Sentiment Analysis of StockTwit Data

Minu Sarraf
University of Rochester
msarraf@ur.rochester.edu

Babu Aman Rai Saxena
University of Rochester
bsaxena@ur.rochester.edu

Abstract

Sentiment Analysis has been studied a lot after the use of Deep Learning methodology in the field of Natural Language Processing as it has better ability to make prediction on the sequential data [6]. Stock market is a place where all financial stock related transaction is made. We propose a Sentiment Analysis model for the stock market which takes financial data from the social media platform named stocktwits and displays the sentiment of the market, The sentiment of the market is referred as "bearish" and "bullish", here "bearish" means that the market is down and "bullish" means the market is up, the word bullish and bearish had been derived from the stock market terms bear and bull which define the stock market. In this proposed model we are using LSTM (Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers) to predict the stock market sentiment, and evaluate their accuracies on the financial Stocktwits data.

1. Introduction

Information gathering has turned out to be an integral part of assessing people's behaviors and actions. The Internet is used as a learning site for sharing and exchanging ideas. People can actively give their reviews and recommendations for variety of products and services using popular social sites and personal blogs. Social networking sites, including Twitter, Facebook, and Instagram, are examples of the sites used to share opinion. The stock market is an essential area of the economy and plays a significant role in trade and industry development. Predicting Stock Market movements is a well-known and area of interest to researchers. Financial news stories are thought to have an impact on the return of stock trend prices and many data mining techniques are used address fluctuations in the Stock Market. Sentiment analysis has attracted significant attention with the increase in the social media platforms. In the current world things are changing rapidly, more importantly stock market is just as unpredictable as it was before, with digital revolution the stock market is more accessible to a majority of the people as trading can be done online. Online trading is beneficial for a lot of people and knowing the news and new topics in the market is also important Stocktwits is a social media platform where posts regarding finance(trading) and stock market are made. On this particular platform many Top-level management people and investors are present and tweets and posts regarding stocks from other social media platform like twitter are shown on the stocktwits platform this makes the stocktwits data reliable for financial sentiment analysis. Analyzing Sentiments of stock market is important because it will be helpful for anyone who intends to invest in the stock market as knowing the market is up or down will help him calculate the flow of stock market which will help him to make stock related decisions. Whereas to do this manually we need to go posts by post of many people to get the sentiment of the market, which is tiring and time taking process. Our project aims to find the sentiment of the stock market and analyzing this sentiment may be helpful to people who wants to invest in the stock media.

Figure 1: This image was big hit on the financial page of twitter platform following the elon musk supporting tweets towards favoring dogecoin cryptocurrency over other coins. This created ripples in the stock market and made Dogecoin value very high.

2. Related Work

In this particular section, we review methods for sentiment prediction using SVM (support machine vectors), Naïve-bayes and Recurrent Neural Network (RNN) which is closely related to our work.

2.1 Sentiment Analysis Using Traditional Algorithms (SVM, Naïve-bayes)

SVM is a supervised machine learning algorithm that can be used for both classification or regression challenges. Classification is predicting a label and Regression is predicting a continuous value. SVM performs classification by finding the hyper-plane that differentiate the classes we plotted in n-dimensional space [3]. Although sentiment analysis using SVM is a good technique deep learning performs fairly better than SVM and other classifiers like Naïve-bayes [6]. The main problem with SVM is that it may not perform well if the data is very noisy and we have a challenging part of analyzing data which we will mention in the section 5 in this paper.

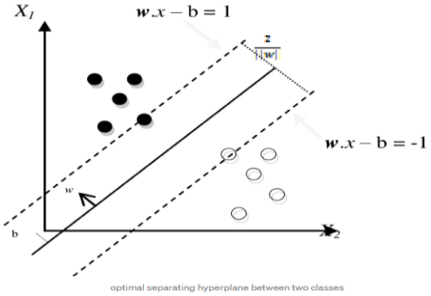


Figure 2: Support Vector Machine Graphical representation

2.2 Sentiment Analysis using RNN (Recurrent Neural Network)

RNN is commonly used neural network architecture for NLP. It has recognized to be comparatively accurate and efficient for construction of language models and in tasks of speech recognition. RNNs are mainly useful if the prediction has to be at word-level, for instance, Named-entity recognition or Part of Speech tagging. As it stores the information for current feature as well neighboring features for prediction. A RNN maintains a memory based on history information, which enables the model to predict the current output conditioned on long distance features. Below is an example of NER using RNN. The idea behind RNN is to remember what information was there in the previous neurons so that these neurons could pass information to themselves in the future for further analysis. One of the major problems of RNN is the Vanishing gradient. In any neural network, the weights are updated in the training phase by calculating the error and back-propagation through the network.

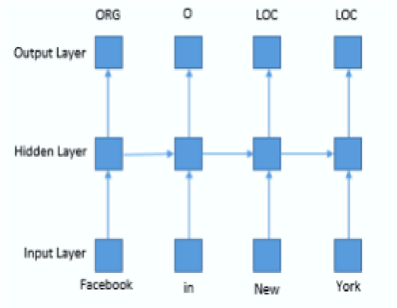


Figure 3: Recurrent Neural Network Layered architecture

But in the case of RNN, it is quite complex because we need to propagate through time to these neurons.

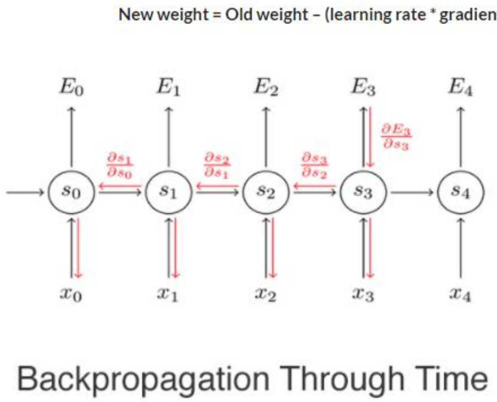

New weight = Old weight – (learning rate * gradient)
E_0
E_1
E_2
E_3
E_4
$\frac{\partial s_1}{\partial s_0}$
$\frac{\partial s_2}{\partial s_1}$
$\frac{\partial s_3}{\partial s_2}$
$\frac{\partial E_3}{\partial s_3}$
s_0
s_1
s_2
s_3
s_4
x_0
x_1
x_2
x_3
x_4
Backpropagation Through Time

Figure 4: This figure shows backpropagation of RNN

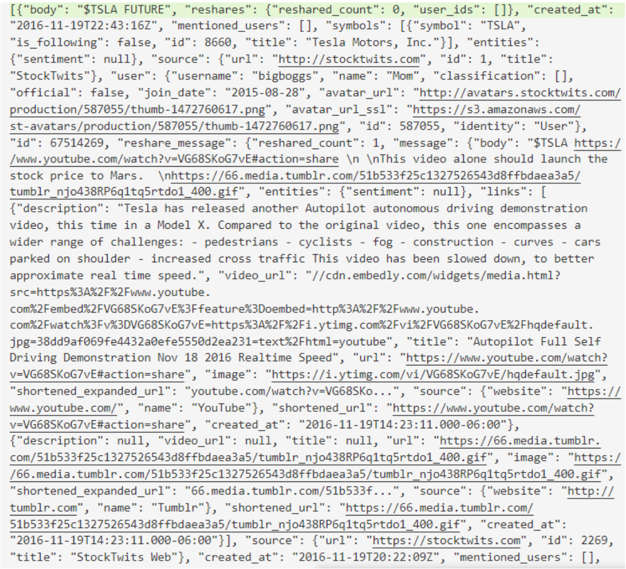
The problem lies in calculating these weights. The gradient calculated at each time instance has to be multiplied back through the weights earlier in the network. So, as we go deep back through time in the network for calculating the weights, the gradient becomes weaker which causes the gradient to vanish. If the gradient value is very small, then it won't contribute much to the learning process. To solve the problem of vanishing gradient LSTM was introduced, which we explained in the 3.5 segment.

3 Methodology

The formal financial language contains terms which are very ambiguous to understand as many terms are co-related to different meaning which will redirect the sentence to a different context which makes it harder for the non-finance related people to understand and comprehend the meaning of this language, for example terms like Liquid, short, bear and bull and etc. these terms can have different meanings when used in a non-finance field this makes the harder for the ability of the pre-trained model to give accurate results. In this method we are using LSTM and BERT approaches due to its ability to predict the sequences better than RNN (Recurrent Neural Network).

3.1 Data Collection

The data which is used to train the model is from the stocktwits platform, this data was collected by web scrapping the from the python API. The size of the data that we have collected was around 19 megabytes and it had around 18000 records of data out of which 13000 has been used to train the model and remaining data is used for testing the model.



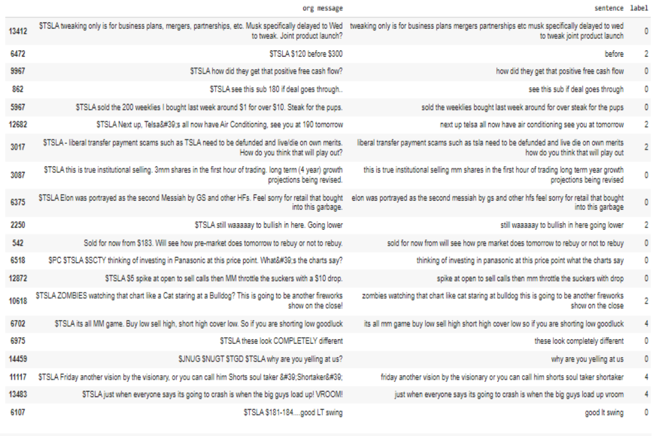
```
[{"body": "$TSLA FUTURE", "reshares": {"reshared_count": 0, "user_ids": []}, "created_at":
"2016-11-19T22:43:16Z", "mentioned_users": [], "symbols": [{"symbol": "TSLA",
"is_following": false, "id": 8660, "title": "Tesla Motors, Inc."}], "entities":
{"sentiment": null}, "source": {"url": "http://stocktwits.com", "id": 1, "title":
"StockTwits"}, "user": {"username": "bigboggs", "name": "Mom", "classification": [],
"official": false, "join_date": "2015-08-28", "avatar_url": "http://avatars.stocktwits.com/
production/587055/thumb-1472760617.png", "avatar_url_ssl": "https://s3.amazonaws.com/
st-avatars/production/587055/thumb-1472760617.png", "id": 587055, "identity": "User"},
"id": 67514269, "reshare_message": {"reshared_count": 1, "message": {"body": "$TSLA https:/
/www.youtube.com/watch?v=VG68SKoG7vE#action=share \n \nThis video alone should launch the
stock price to Mars.  \nhttps://66.media.tumblr.com/51b533f25c1327526543d8ffbdaea3a5/
tumblr_njo438RP6q1tq5rtdo1_400.gif", "entities": {"sentiment": null}, "links": [
{"description": "Tesla has released another Autopilot autonomous driving demonstration
video, this time in a Model X. Compared to the original video, this one encompasses a
wider range of challenges: - pedestrians - cyclists - fog - construction - curves - cars
parked on shoulder - increased cross traffic This video has been slowed down, to better
approximate real time speed.", "video_url": "//cdn.embedly.com/widgets/media.html?
src=https%3A%2F%2Fwww.youtube.
com%2Fembed%2FVG68SKoG7vE%3Ffeature%3Doembed=http%3A%2F%2Fwww.youtube.
com%2Fwatch%3Fv%3DVG68SKoG7vE=https%3A%2F%2Fi.ytimg.com%2Fvi%2FVG68SKoG7vE%2Fhqdefault.
jpg=38dd9af069fe4432a0efe5550d2ea231=text%2Fhtml=youtube", "title": "Autopilot Full Self
Driving Demonstration Nov 18 2016 Realtime Speed", "url": "https://www.youtube.com/watch?
v=VG68SKoG7vE#action=share", "image": "https://i.ytimg.com/vi/VG68SKoG7vE/hqdefault.jpg",
"shortened_expanded_url": "youtube.com/watch?v=VG68SKo...", "source": {"website": "https://
www.youtube.com/", "name": "YouTube"}, "shortened_url": "https://www.youtube.com/watch?
v=VG68SKoG7vE#action=share", "created_at": "2016-11-19T14:23:11.000-06:00"},
{"description": null, "video_url": null, "title": null, "url": "https://66.media.tumblr.
com/51b533f25c1327526543d8ffbdaea3a5/tumblr_njo438RP6q1tq5rtdo1_400.gif", "image": "https:/
/66.media.tumblr.com/51b533f25c1327526543d8ffbdaea3a5/tumblr_njo438RP6q1tq5rtdo1_400.gif",
"shortened_expanded_url": "66.media.tumblr.com/51b533f...", "source": {"website": "http://
tumblr.com", "name": "Tumblr"}, "shortened_url": "https://66.media.tumblr.com/
51b533f25c1327526543d8ffbdaea3a5/tumblr_njo438RP6q1tq5rtdo1_400.gif", "created_at":
"2016-11-19T14:23:11.000-06:00"}], "source": {"url": "https://stocktwits.com", "id": 2269,
"title": "StockTwits Web"}, "created_at": "2016-11-19T20:22:09Z", "mentioned_users": [],
```

Figure 5: Raw scrapped stocktwits data in JSON format

The data that we have scrapped was raw data and it was stored in a JSON file, so we had to do some preprocessing steps to remove the redundancies in the data of the collected stocktwits platform.

3.2 Data Pre-Processing

The data retrieved from the stocktwits platform is a raw data and before training period the input text has to be cleaned which will remove URL, @(sign), single letter words and apart from that we had to convert the data into lower case, remove punctuations and we also had to do labelling of the data

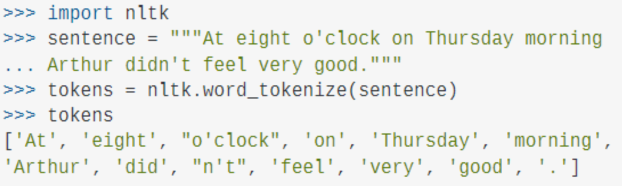


	org message	sentence	label
13412	$TSLA tweaking only is for business plans, mergers, partnerships, etc. Musk specifically delayed to Wed to tweak. Joint product launch?	tweaking only is for business plans mergers partnerships etc musk specifically delayed to wed to tweak joint product launch	0
6472	$TSLA $120 before $300	before	2
9967	$TSLA how did they get that positive free cash flow?	how did they get that positive free cash flow	0
862	$TSLA see this sub 180 if deal goes through..	see this sub if deal goes through	0
5967	$TSLA sold the 200 weeklies I bought last week around $1 for over $10. Steak for the pups.	sold the weeklies bought last week around for over steak for the pups	0
12682	$TSLA Next up, Telsa&#39;s all now have Air Conditioning, see you at 190 tomorrow	next up telsa all now have air conditioning see you at tomorrow	2
3017	$TSLA - liberal transfer payment scams such as TSLA need to be defunded and live/die on own merits. How do you think that will play out?	liberal transfer payment scams such as tsla need to be defunded and live die on own merits how do you think that will play out	2
3087	$TSLA this is true institutional selling. 3mm shares in the first hour of trading. long term (4 year) growth projections being revised.	this is true institutional selling mm shares in the first hour of trading long term year growth projections being revised	0
6375	$TSLA Elon was portrayed as the second Messiah by GS and other HFs. Feel sorry for retail that bought into this garbage.	elon was portrayed as the second messiah by gs and other hfs feel sorry for retail that bought into this garbage	0
2250	$TSLA still waaaaay to bullish in here. Going lower	still waaaaay to bullish in here going lower	2
542	Sold for now from $183. Will see how pre-market does tomorrow to rebuy or not to rebuy.	sold for now from will see how pre market does tomorrow to rebuy or not to rebuy	0
6518	$PC $TSLA $SCTY thinking of investing in Panasonic at this price point. What&#39;s the charts say?	thinking of investing in panasonic at this price point what the charts say	0
12872	$TSLA $5 spike at open to sell calls then MM throttle the suckers with a $10 drop.	spike at open to sell calls then mm throttle the suckers with drop	0
10618	$TSLA ZOMBIES watching that chart like a Cat staring at a Bulldog? This is going to be another fireworks show on the close!	zombies watching that chart like cat staring at bulldog this is going to be another fireworks show on the close	2
6702	$TSLA its all MM game. Buy low sell high, short high cover low. So if you are shorting low goodluck	its all mm game buy low sell high short high cover low so if you are shorting low goodluck	4
6975	$TSLA these look COMPLETELY different	these look completely different	0
14459	$JNUG $NUGT $TGD $TSLA why are you yelling at us?	why are you yelling at us	0
11117	$TSLA Friday another vision by the visionary, or you can call him Shorts soul taker &#39;Shortaker&#39;	friday another vision by the visionary or you can call him shorts soul taker shortaker	4
13483	$TSLA just when everyone says its going to crash is when the big guys load up! VROOM!	just when everyone says its going to crash is when the big guys load up vroom	4
6107	$TSLA $181-184....good LT swing	good lt swing	0

Figure 6: Cleaned labeled data of stocktwits

3.3 Tokenizing

Tokenization is the process of splitting a sentence, paragraph or an entire text document into smaller units of words. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. There are different methods and libraries available to perform tokenization. NLTK, Gensim, Keras are some of the libraries that can be used to accomplish the task. In this method we are using NLTK library to tokenize the input data. In this method we useword_tokenize method which basically returns the individual words from the string.



```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
```

Figure 7: In the above figure we can see the words are separated and tokenized using the word_tokenizer from NLTK library in python

3.4 Data Visualization:

After preprocessing the data, we need to view the most frequent words so we are creating the word cloud of the most frequent repeating words after tokenization of those words.

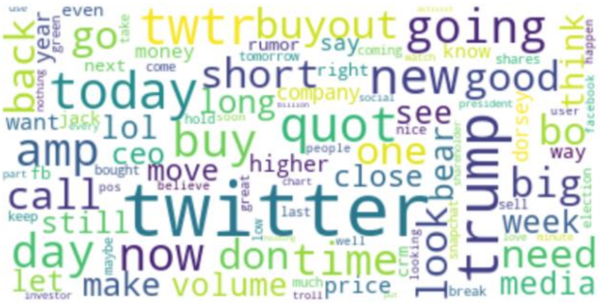


Figure 8: The above figure shows the word cloud of the most frequent words that are occurring in the dataset. Apart from this wea re also seeing the distribution of the labels across the dataset and we are viewing that in a bar graph.

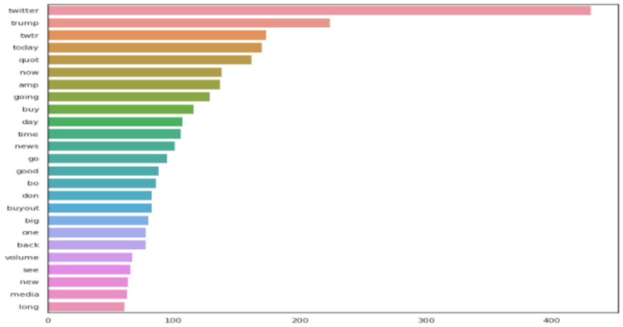


Figure 9: The above figure shows the bar chart of the most frequent words on the corpus

3.5 LSTM(Long Short Term Memory)

Long-Short-Term-Memory are usually just called "LSTMs" are a special kind of RNN, they are skilled in learning long-term dependencies [2]. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people. They work extremely well on a large variety of problems, and are now extensively used. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior This is the reason we are using LSTM over a basic feed forward neural network or RNN. All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer. The reason for using LSTM is to overcome the vanishing gradient problem which could be encountered if we would have used RNN(Recurrent neural network). Below figure 10 shows the architecture of LSTM.

It has a memory cell at the top which helps to carry the information from a particular time instance to the next time instance in an efficient manner. So, it can able to remember a lot of information from previous states when compared to RNN and overcomes the vanishing gradient problem. Information might be added or removed from the memory cell with the help of valves.LSTM network is fed by input data from the current time instance and output of hidden layer from the previous time instance.

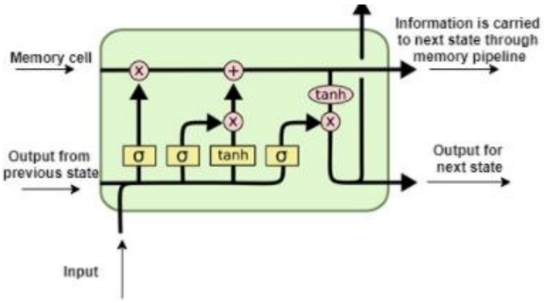

Memory cell
Information is carried to next state through memory pipeline
Output from previous state
Output for next state
Input
σ
σ
tanh
σ
tanh
×
+

Figure 10: The above figure shows the LSTM cell

These two data passes through various activation functions and valves in the network before reaching the output.

3.6 BERT

BERT stands for Bidirectional Representation for Transformers. It was proposed by researchers at Google Research in 2018. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words or sub-words in a text. It learns how important all of the words in the sentence are by looking at their specific positions in the sentence [1]. The Transformer reads entire sequences of tokens at once and does not take into account directionality while the attention mechanism pay attention to one specific word in the sentence simultaneously.

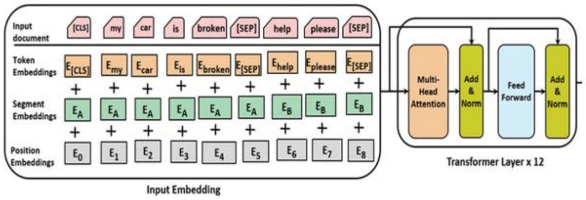

Input document
[CLS]
my
car
is
broken
[SEP]
help
please
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{car} E_{is} E_{broken} $E_{[SEP]}$ E_{help} E_{please} $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8
Input Embedding
Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Transformer Layer x 12

Figure 11: The above figure shows the BERT Architecture

BERT is a Transformer and Transformer is a popular attention mechanism used to learn contextual relations between words in a text using artificial intelligence. Using BERT was important for us because it has the ability to know the context of the data. Apart from that BERT can read a text (or a sequence of words) all at once, with no specific direction. Thanks to its bidirectionality, this model can understand the meaning of each word based on context both to the right and to the left of the word; this represents a clear advantage in the field of context learning.

4.0 Experiments and Evaluations

As described earlier we have done all the steps which involved pre-processing the data and after pre-processing we are running the data on the LSTM and the BERT model

4.1 Performance Evaluation of the LSTM

To evaluate the model, we have used linear schedule and AdamW optimizer to avoid the exploding gradient problem we have introduced the concept of gradient clipping also. We have trained the model on 5 epochs with batch size =64.

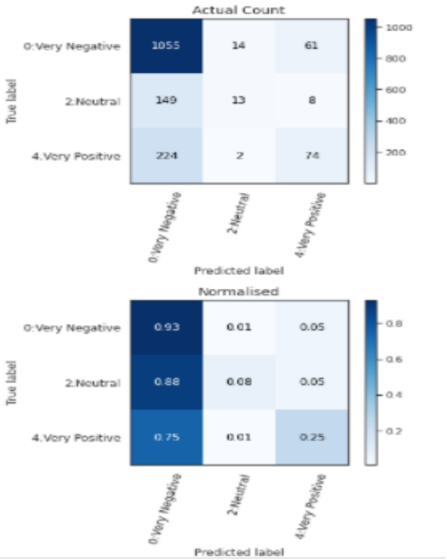


Figure 12: The above figure shows the confusion matrix for the predictions of LSTM model

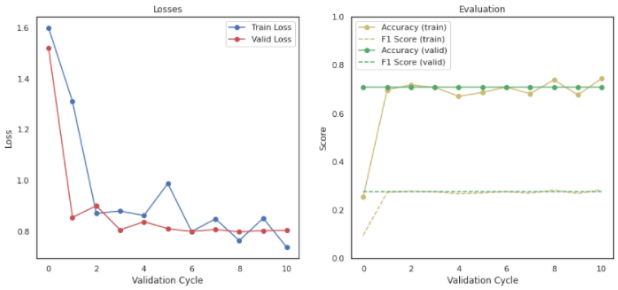
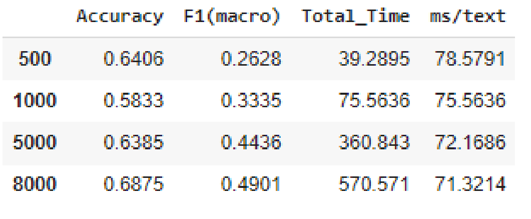


Figure 13: This is the results of the model on the data



	Accuracy	F1(macro)	Total_Time	ms/text
500	0.6406	0.2628	39.2895	78.5791
1000	0.5833	0.3335	75.5636	75.5636
5000	0.6385	0.4436	360.843	72.1686
8000	0.6875	0.4901	570.571	71.3214

Figure 14: These are the accuracies of the LSTM model

In the above figure we can see the accuracies of the LSTM network and by increasing the data the performance was increasing significantly.

4.2 Performance Evaluation of the BERT

Even in this section we have used the same Adam optimizer and linear schedule to evaluate our model.

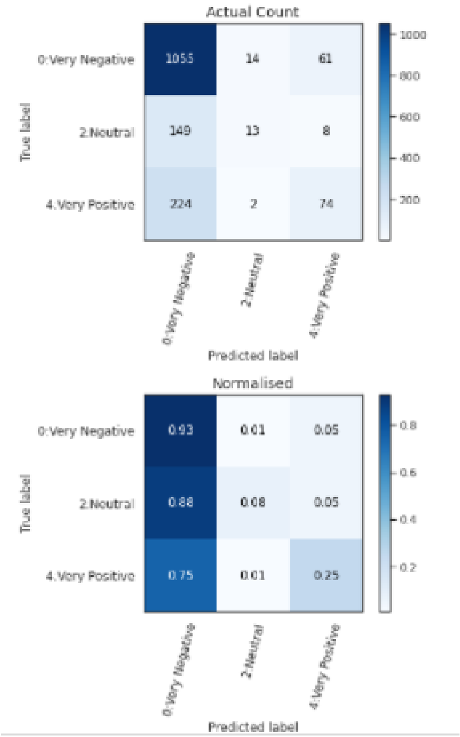


Figure 15: The above figure shows the confusion matrix of the predicted label of BERT model

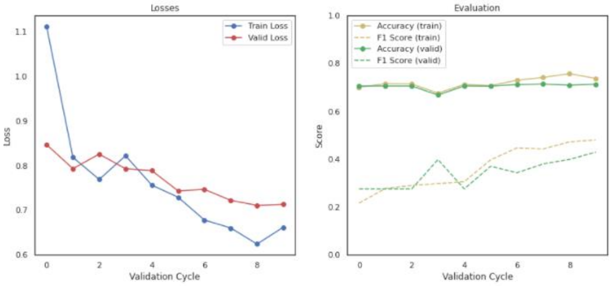
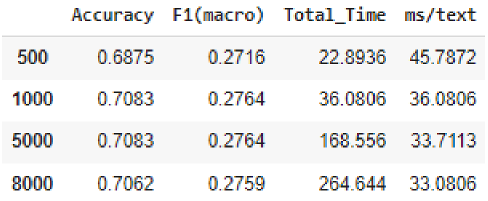


Figure 16: This is the results of the BERT model on the data



	Accuracy	F1(macro)	Total_Time	ms/text
500	0.6875	0.2716	22.8936	45.7872
1000	0.7083	0.2764	36.0806	36.0806
5000	0.7083	0.2764	168.556	33.7113
8000	0.7062	0.2759	264.644	33.0806

Figure 17: These are the accuracies of the BERT model

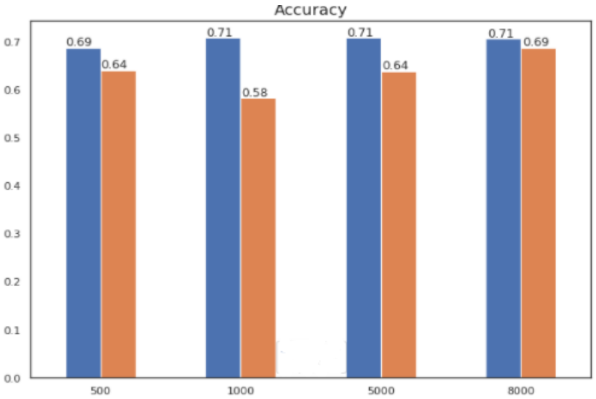
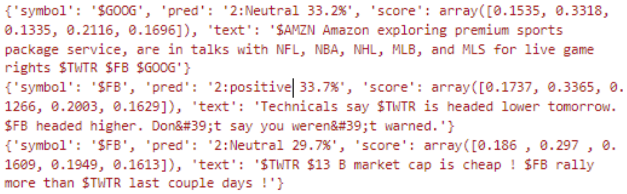


Figure 18: The above bar charts shows the difference between accuracies of LSTM (Red color) and BERT (Blue color)



```
{'symbol': '$GOOG', 'pred': '2:Neutral 33.2%', 'score': array([0.1535, 0.3318,
0.1335, 0.2116, 0.1696]), 'text': '$AMZN Amazon exploring premium sports
package service, are in talks with NFL, NBA, NHL, MLB, and MLS for live game
rights $TWTR $FB $GOOG'}
{'symbol': '$FB', 'pred': '2:positive 33.7%', 'score': array([0.1737, 0.3365, 0.
1266, 0.2003, 0.1629]), 'text': 'Technicals say $TWTR is headed lower tomorrow.
$FB headed higher. Don&#39;t say you weren&#39;t warned.'}
{'symbol': '$FB', 'pred': '2:Neutral 29.7%', 'score': array([0.186 , 0.297 , 0.
1609, 0.1949, 0.1613]), 'text': '$TWTR $13 B market cap is cheap ! $FB rally
more than $TWTR last couple days !'}
```

Figure 19: These are the Prediction made using the BERT model

5.Challenges and Future work:

The main challenge faced during the implementation of this model was pre-processing data and labelling them and apart from that we faced problems during the training of the model due to the size of the data, as after cleaning and pre-processing the size of the data significantly reduces and this makes the classifying the data hard. Another major problem is that financial data usually contains frequent terms like “short”, “liquid”, “ bear” and etc. As we can see that these terms are interchanged between who is using them and what context it is being used and sometimes it changes the entire meaning of the sentence and these terms can cause the model to give some false predictions.
In the future work, we are hoping to increase the dataset size by scrapping more data and increase the computational time of the model by giving more time to train the model, apart from that we can also create a model which can have a lookup dictionary of certain words that are being used interchangeable in the field of finance and Apart from that we can try new algorithms such as Albert and Bi-direction LSTM and evaluate their performance.

6. Conclusion

Given the same data and the resources we have seen that BERT performs better than the LSTM, although the performance is better it is not a big difference between both of them, in our case we have seen with the Increase of the amount of data the accuracies have been improving, In the future work, to increase the accuracy of the model we need to have more data.

7. Acknowledgement

we take this opportunity to express my profound gratitude and deep regards to my teachers Professor Chenliang Xu for his exemplary guidance, monitoring and constant encouragement throughout the course of this project.. Apart from that we would like to thank Graduate Teaching Assistant Chao Huang and Sizhe le for their overall support through the course and completion of the project.

8.**References**

[1] Sousa, M. G., & Sakiyamav, K. (2019). BERT for StockMarket Sentiment Analysis. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI).

[2] S. Wen et al., "Memristive LSTM Network for Sentiment Analysis," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 3, pp. 1794-1804, March 2021.

[3] Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment Analysis of Tweets using SVM. International Journal of Computer Applications (0975 – 8887).

[4] M. Day and C. Lee, "Deep learning for financial sentiment analysis on finance news providers," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

[5] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019

[6]Ahmed Sulaiman M. Alharbi, Elise de Doncker,
Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information.