Sentiment Analysis using DistilBERT

Tahnia Sabah
Computer and Information Science
University of Strathclyde
Glasgow, United Kingdom
tahnia.sabah.2021@uni.strath.ac.uk

Abstract – The use of social media has become omni present in our everyday life since the last decade. Twitter, one of the most popular social networking websites is becoming the go to site for sharing opinions for millions of people every day. The users express their sentiments and points of view by tweets regarding different topics such as, politics, religion, current trends, commercial products, etc. Sentiment analysis is an effective way that allows the monitoring of changes in opinion in social media towards recent event, entities, services, trends etc. This is important information for any personnel, company or organization working with decision making, legislation, regulation and maintenance of anything that is related to the people. Research on Twitter sentiment analysis which analyzes Twitter data or 'tweets' to extract people's sentiments about a topic has grown rapidly in the recent years.

Sentiments are generally divided into three categories: positive, negative, and neutral, we are using textblob library to categorize Christmas tweets based on its sentiment. We are training our model with DistilBERT for sentiment predictions for each tweet. We are using twitter API to fetch 1000 tweets to train the model for the simplicity of implementation. The model has reached an impressive accuracy of 84%.

Keywords – DistilBERT, Sentiment, Textblob, NLP

I. INTRODUCTION

With the rapid growth of social networking sites, these sites have become the largest web destination for people to express their raw unfiltered opinion. Twitter users are posting hundreds of tweets every day to discuss or express their opinions on various topics. Twitter generates vast amounts of information that makes twitter a rich source of data that can be used in uncountable way.

In recent years, research on social data for the sentiment analysis of people's opinions on a product, topic, or event has increased and has proved to be very useful. Sentiment analysis, also known as opinion mining, is an important application of natural language that is becoming more popular with the use of extensive pretrained language models. This process determines the sentiment orientation of a text as positive, negative, or neutral.

Currently, there are many transformer-based language models available like, Megatron – Nvidia, BERT – Google Research, GPT-2 – OpenAi and many more.

Sentiment analysis is used in fields like social media analysis since sentiments are one of the most crucial properties to determine the person's behavior.

Sentiment Analysis is a technique that deals with opinions, feelings, feedbacks, emotions along with perception attached to the data. It is used to determine whether a text is positive, negative or neutral. In text analytics, natural language processing (NLP) and machine learning (ML) techniques are used to label sentiment scores to the topics, categories or entities within a phrase. This field is of interest to both researcher and industry communities because is holds raw unfiltered mass data.

II. METHODOLOGY

A. DATASET

The dataset used in this paper is a real time dataset which is been generated from the Twitter API to fetch the tweets about a particular query. Here in this paper, we are generating a dataset which contain 1000 rows and 11 columns. The data also contain many special symbols and different emoji which needs to be treated during spreprocessing. Before using it to analyze the sentiment. This process of Data Preprocessing is known as Data Cleaning.

B. TRANSFER LEARNING

In this paper we are using the pre trained model DistilBERT which was reduced by 40% from the original BERT model which was created by google. This pre-trained model is a type of Transformer model and is fast and small when compared to the original model. This model runs 60% faster, and it also preserves 95% of the performance of BERT model. This model can be loaded from the Transformer python library, and it also contains the tokenizer which helps us to tokenize the text in the required format by adding few of the special parameters.

The Idea of DistilBERT is when a large model is trained with huge networks its full output distribution can approximated from the smaller network.

C. MODEL

The DistilBERT Model contains of 6 different layers and there are about 758 hidden layers which contain 66 million parameters.

D. PROPOSED APPROACH

In this paper, we are using the tweepy and the textblob python library. The tweepy library makes it easy to access the tweets from the twitter by passing on the API Key and the API Secret Key on to the Authentication Handler. The fetched tweets are in the form of Dictionary. So, this makes it easy to fetch the tweet and store it on to a Dataframe.

Using the TextBlob we get the polarity of each tweet like Greater than 0 being a positive sentiment, equal to 0 being a neutral sentiment and less than 0 being a negative sentiment.

From the Transformer Library we are calling two function one is a tokenizer and the other for Sequence Classification. And we are splitting the Data in train and test and here we are splitting 30% of the data for testing and the remaining 70% data for training.

We are then using the Tokenizer from the transformer library and tokenizing on the tweets on both train and test data. We are tokenizing it with a max length of 64. After the tokenizing is complete, we get a Numpy ndarray which is the required format for the input to the DistilBERT model. We also need to reshape the labels onto the required format.

The optimizer used here to train the DistilBERT is Adam, and the learning rate is 0.01. The loss function used here is mean squared error. We achieved an accuracy of 84.14% using just 3 epochs on the training set.

E. LOSS FUNCTION

The Loss Function used in this paper is mean squared error which is usually used for regression problems.

This loss function is very much sensitive towards outliers.

$$L(y, y') = \frac{1}{N}\sum_{i=0}^{N}(y - y'_i)^2$$

III. Results

In this paper, we obtained a result of 82.14% with 20 epochs. During training the learning rate was set 0.01. In the below fig 1, it depicts the graph of Accuracy vs Epochs. We can see that after few epochs the accuracy is constant.

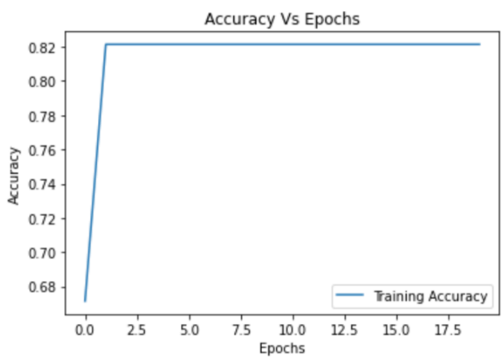

Fig 1 – Accuracy vs Epochs

In the below fig 2, it depicts the graph of Loss vs Epochs. The loss is not constant there is a slight variation in the loss.

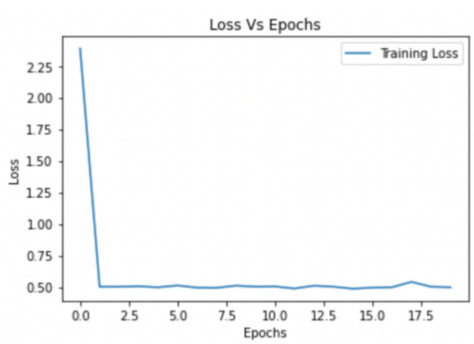

Fig 2 – Loss vs Epochs

IV. COMPARATIVE STUDY

In Comparison with the paper in Reference [2], the approach used is the same but they have used Lemmatization technique and Word Tokenization for Data Preprocessing, the accuracy obtained in this paper is around 63% and very much less when compared to that of my model which got an accuracy of 82.14%. And from the paper in reference [4] we can that see that the accuracy obtained was 93.6%. and the evaluation loss was found to be 39.5%.

	BERT	DISTILBERT
Size (millions)	Base:110 Large:340	Base:66
Performance	Outperform	3% degradation from BERT
Data	16gb BERT data	16gb BERT data
Method	BERT	BERT Distillation

From the Above table we can compare that DistilBERT is significantly faster and there is only 3% loss when compared to BERT's performance. And in DistilBERT the actual BERT is distilled.

V. DISCUSSION AND FUTURE WORK

In the future, I plan to use other variations of BERT, ELMo Variations, Open AI GPT Series and add detailed classification of text to address more complex problems related to Sentiment Analysis using NLP.

VI. REFERENCES

[1] Twitter Sentiment Analysis Based on Ordinal Regression Shihab Elbagir and Jing Yang

[2] Detecting Topic and Sentiment Dynamics Due to COVID-19 Pandemic Using Social Media Hui Yin, Shuiqiao Yang , and Jianxin Li
[3] New Clustering Algorithms for Twitter Sentiment Analysis Hajar Rehioui and Abdellah Idrissi
[4] Sentiment Analysis of News Headlines Using Simple Transformers