopo una rapida esplorazione del dataset, il campo Review è stato filtrato rimuovendo hyperlinks, emoji, stopwords e i caratteri sono stati convertiti in caratteri minuscoli.

Successivamente, visto che il campo Sentiment contiene al suo interno variabili categoriche, quali Positive e Negative, queste sono state trasformate in variabili binarie, con valori 1 per Positive e 0 per Negative.

Infine, a causa della grandezza del dataset che ha portato a problemi di overflow e di esecuzione in locale, al fine di mantenere invariata la grandezza del dataset di 50mila righe abbiamo deciso di eseguirlo attraverso Google Colab.

Il dataset, alla fine della fase di pre-processing assume il seguente aspetto:

	review	sentiment
36457	Come on Tina Fey you can do better then this. ...	0
48521	This is a very beautiful and almost meditative...	1
5586	What an embarassment...This doesnt do justice ...	0
40152	To begin with, I really love Lucy. Her TV show...	0
9437	I haven't seen this film in years so my knowle...	1
...	...	...
16180	I can understand how fans of filmmaker Roman P...	1
46643	Rita Hayworth lights up the screen in this fun...	1
43178	The operative rule in the making of this film ...	0
17408	I expected this to be a lot better. I love Tim...	0
42625	First of all, I have to say that I am not gene...	0

50000 rows × 2 columns

Figura 2 Dataset

3 Il modello

Il modello è stato addestrato suddividendo il dataset in:

- training e validation: rappresenta l'input per l'addestramento del modello e contiene il 90% dei dati, ovvero 5.400 valori. Di questi, l'80% (17292) sono stati usati per la fase di addestramento del modello ed il restante 10% (4323) per la fase di validazione;
- test: rappresenta l'input per la fase di test del modello e contiene il restante 10% dei dati.

3.1 Architettura utilizzata

BertForSequenceClassification e il modello di partenza, sul quale viene effettuato fine- tuning, ed è composto da:

- bert-base-uncased: modello BERT composto da 12 layer, 768 nodi nascosti, 12 attention heads e 110 milioni di parametri;
- uno strato di classificazione a due classi a valle di bert-base-uncased. Il tokenizer usato è BertTokenizer ed è basato su bert-base-uncased.

3.2 Fase di addestramento

Per l'addestramento del modello sono stati impostati i seguenti parametri:

- batch: 16;
- epoche: 4;
- ottimizzatore: Adam;
- learning rate: $2 \times 10^{-}5$;
- epsilon: $1 \times 10^{-}8$.

Nell'immagine di seguito viene mostrato l'andamento dei valori di loss e di accuracy per la fase di train e validation nel corso delle epoche di addestramento.

epoch	Training Loss	Valid. Loss	Valid. Accur.	Training Time	Validation Time
1	0.380019	0.303208	0.876845	0:06:27	0:00:33
2	0.208087	0.401484	0.874539	0:06:31	0:00:33
3	0.108730	0.498803	0.887454	0:06:31	0:00:33
4	0.055709	0.552910	0.887685	0:06:32	0:00:33

Figura 3 Andamento epoch

Graficamente viene anche mostrato l'andamento dei valori di loss e accuracy per entrambe le fasi di training e validation:

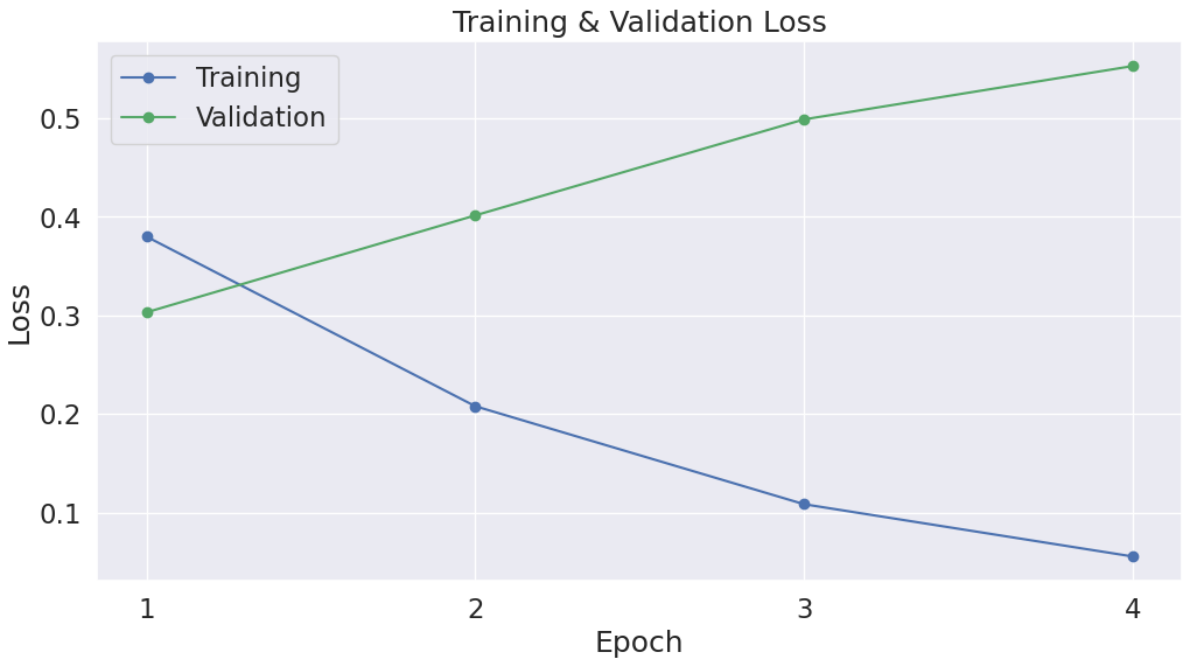

Figura 4 Andamento del Training e Validation Loss

Come si può osservare dalla figura 3.2, il numero ideale di epoche per l'addestramento, per non incorrere in situazioni di overfitting, è pari a 1.

4 Metriche

Di seguito le metriche di valutazione utilizzate per il modello.

4.1 Accuracy, Precision, Recall e F1 Score

Per la valutazione del modello sono state prese in considerazione le quattro metriche principali quali Accuracy, Precision, Recall e F1 Score.

L'Accuracy indica la percentuale di elementi etichettati correttamente rispetto al totale degli elementi e viene calcolata come segue:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

La Precision indica la percentuale di elementi classificati correttamente con classe C_i rispetto al totale degli elementi classificati con C_i e viene calcolata come segue:

$$Precision(C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)}$$

La Recall valuta la percentuale di elementi classificati correttamente con classe C_i rispetto al totale degli elementi realmente di classe C_i e viene calcolata come segue:

$$Recall(C_i) = \frac{TP(C_i)}{TP(C_i) + FN(C_i)}$$

La F1 Score mette in relazione Precision e Recall e corrisponde alla media armonica tra le due, viene calcolata come segue:

$$F1(C_i) = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Nella seguente tabella vengono mostrate le metriche ottenute.

Metrica	Valore
Accuracy	0.8950
Precision	0.8884
Recall	0.9075
F1	0.8860

4.2 Confusion Matrix

La matrice di confusione è uno strumento grafico usato per valutare le performance di un modello di classificazione binaria.
Ogni colonna della matrice rappresenta i valori predetti, Predicted label, mentre ogni riga rappresenta i valori reali, True label. L'elemento sulla riga i e sulla colonna j è il numero di casi in cui il classificatore ha classificato la classe "vera" i come classe j.
Sulla diagonale principale sono presenti gli elementi classificati correttamente.

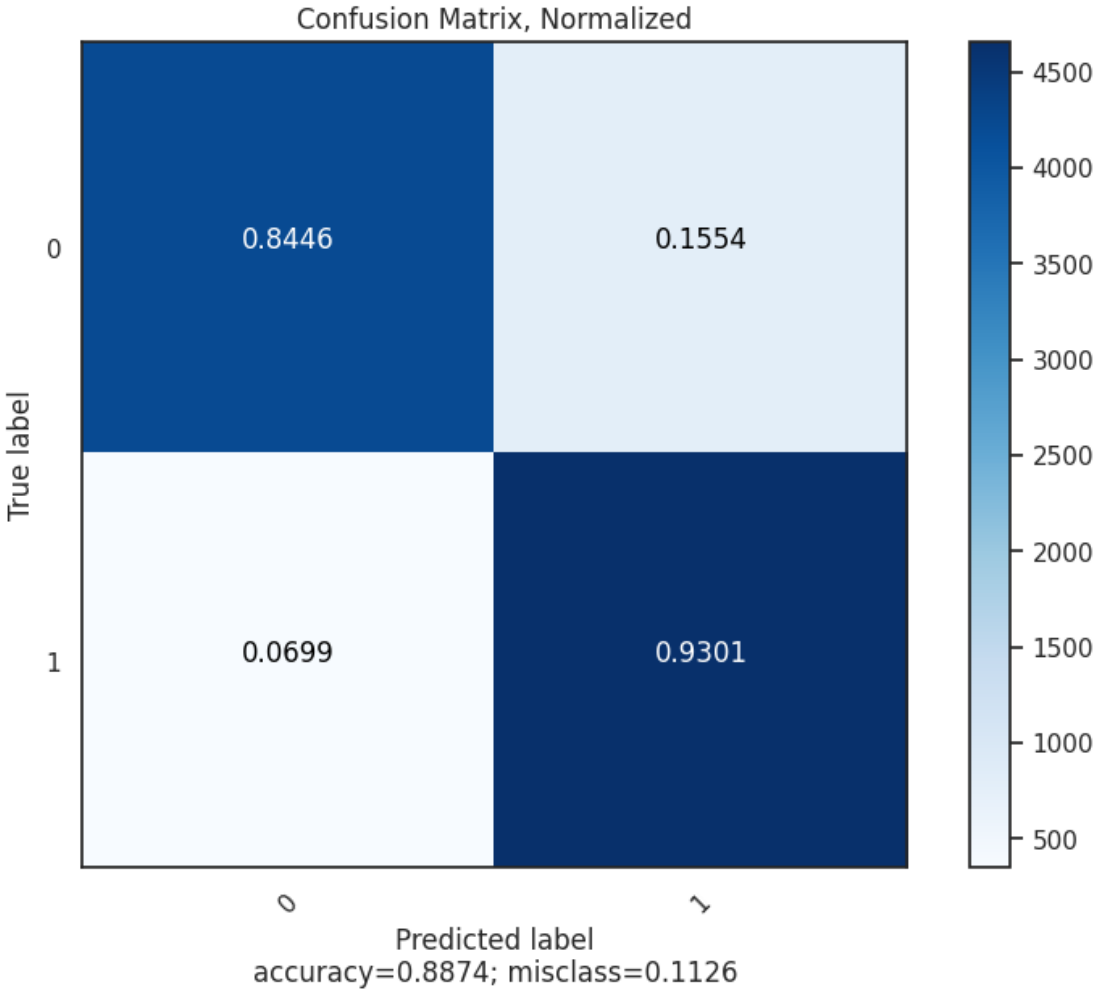

Figura 5 Matrice di confusione