European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 02, 2020

Stock Prediction using Sentiment analysis and Long Short Term Memory

Harsh Panday[1], V. Vijayarajan[2*], Anand Mahendran[3], A. Krishnamoorthy[4], V.B. Surya Prasath[5]

[1]*Student, SCOPE, Vellore Institute of Technology, Vellore.*
[2*]*Associate Professor, SCOPE, Vellore Institute of Technology, Vellore.*
[3]*Associate Professor, SCOPE, Vellore Institute of Technology, Vellore.*
[4]*Assistant Professor Senior, SCOPE, Vellore Institute of Technology, Vellore.*
[5]*Department Electrical Engineering and Computer Science, University of Cincinnati, USA.*

Abstract: The Stock market is a shambolic place for prediction as there are plenty of factors that affect the stock market simultaneously. Numerous studies have been conducted regarding this field, in hopes that one day accurate stock values can be predicted. This paper introduces a hybrid algorithm that incorporates Twitter sentiment analysis and Long Short Term Memory to predict next day closing values of a stock. Our proposed algorithm exploits the temporal correlation between public sentiment and its effect on stock values. We use Part-of-speech tagging to perform sentiment analysis and Long Short Term Memory for foretelling the next day closing price of the stock, both of these combined gives us a decent picture regarding the future of the stock.

Index Terms: Sentiment analysis, LSTM, Machine Learning, Stock prediction, Twitter, TextBlob, Public Sentiment, Positive Sentiment, Negative Sentiment.

1. INTRODUCTION

Stock market prediction is often deemed as one of the most arduous tasks in the financial world. This is because of the many agents that are involved in changing the trend of stock values. Agents like political events, the general sentiment of people towards the company, economic conditions and many more are simultaneously in play in the changing the trend of stock values. Hence, it is quite toilsome to predict stock values. During recent years machine learning and artificial intelligence have played a considerable role in creating algorithms that help to predict the direction of the stock market.

In this paper, we combine machine learning and public sentiment. This has been achieved by using a hybrid algorithm that uses sentiment analysis and LSTM to predict the next day stock values and the public's sentiment, which helps us to correlate the market conditions and public sentiment. Publicly available Twitter data is used to perform sentiment analysis and yahoo finance to get stock values.

2. LITERATURE SURVEY

An ample amount of machine learning approaches have been used over the past decade or so to foresee stock market values. Support vector regression is used in [9] to predict next days of data relating to stock market. In [3] Artificial Neural Network is used to speculate stock market indices; particularly they predicted that whether the trend movement will go up or down. Textual analysis has been in done in [10] in which three different words based representations that are Bag of Words, Noun phrases, Named entities they used the ability of

these textual representation's ability to foresee discrete amount of stock prices 20 minutes after any write-up is released regarding that stock or company. In [6] Backpropogation and multilayer feed forward network to predict and calculate stock values.

Recent research trends tend to combine two or more different algorithms together to make a hybrid algorithm. Rohit Choudhary, Kumkum Garg [2] has proposed a machine learning system that is a blend of Genetic Algorithm and Support Vector Machine (SVM) and they found that Genetic Algorithm improved SVM to give more accurate results. The study in [1] shows a machine learning model that integrates Particle Swarm Optimization (PSO) and Least Square Support vector machine (LS-SVM). They found that the hybridized algorithm they used converged to the global minimum and overcame the over fitting problem that was being faced by Artificial Neural Network, especially during fluctuations.

In last few years analysis and correlation are becoming a big part of these algorithms. The study conducted in [5] shows that conducting fundamental and technical analysis helped them to incorporate various features which helped them to increase their accuracy. In [7] they proposed a advanced algorithm that make use of the temporal correlation among global stock market and diverse financial products to foresee the next day stock trend with the assistance of SVM.

In this work LSTM and sentiment analysis based model is proposed to predict next day closing value of stock market. LSTM is used because predictions made by it are always conditioned by the past experience of the inputs, as we know stock market is affected by many things and this is where LSTM plays a huge role as it has some contextual state cells that act as long-term and short-term memory cells which allows it to take both past inputs and current inputs in account to predict the next value, because of this it can predict stock values with minimal error. On the other hand sentiment analysis helps include public sentiment in our study and how it affects the stock market. In this work we have used Apple, Microsoft and Google's data and predicted there day closing values.

Rest of the work is organized as follows. In Section 2 we describe our dataset and we introduce the LSTM driven prediction model. Section 3 provides a case study based on the model using Apple, Microsoft and Google's stock data. Finally, Section 4 conclusions are given regarding this work.

3. METHODOLOGY

Dataset

In this project, two main datasets have been used –

1. Yahoo finance stock data from 01-01-2012 to 18-02-2020. The data includes open, close, high, low, values of a given day.
2. Publicly available Twitter data. This incorporates the timestamp and tweet text for every tweet of a particular period. Since predictions are being made on daily basis, tweets are split by day using their timestamps.

Prediction

The training and prediction is done using Long Short term memory (LSTM) architecture. LSTM architecture is a part of recurrent neural network and is generally used in the field of deep learning. LSTM has feedback connections, which makes it very useful to process entire sequences of data. Before feeding data to LSTM, data should be processed and normalized. The other dataset that is used for sentiment analysis is obtained from Twitter; this data should also be processed before using it to do sentiment analysis.

Data Processing

a. First, we use pandas to obtain data from yahoo finance. As we our target value is the Close value, a target dataframe is created with only close column. The data is then normalized and converted so that all the values lie between 0 and 1.Data is then divided into two parts i.e. training data (70 percent) and testing data (30 percent).
b. Second, the data form Twitter is obtained using Tweepy. Tweepy is a library that is used for accessing the Twitter API. After obtaining the tweets from the API they are cleaned that is all links and special characters are removed. After cleaning they are then divided according to their polarity that is positive polarity means the tweet is positive and negative polarity means the tweet is negative.

Stock and Date Selection

- For this project Apple, Google, Microsoft's stock values have been used. Their stock values are publicly available. The values used are from 01- 01-2012 to 18-02-2020.

4. PREDICTION MODELS

Sentiment Analysis

Social media in the past few years has changed the way investors predict the stock market. As real time user opinion is present on social media, investors exploit this data to predict stock prices. In [11] a study has been done using 18 million tweets relating to stocks, there purpose was to find out whether user sentiment is reflected in the market or not. The experiment proved that tweets do fluctuate with stock values; it was found that the effect of negative sentiment is more crucial than 4 Harsh Panday, V. Vijayarajan, and V. B. Surya Prasath positive sentiment. They concluded that 1% increase in negative tweets leads to 0.03% drop in stock returns. Meanwhile positive sentiment does not show its effect on stock return on daily basis as drastically as the negative sentiment does, but it does have repercussion in the long term over the stock.

In this project Part of speech tagging is used in order to perform sentiment analysis. First of all we have used the Twitter API to collect the latest tweets regarding the company whose stock is being predicted. After collection, the tweets are cleaned that is any special character or links are removed from the tweet. Then the tweet is stored as a bag of words. The next step is using TextBlob an extremely powerful NLP library to perform lexicon based part of speech tagging, TextBlob assigns scores to all the words based on a pre-defined dictionary. After assigning scores to all the words, sentiment is calculated by taking average of the polarity of the word. A word will have a sentiment based on its polarity and its polarity depends on the context in which the words are used. The table below shows how the word outstanding can be used in various contexts and have different polarity.

Table 1 Different Polarities of the Word Outstanding

Word	Sense	Polarity
Outstanding	"having a quality that thrusts itself into attention"	0.5
Outstanding	"of major significance or importance"	1.0
Outstanding	"owed as a debt"	-0.5
Outstanding	"distinguished from others in excellence"	1.0

We have used this to determine the polarity of the tweets that is whether the tweet is positive or negative. Aggregating and comparing the number of positive tweets to the number of negative tweets we can determine the public sentiment.

LSTM

Long Short Term Memory or LSTM is a type of recurrent neural network that is capable of learning order dependencies in sequence prediction problems. It has feedback connections which help it to process sequence of data. It also has internal state cell which function as long or short term memory cells. The output is regulated by these state cells. This property is very useful when we need to depend on prior inputs rather than the latest ones. As time passes it is less likely that the output will be dependent on very old inputs, these inputs are forgotten by it through their forget gates, the forget gates is just a multiplicative factor of 0.9, that is within 12 steps the factor becomes $0.9^{12}= 0.282$. Equations used by forget, input and output gates in LSTM are given below. These equations help LSTM to consider previous inputs in predicting the next output and also the forget gate equation helps in forgetting very old inputs.

$$f_t = \sigma(w_f (h_t-1, x_t) + b_f)$$
$$i_t = \sigma(w_i(h_t-1, x_t) + b_i)$$
$$o_t = \sigma(w_o(h_t-1, x_t) + b_0)$$

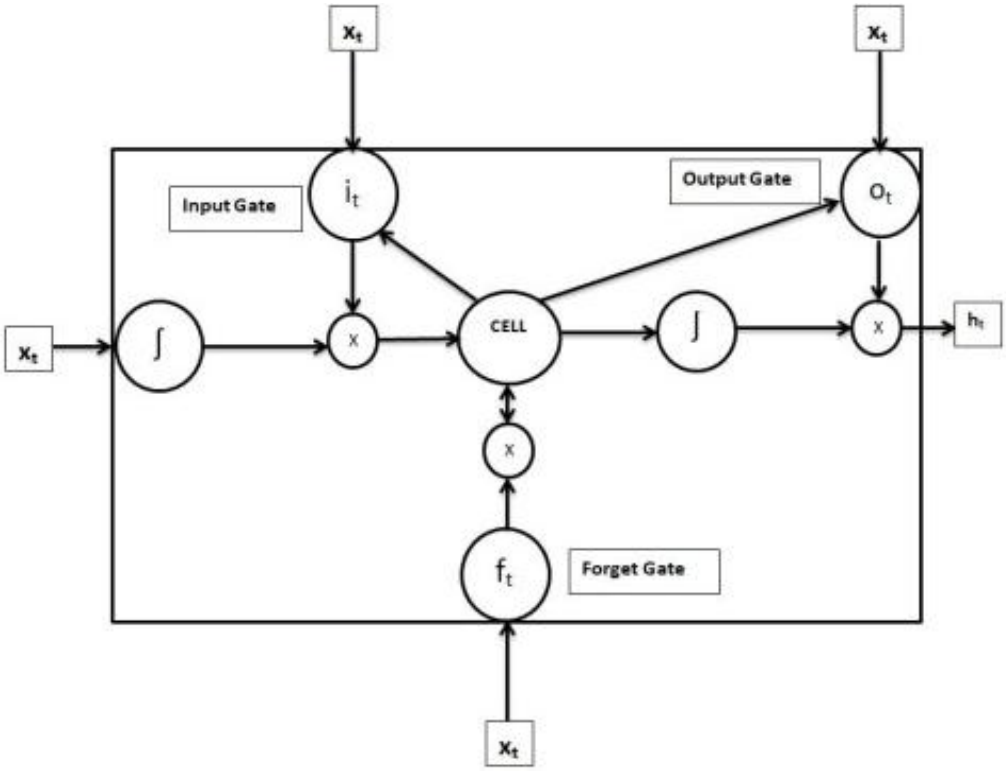

Fig. 1 LSTM Internal Wiring showing different Gates

In the proposed algorithm we use LSTM to predict next day closing values of stock. The first step is to scrape the data form yahoo finance. The data is then scaled in order to fit the values between 0 and 1.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Fig. 2 Formula for Rescaling

The data is then split into training and testing data set. After converting and storing the data as numpy arrays it is then trained under the LSTM model. We used Adam as our optimizer and trained 4 epochs. The predicted close values were then compared to the actual close values and Root Mean Squared Error was calculated in order to determine our accuracy and a chart is prepared in order to visualize the comparison between the predicted and actual values.

5. EXPERIMENTS AND RESULTS

We tested our approach with three stocks Apple, Google and Microsoft. The data used in this study were obtained from Twitter and yahoo finance. We collected in all 2546 trading days' data from January 2012 to February 2020. For each day, the opening, highest, lowest and closing values of the stock price were obtained. After training the data the data we test our model on the remaining. We then prepare a chart in order to better visualize the difference between out test and trained prediction values. The below charts shows that comparison

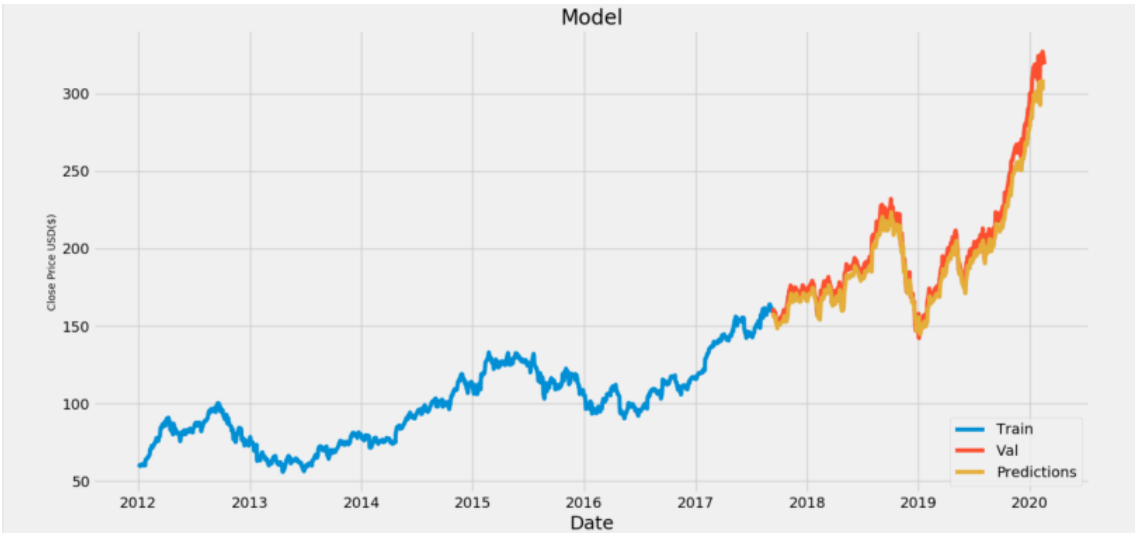

Fig. 3 Comparison between Trained Data and Actual Data of Apple

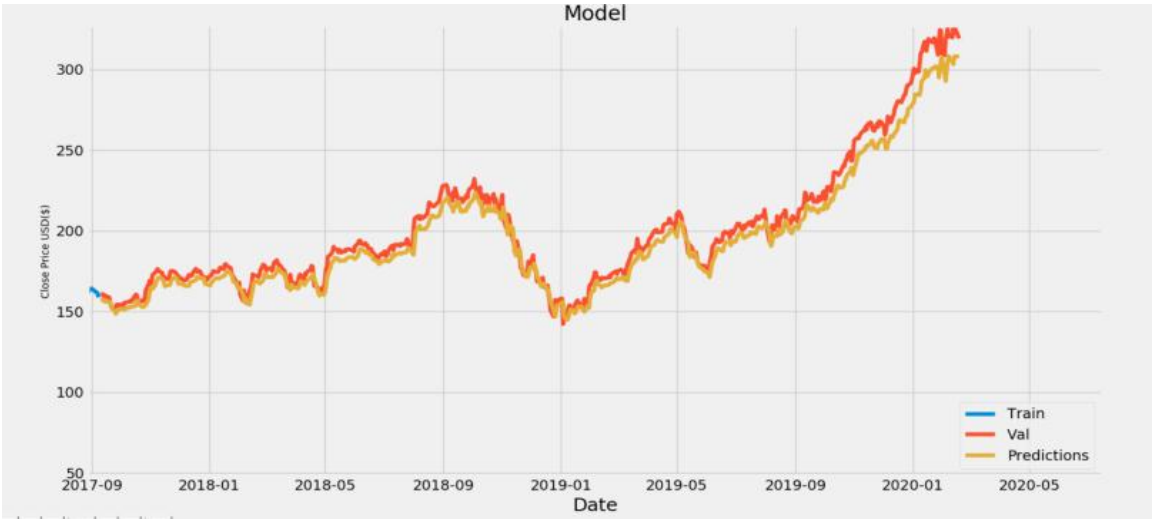

Fig. 4 Comparison between Trained Data and Actual Data of Apple (Zoomed)

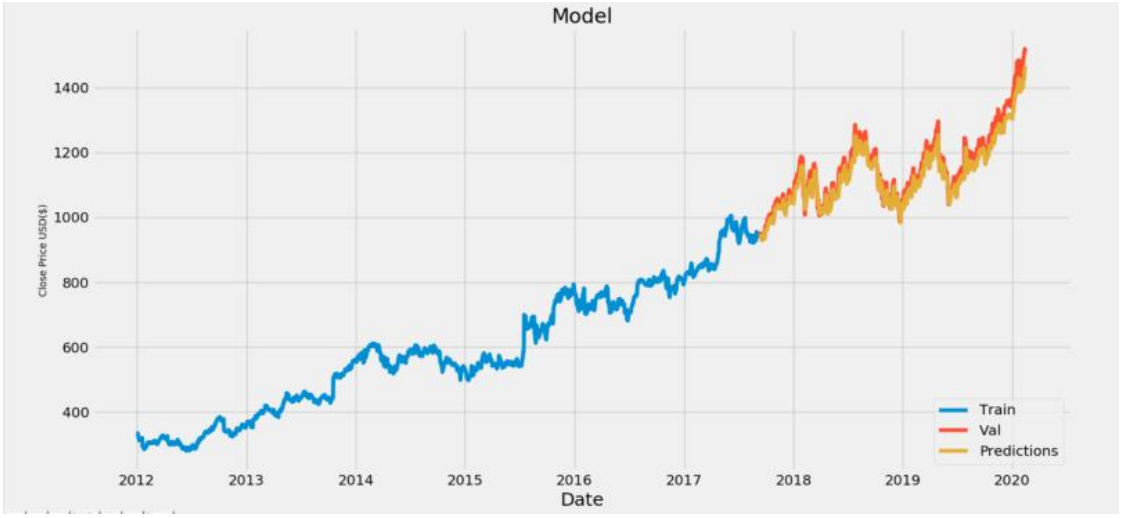

Fig. 5 Comparison between Trained Data and Actual Data of Google

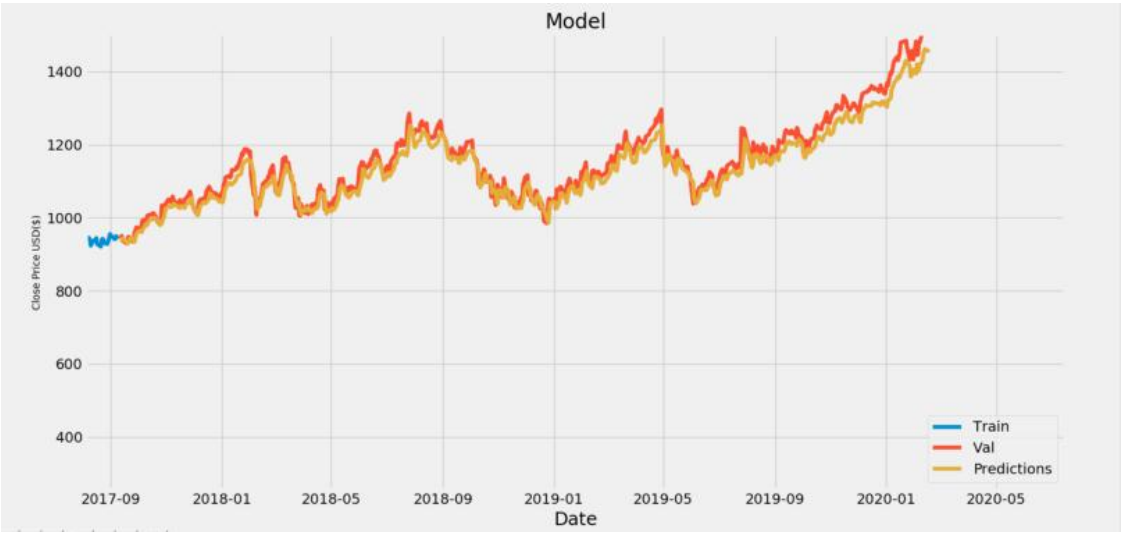

Fig. 6 Comparison between trained data and actual data of Google (zoomed)

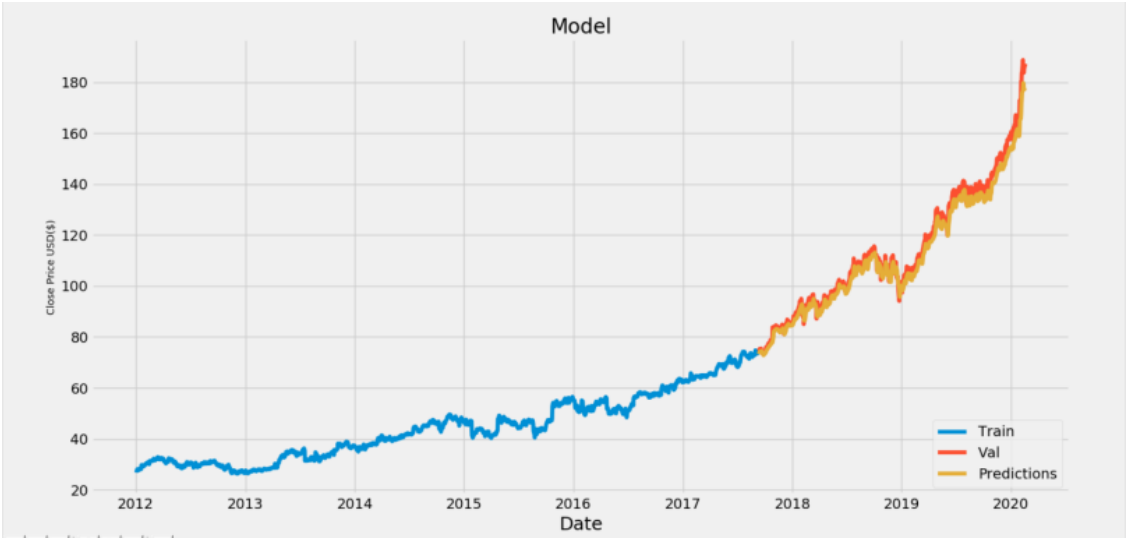

Fig. 7 Comparison between Trained Data and Actual Data of Microsoft

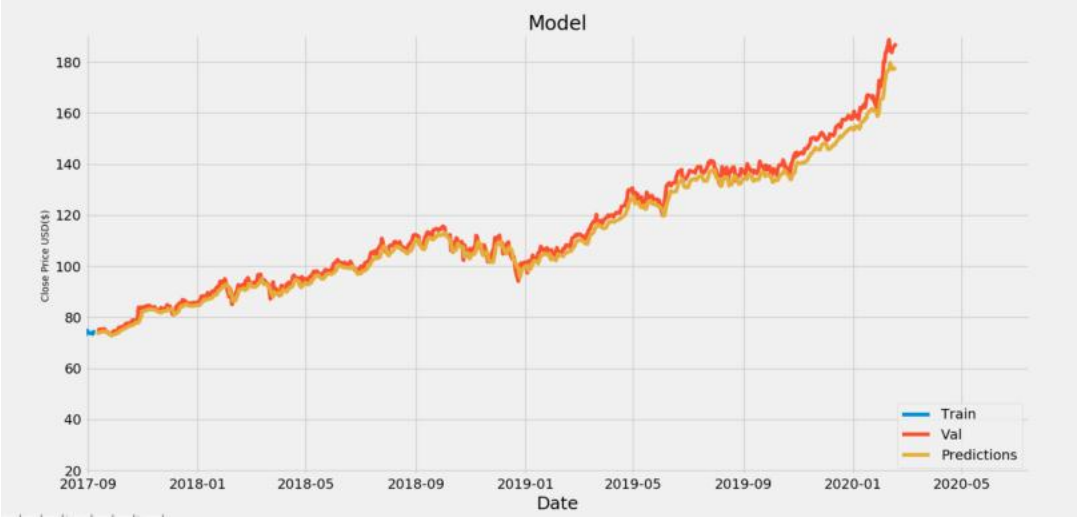

Fig. 8 Comparison between trained data and actual data of Microsoft (zoomed)

The below tables shows the predictions made by our algorithm of the recent dates and the actual stock values of those dates. The tables below show the last five predictions. As we can see the difference between the predicted and actual value is very less, hence it shows the LSTM can predict the next day close value very accurately.

Table 2 Prediction vs. Actual stock values of Apple

Date	Prediction	Actual
2020-02-11	190.08	184.44
2020-02-12	186.66	184.71
2020-02-13	185.76	183.71
2020-02-14	185.15	185.35
2020-02-18	187.09	187.22

Table 3 Prediction vs. Actual Stock Values of Google

Date	Prediction	Actual
2020-02-11	1504.32	1510.06
2020-02-12	1512.20	1518.63
2020-02-13	1500.17	1513.39
2020-02-14	1505.36	1518.72
2020-02-18	1511.06	1519.43

Table 4 Prediction vs. Actual stock values of Microsoft

Date	Prediction	Actual
2020-02-11	317.43	319.6
2020-02-12	315.47	320.2
2020-02-13	323.33	324.86
2020-02-14	322.04	324..95
2020-02-18	321.13	319.00

We calculate root mean square error (RMSE) to get the error value in our prediction.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

Fig. 9 Formula for RMSE

The average RMSE error for our model is 1.2583.

This low RMSE error shows that LSTM is extremely accurate in predicting stock values. After the LSTM model compilation the sentiment analysis begins. The latest tweets are collected in order to get the up to the minute public sentiment regarding the stock. The increment in the negative sentiment can be juxtaposed with the values of stock and the effect of the negative sentiment can be clearly seen. The below pie charts show the public sentiment of Apple, Google and Microsoft.

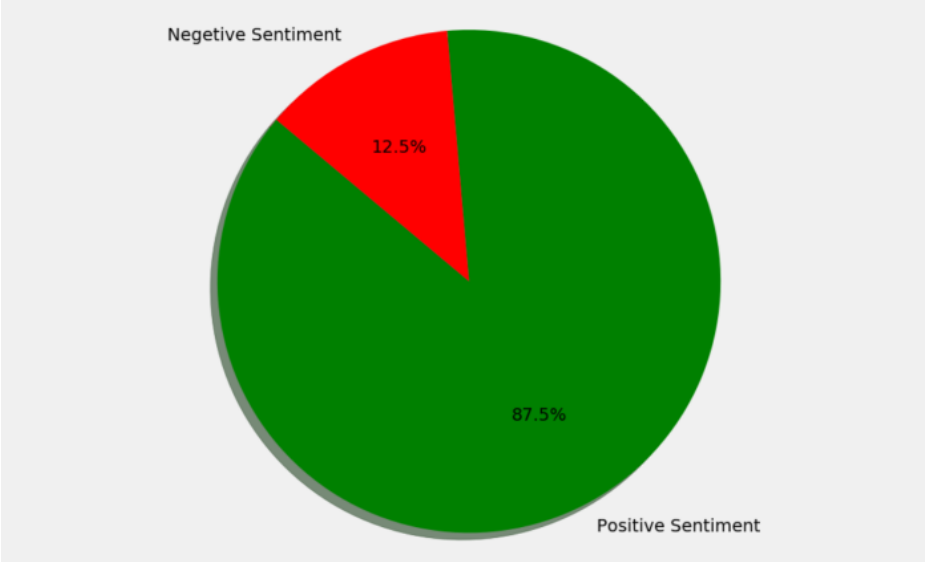

Fig. 10 Sentiment Analysis of Apple

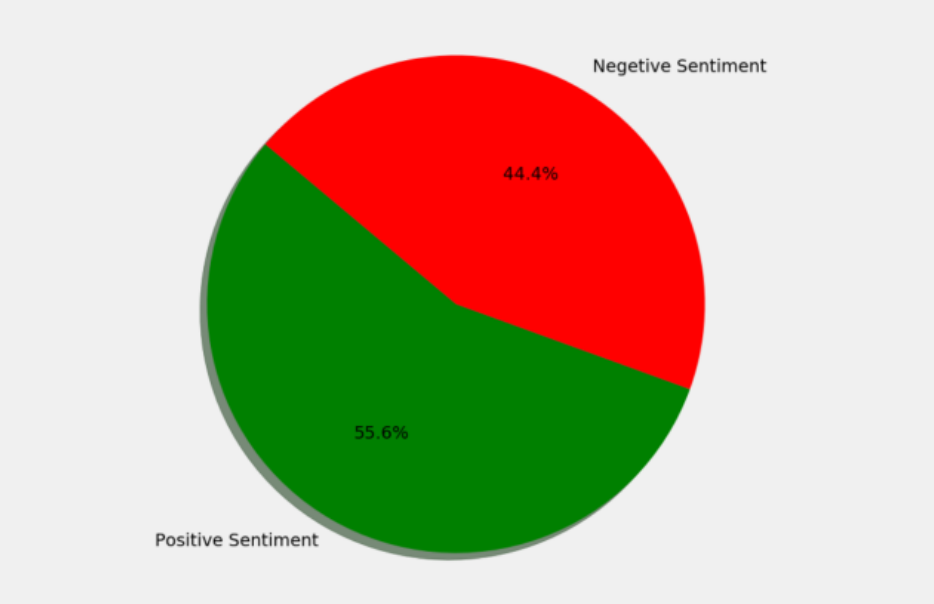

Fig. 11 Sentiment Analysis of Google

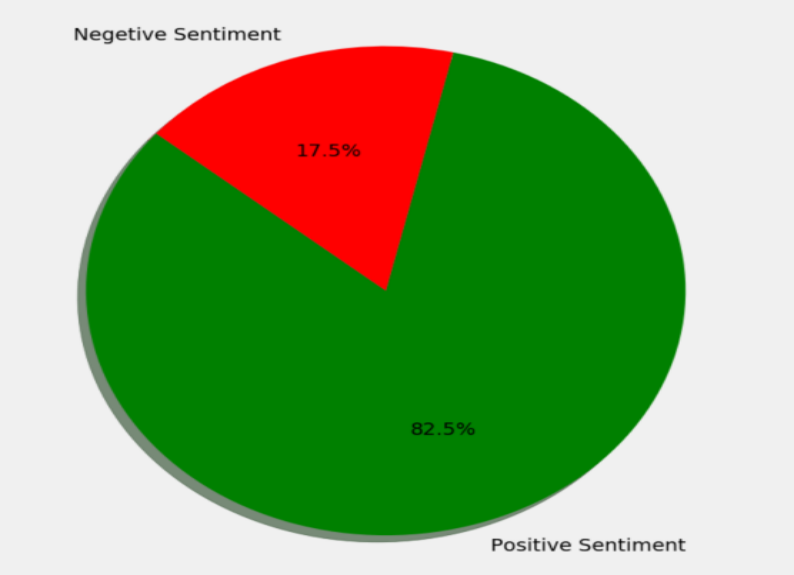

Fig. 12 Sentiment Analysis of Microsoft

6. CONCLUSION

In this paper, we proposed a hybrid system which uses neural network LSTM to predict stock values and sentiment analysis to verify our predictions. Sentiment analysis helps us to take various political and economic factors in consideration, as these factors have a major effect on the stock market. Our results show that there indeed is a relation linking public sentiment and stock values. On the other hand LSTM has proven to be the best algorithm to predict stock values as it takes previous values into consideration but also uses forget gates to remove very old values as it is not likely that the next outcome will depend upon them, this makes it very efficient.

There are number of further directions that can be investigated beginning from this project. The first one is to include other social media sites along with Twitter to get the public sentiment. Second, our dataset only considers English speaking people, in order to map out real public sentiment all languages must be included. Finally LSTM could be optimized even more in order to predict more accurate values.

7. ACKNOWLEDGMENT

None.

8. CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

9. REFERENCES

[1] Hegazy, O., Soliman, O.S., Salam, M.A. (2014). A machine learning model for stock market prediction. *arXiv preprint arXiv:*1402.7351.

[2] Choudhry, R., Garg, K. (2008). A hybrid machine learning system for stock market forecasting. *World Academy of Science, Engineering and Technology,* 39(3), 315-318.

[3] Di Persio, L., Honchar, O. (2016). Artificial neural networks architectures for stock price prediction: Comparisons and applications. *International journal of circuits, systems and signal processing*, 10(2016), 403-413.
[4] Mittal, A., Goel, A. (2012). Stock prediction using Twitter sentiment analysis. Standford University, CS229 (2011 http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingtwitterSentimentAnalysis.pdf), 15.
[5] Shah, V.H. (2007). Machine learning techniques for stock prediction. *Foundations of Machine Learning— Spring*, 1(1), 6-12.
[6] Khan, Z.H., Alin, T.S., Hussain, M.A. (2011). Price prediction of share market using artificial neural network (ANN). *International Journal of Computer Applications*, 22(2), 42-47.
[7] Shen, S., Jiang, H., Zhang, T. (2012). *Stock market forecasting using machine learning algorithms.* Department of Electrical Engineering, Stanford University, Stanford, CA, 1-5.
[8] Tsai, C.F., Wang, S.P. (2009, March). Stock price forecasting by hybrid machine learning techniques. *In Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, No. 755, p. 60).
[9] Nunno, L. (2014). *Stock market price prediction using linear and polynomial regression models.* Computer Science Department, University of New Mexico: Albuquerque, NM, USA.
[10] Schumaker, R., Chen, H. (2006). Textual analysis of stock market prediction using financial news articles. *AMCIS 2006 Proceedings*, 185. Oct 2017.
[11] Deng, S., Huang, Z.J., Sinha, A.P., Zhao, H. (2018). The interaction between microblog sentiment and stock return: An empirical examination. *MIS quarterly*, 42(3), 895-918.
[12] K. Greff, R.K. Srivastava, J. Koutn´ık, B.R. Steunebrink and J. Schmidhuber. LSTM: A Search Space Odyssey," *in IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222-2232, Oct. 2017.
[13] Gers, F.A., Schmidhuber, J., and Cummins, F. (1999). *Learning to forget: Continual prediction with LSTM.*
[14] Xingjian, S.H.I., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., and Woo, W.C. (2015). Convolutional LSTM network: A machine learning approach for precipitation now casting. *In Advances in neural information processing systems* (pp. 802-810).
[15] Sundermeyer, M., Schl¨uter, R., and Ney, H. (2012). LSTM neural networks for language modeling. *In Thirteenth annual conference of the international speech communication association.*