Immagine che contiene testo

Descrizione generata automaticamente

Corso di Laurea Magistrale in Ingegneria Informatica

Analisi di Social Network e Media

Progetto:

BERT Sentiment Analysis

Ivan Scuderi 216635

Anno Accademico 2021-2022

**1: Scopo dell’analisi**

L’obiettivo del seguente progetto didattico è quello di svolgere e risolvere dei task di *sentiment analysis* mediante l’impiego dell’architettura *transformers* **BERT**, in particolare si impiegheranno *dataset* testuali di varia natura e forma al fine di individuare il contesto, o caratteristiche dei dati, su cui si ottengono le performance migliori, andando anche ad indagarne le motivazioni.

Come detto in precedenza si impiegheranno differenti *dataset*, su cui si andranno a costruire modelli di reti neurali al fine di risolvere sostanzialmente dei problemi di classificazione binaria o multi-classe. Infatti la scelta delle collezioni impiegate è stata guidata anche dalla numerosità del relativo insieme delle etichette, oltre che ovviamente dalle caratteristiche e differenze tra i contesti rappresentati dai vari documenti.

Un ulteriore parametro di scelta è stato anche il numero di parole contenute nei vari record, questo al fine di analizzare se, ed in che misura, la presenza di testi più lunghi in alcune collezioni vada a favorire le performance di **BERT** dal punto di vista dell’ottenimento di vettori di *embedding* maggiormente correlati al contesto.

Di seguito si procederà andando ad analizzare singolarmente i vari *dataset* descrivendone il perché della scelta e le caratteristiche specifiche del contesto, andando poi ad esplorarli maggiormente nel dettaglio insieme alle operazioni di pulizia e *preprocessing* effettuare per rendere l’input compatibile con l’architettura che si intende impiegare.

**2: Dataset Selezionati e Caratteristiche**

Sono stati scelti per lo sviluppo del seguente lavoro progettuale tre collezioni testuali:

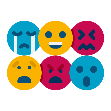
1. **Twitter Dataset**

Collezione di *tweet* estratti dalla piattaforma di social network Twitter che possono essere categorizzati come testi appartenenti al contesto di ‘*social life*’, in particolare il tema ricorrente in tali post fa riferimento alla pandemia di Corona Virus, inoltre sono presenti tre etichette di classe.

1. **Immagine che contiene testo, clipart

   Descrizione generata automaticamenteIMDB Dataset**

Insieme di recensioni di film o serie tv caricate sul noto portale web IMDb, il contesto specifico risulta quindi essere quello cinematografico, inoltre sono presenti solo due classi e si può notare come la lunghezza dei testi sia nettamente superiore rispetto alle altre collezioni.

1. **Emotion Dataset**

Insieme di frasi etichettate con la sentiment vera e propria che queste esprimono, i testi di tale collezione sono molto più semplici rispetto alle altre ma, nonostante ciò, è stato selezionato proprio per via della numerosità delle sue etichette di classe che risultano essere sei.

In generale le operazioni di pulizia che sono state effettuate su tali dataset sono le medesime: sono stati rimossi caratteri codificati non nello standard *ASCII/UTF-8*, sono state rimosse eventuali *emoji*, *link*, indirizzi *e-mail*, *hashtag*, nomi utenti, eventuali costrutti dei linguaggi di *markup*, sono state accorpare le parole con apostrofo, sono stati eliminati i numeri presenti singolarmente (non collegati a parole del testo) ed infine ogni record è stato portato in lettera minuscola. In particolare solo per l’**Emotion Dataset** tali operazioni sono risultate pressoché inutili, poiché si trovava già di partenza in uno stato molto buono dal punto di vista della pulizia e qualità dei dati.

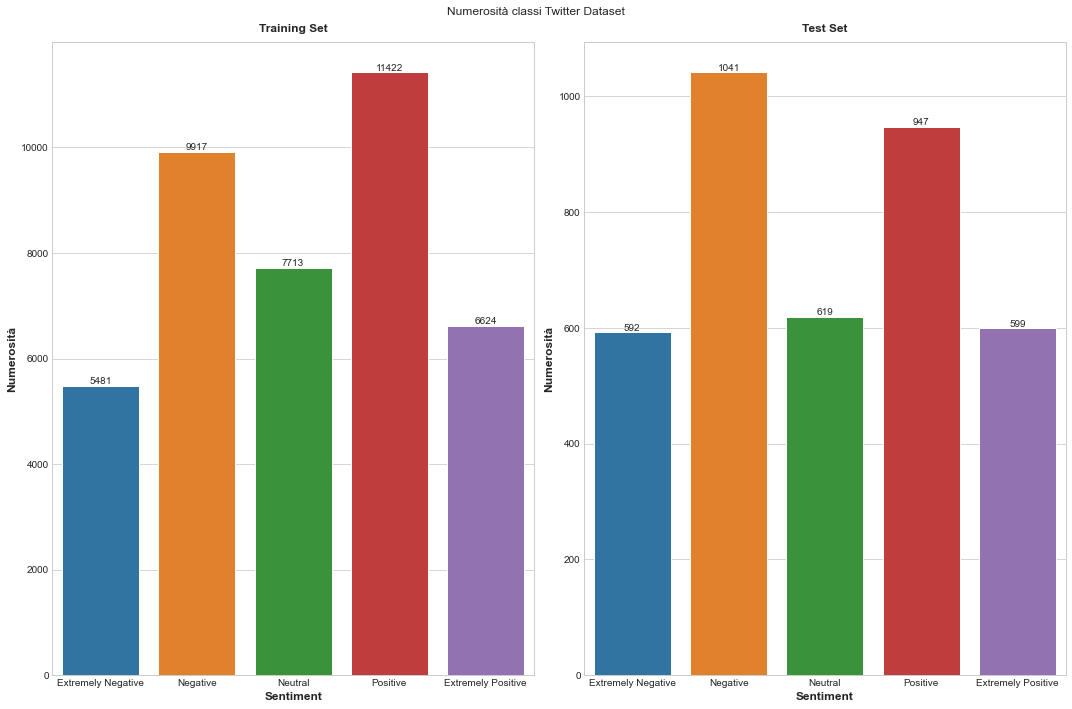
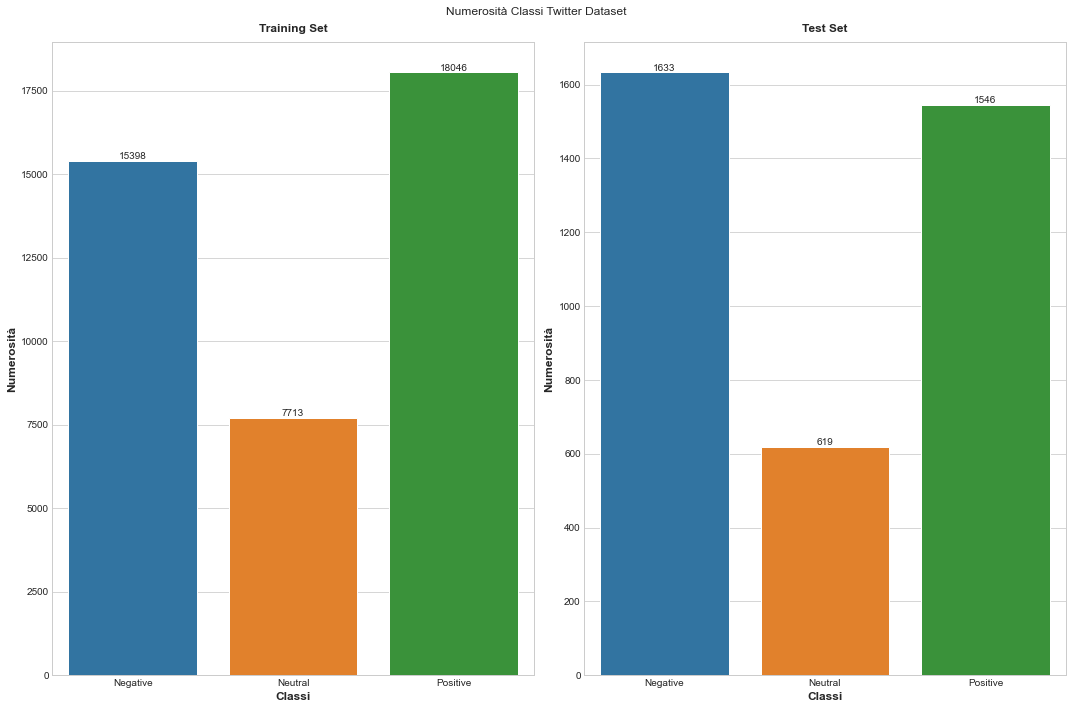
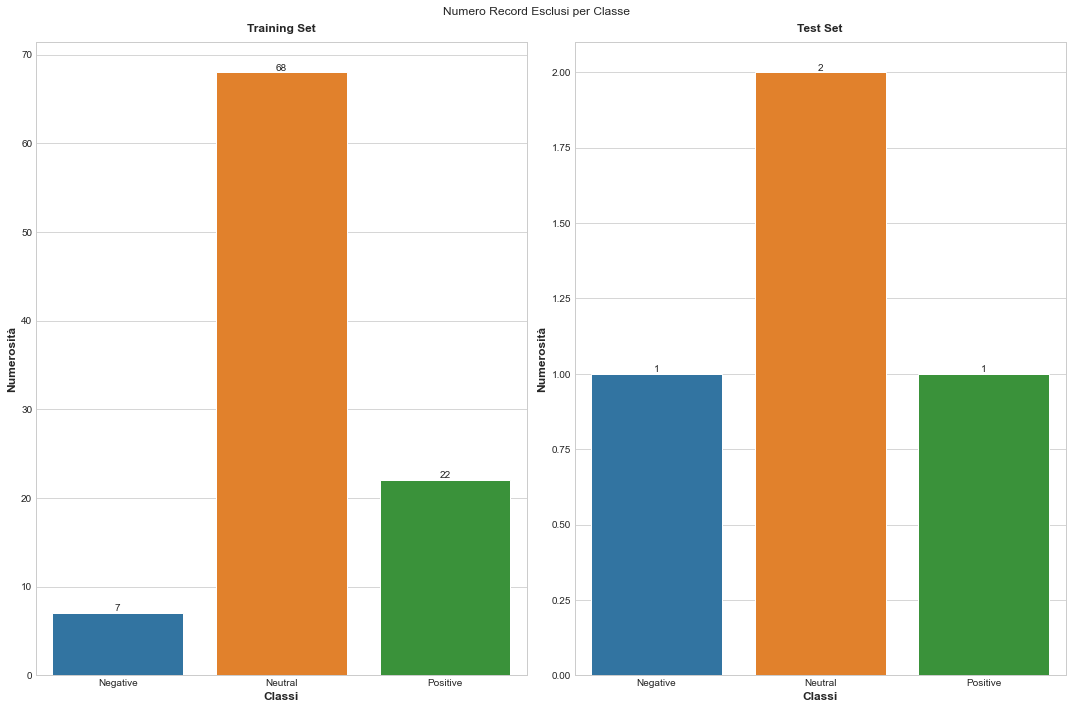
Da sottolineare il fatto che per l’analisi condotta scelta progettuale è stata quella di non rimuovere eventuali *stopwords* nel testo (a meno di alcuni grafici visualizzati in cui non sono state considerate), e non sono state nemmeno implementate le operazioni di *stemming* e *lemmatization*. Questo al fine di evitare di perdere informazioni di contesto che potrebbero portare ad un impoverimento delle performance di **BERT**. Inoltre le etichette di classe sono state trattate mediante l’utilizzo della strategia *OneHotEncoding*, questo perché non essendoci una relazione d’ordine l’obiettivo è quello di mantenerle scorrelate agli occhi del modello.

Si procede andando ad analizzare singolarmente i vari *dataset*.

**2.1 Twitter Dataset**

Per il seguente *dataset* è stato possibile ottenere sia l’insieme di *train* che di *test*, inizialmente questi comprendevano rispettivamente 41'157 e 3'798 elementi che sono diventati 40'847 e 3'786 dopo le operazioni di eliminazione di valori nulli, valori duplicati e record che non raggiungevano la soglia di lunghezza pari a 4.

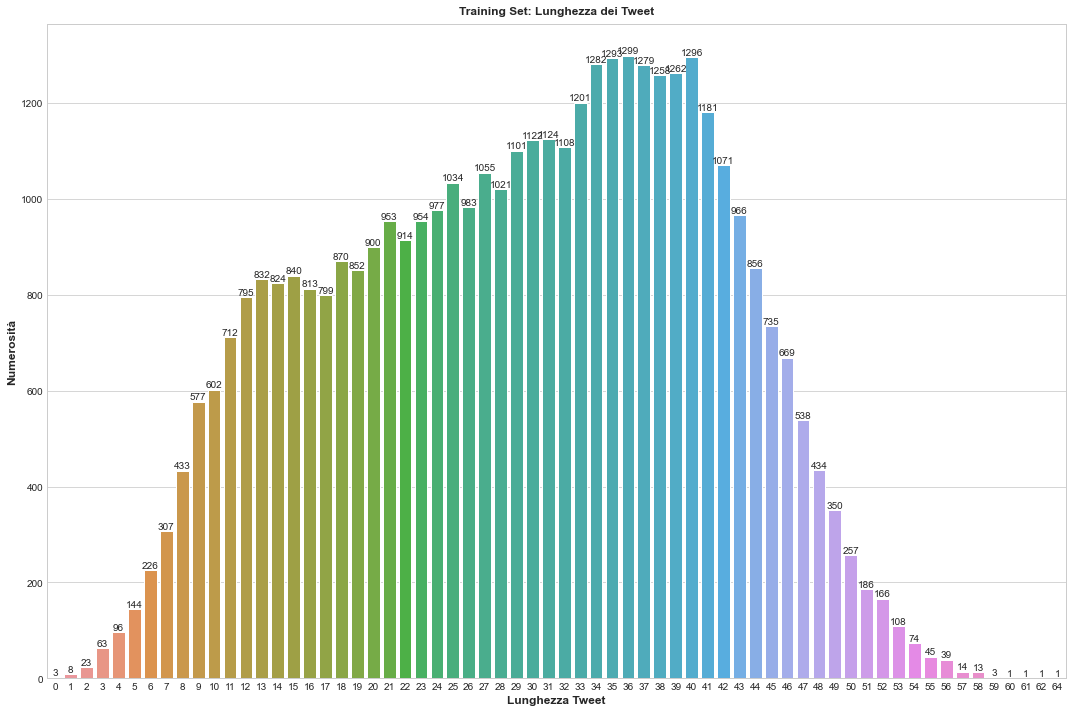
Erano presenti 5 etichette di classe, sono state successivamente portate a 3 dopo l’accorpamento delle classi ‘*border line’*, le risultanti sono: ‘*Positive’*, ‘*Negative’* e ‘*Neutral’.* Si può facilmente notare come, nonostante l’operazione di accorpamento, sia presente comunque sbilanciamento nella numerosità delle varie *label*.

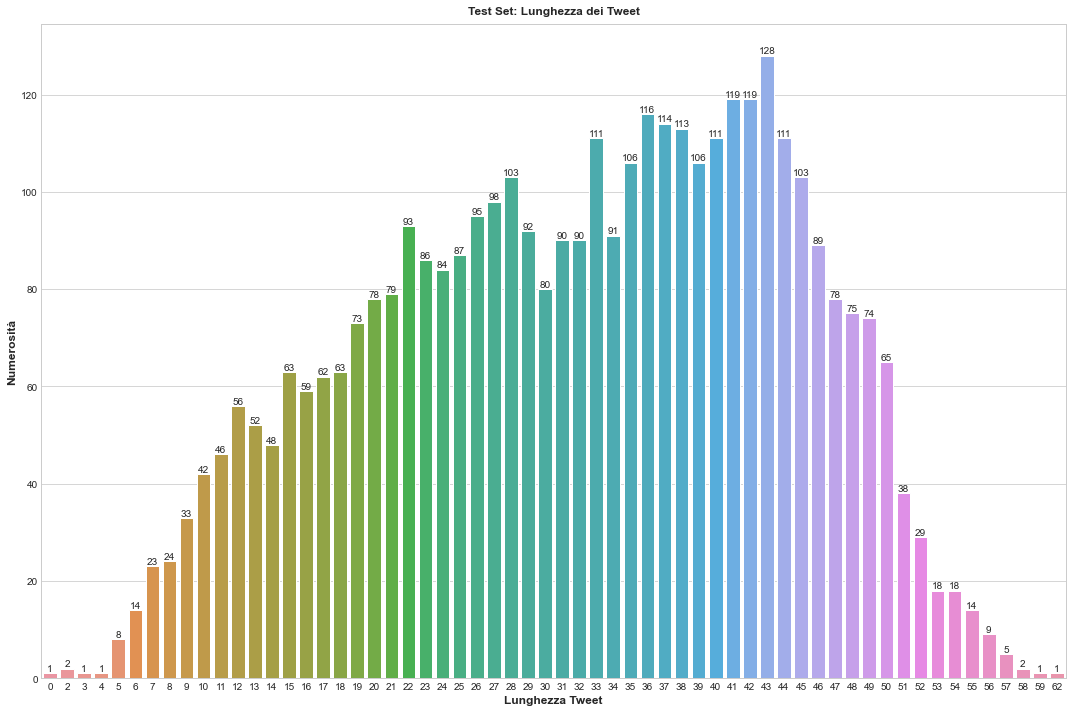
Come detto in precedenza all’interno del dataset vi erano record che, a causa delle operazioni sopracitate di pulizia, sono risultati essere costituiti da pochissimi elementi o addirittura zero; scelta progettuale è statata quella di rimuovere tali elementi che risultano avere pochissimo valore dal punto di vista semantico e contenutistico, inoltre come ci si poteva aspettare la maggior parte di tali elementi provengono dalla classe ‘**Neutral’**.

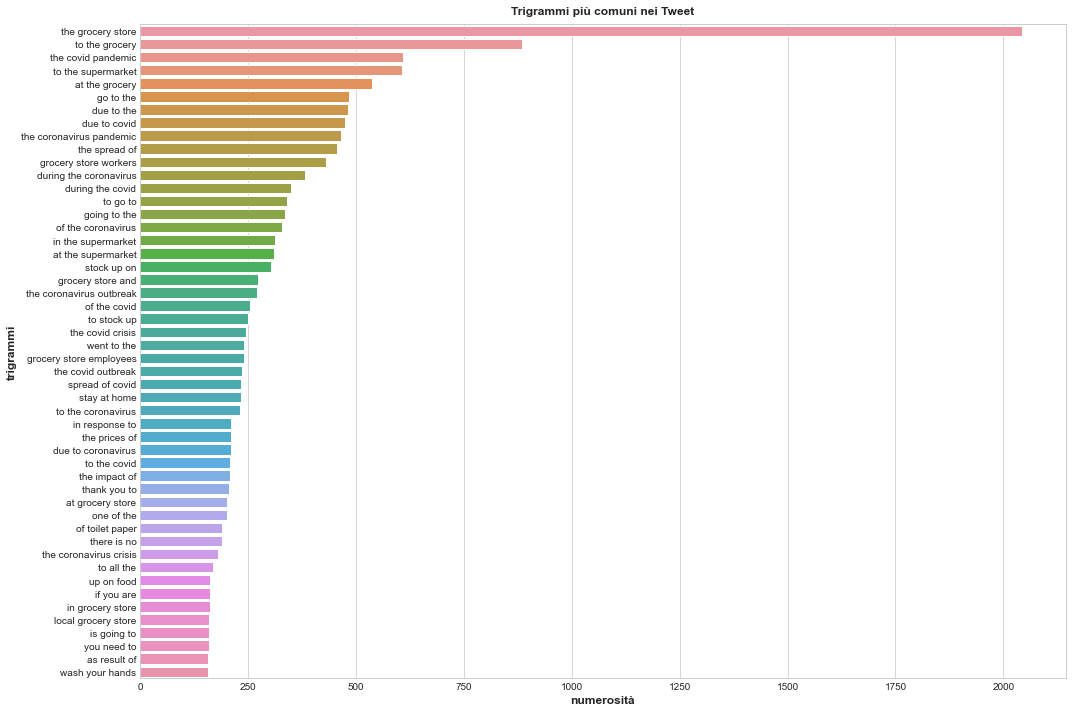
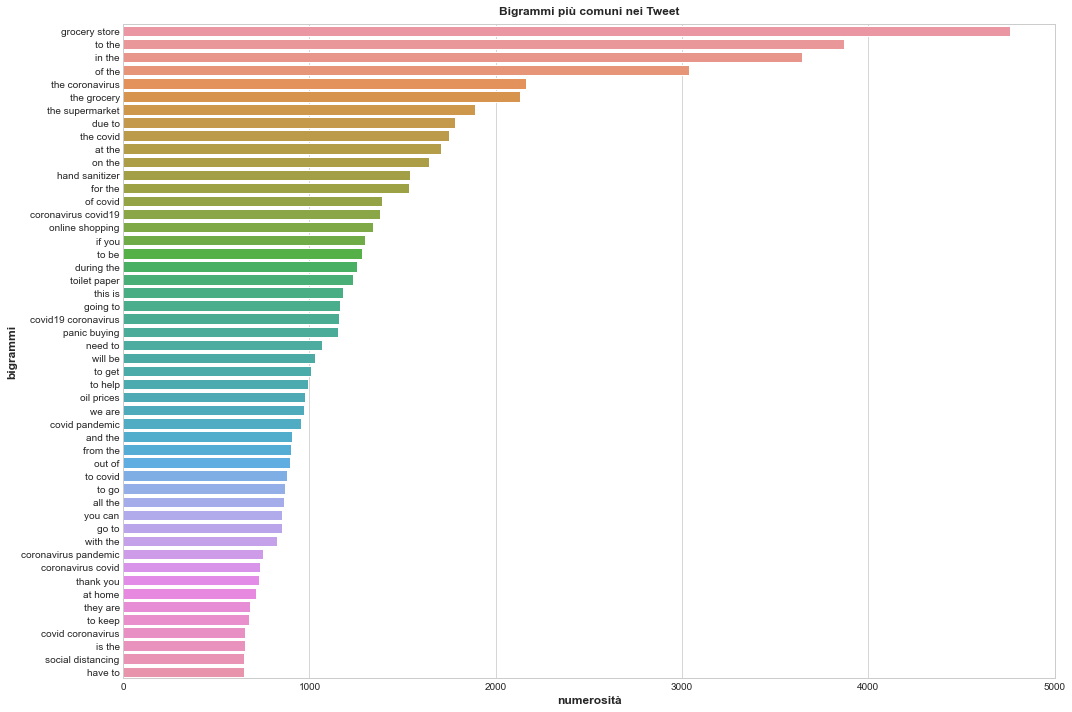
Al contrario nei casi normali si trovano mediamente *tweet* con una lunghezza pari a circa 29, comunque la maggior parte delle entry non risulta avere più di 40 parole, inoltre tali valori risultano pressocchè simili a prescindere se si considera il *training* o il *test set*.

Immagine che contiene tavolo

Descrizione generata automaticamente





Immagine che contiene testo, lavagnabianca

Descrizione generata automaticamenteInfine l’analisi delle *wordcloud* contenti le parole maggiormente frequenti divise per classe, insieme alla visualizzazione delle coppie e triplette di termini più frequenti, sottolinea ancora una volta quale sia il contesto specifico di tale insieme di dati. I termini più frequenti sono infatti ‘*Covid’* e ‘*Coronavirus’*, insieme ad ulteriori parole che molto probabilmente all’inizio corrispondevano ad *hashtag* impiegati dagli utenti.

Le operazioni di *preprocessing* sono terminate andando a costruire il validation set mediante un’estrazione casuale del 10% delle *entry* del *training set*. Inoltre è stato applicato ai vari insiemi il **BertTokenizerFast**, oggetto che permette di ottenere per ogni record testuale un primo vettore di numeri nella forma accettata dall’architettura encoder **BERT** utilizzata nel progetto. In particolare sono stati impiegati due vettori per ogni insieme, ossia: il vettore di **input\_id**, che rappresenta l’indice della parola nel corpus utilizzato per l’addestramento delle reti citate precedentemente, ed il vettore binario di **attention\_mask**, il cui scopo è quello di identificare al layer di embedding quali posizioni dell’input corrispondono ad elementi aggiunti per via dell’operazione di *padding* (nel caso in cui il testo non raggiunga la lunghezza massima impiegata per standardizzate la dimensione dei record) .

In particolare per tale fonti dati è stata impiegato 128 come valore di lunghezza massima, ottenendo un risultato della forma seguente:

Immagine che contiene tavolo

Descrizione generata automaticamente

**2.1 IMDB Dataset**

Per tale dataset