Stock Market analysis Using Twitter

BAN 675

(Text Mining)

PREPARED BY:

Abhisha Burande, Anshika Sharma, Maitreyee Das, Priyanka Kushwaha and Shweta arande

College of Business & Economics

California State University, East Bay

ABSTRACT

Predicting stock market movements is a well-known problem of interest. Now-a-days social media is perfectly representing the public sentiment and opinion about current events. Twitter has attracted a lot of attention from researchers for studying the public sentiments. The thesis of this work is to observe how well the changes in stock prices in the market, the rises and falls, are correlated with the public opinions being expressed in tweets about the market. Understanding the author's opinion from a piece of text is the objective of sentiment analysis. The present project has employed two different techniques, loughran and textblob, for analyzing the public sentiments in tweets. We have applied sentiment analysis and supervised machine learning principles to the tweets extracted from Twitter and analyze the correlation between stock market movements and sentiments in tweets. In an elaborate way, positive or negative news and tweets in social media about the stock market would definitely encourage or discourage people to invest and as a result the stock price would increase or decrease. So different regression models were constructed between the stock prices and the public sentiments. At the end, correlation was evaluated between public sentiments and stock market prices through random forest regression and linear regression technique and later the performance based on the parameters, accuracy of the models were calculated. Concludingly, it is shown that a strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets.

I. INTRODUCTION

Microblogging today has become a very popular communication tool among Internet users. Millions of messages are appearing daily in popular web-sites that provide services for microblogging such as Twitter, Facebook, Instagram. Authors of those messages write about their life, share opinions on a variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to microblogging services. Recent extreme events show that Twitter, a microblogging service, is emerging as the dominant social reporting tool to spread information on social crises. It is elevating the online public community to the status of first responders who can collectively cope with social crises. Especially, Twitter has attracted a lot of attention from researchers for studying the public sentiments. In previous research, it was implied that if it is properly modeled, Twitter can be used to forecast useful information about the market. Stock market prediction has been an active area of research for a long time. Stock market prediction on the basis of public sentiments expressed on Twitter has been an intriguing field of research.

In this project, we test a hypothesis based on the premise of behavioral economics, that the emotions and moods of individuals affect their decision-making process, thus, leading to a direct correlation between "public sentiment" and "market sentiment". We perform sentiment analysis on publicly available Twitter data to find the public mood. Further, we use these moods and actual Dow values to compare with the predicted stock price values and then analyze the accuracy of our model.

Our project is divided in the following sections. Firstly, we have described the methodology of data collection i.e., extraction of tweets from twitter for three months from Jan 2020 to march 2020. In the second section, the data cleaning process is explained along with obtaining the final dataset. The third section deals with the generation of the word cloud to check the frequency of the words. Further section includes the calculation of tf-idf which is a document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. Later in the sections, sentimental analysis was done using loughran negative, vader sentiment analyzer, textblob and unsupervised sentiment analysis using k-mean clustering and then correlation was evaluated between 'public sentiment' and 'stock market prices' using different models. For the first part, public sentiments were evaluated based on tf-idf and loughran negative.and linear regression model was constructed to check the correlation between the 'public sentiment' (i.e., negative sentiment) and the Dow stock market values. For the other part of the section, sentiment scores based on the vader sentiment intensity analyzer was obtained. A prediction model was created using random forest regression to evaluate correlation between sentiment scores and the Dow stock market values by running the training and testing set over the dataset. In the last section, the different created models were compared on the basis of performance and accuracy parameters and further conclusions were made on these observations.

II. LITERATURE CITED

During decades analyzing the stock market was just based on historical market prices. Autoregressive, moving average (Hellström, T. and Holmström, K.,1998), Genetic Algorithm (K. Kyong-jae, K., and Han, I., 2000), Neural Networks (Quah. T.S., and Srinivasan, B., 1999)and other techniques have been examined for analyzing stock market behavior. Sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. Sentiment analysis of Twitter messages has been used to study behavioral finance, specifically, the effect of sentiments driven from social media on financial and economic decisions. For example, Bollen and Pepe 2011 used social-media sentiment analysis to predict the size of markets, while Antenucci et al. 2014 used it to predict

unemployment rates over time. Twitter sentiment analysis in particular, is a challenging task because its text contains many misspelled words, abbreviation, grammatical errors, and made up words. Therefore, it contains limited contextual information. In previous research, it was implied that if it is properly modeled, Twitter can be used to forecast useful information about the market. Tharsis et al. used a Twitter sentiment analysis from (Kolchyna et al., 2015) which was SVM approach, then compared them to different industries and showed that by adding the sentiments to their predictive models, the error rate reduced between 1 to 3 percent, in predicting the Expected Returns of different industries (Souza et al., 2015). Alanyali et al. found a positive correlation between the number of mentions of a company in the Financial Times and the volume of its stock (Alanyali et al., 2013).

III. PROPOSED MODEL

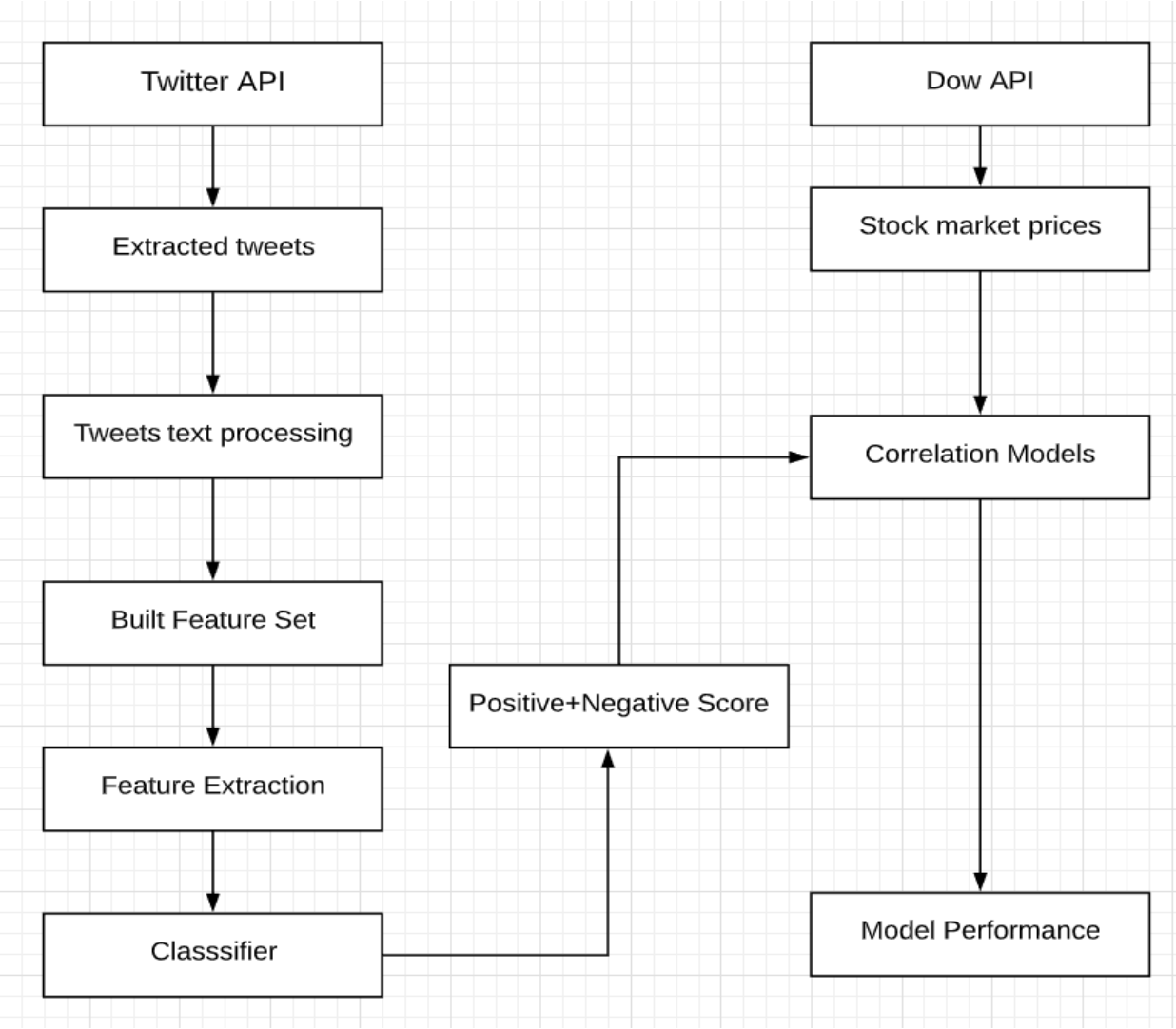

Figure 3.1: Proposed Research Model

IV. METHODOLOGY

4.1. Web Scraping

From the social media platform, twitter was selected to obtain data for text mining analysis as it is one of the biggest public platforms for exchanging thoughts and views of mass population round the globe.So, by keeping stock market as a keyword tweets were extracted using GetOldTweets library from January 1,2020 to March 31,2020. Only the top tweets for every week were extracted to have better knowledge about the current scenario. As a result, a dataset was obtained with four columns i.e., date, text, hashtags and link. Additionally, Dow values of the stock market were also extracted. Further this process is followed by data cleaning.

4.2. Data Cleaning

After obtaining the data set of four columns, different methods were used for data cleaning. So, we started with HTML decoding, as HTML encoding did not get converted to text, it ended up in the text field as '&','&quot', etc. We used BeautifulSoup to clean HTML encoding. Next we removed @mention as this information doesn't add value to build the sentiment analysis model. Further we removed URL links as this can be ignored for sentiment analysis so we decided to remove url links from text. Additionally we dealt with UTF-8 BOM (Byte Order Mark) where BOM was replaced with '?'. Hashtag information could provide some useful information about the tweet so removing it could have been risky so we decided to remove '#' and keep the text intact. However, we realized there were few more issues with the text. Issues:

- Negation words were split into two parts, and the 't' after the apostrophe vanished after cleaning. This made words like 'can't' end up as 'can'. This was risky for sentiment analysis.
- We realized some url links didn't start with 'http' but with 'www'. Our cleaning code failed to recognize special characters like '=', '_', '~', etc.

So, finally we added a few more codes to get rid of the above mentioned issues and to get better sentiment analysis results, following steps were done:

- Converted all text to lower-case
- Added code in order to handle negations.
- Removed numbers and special characters
- Removed stop words
- Tokenizing and converting into stem words using nltk porterstemmer.
- For Dow values, some of the weekend values were missing so we adopted the forward filling method to fill those values for further analysis.

4.3. Word Cloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

After the web scraping and data cleaning process, we proceed with the generation of word clouds for each month (January, February and March). First of all, creation of a mask from the image was done and then later the word clouds were generated. By generating word clouds, we got an overview of how words are distributed and what is the frequency of the word in each month. The purpose of generation of word clouds was to check whether we have enough words for conducting sentiment analysis for tweets or not.

To accomplish this job, we imported stop words and extended the list of stop words. We used libraries like pandas, word cloud and matplotlib. After that we converted all words to lowercase and took out the frequency of each word. In the next step we imported word cloud adjusted background color, height and width and generated three-word clouds.

4.4. Tf-idf Calculation

After obtaining the dataset with the extracted data, we clubbed the tweets of each day and then calculated the tf and idf of tweets of each day. From these values Tf-idf of each word was calculated. Later, this Tf-idf was utilized for the calculation of weighted sentiment using loughran negative.

4.5. Sentimental Analysis

The sentiment analysis was done using three different methods as mentioned below:

1. **Loughran negative with tf-idf:** For this analysis, the tf-idf was taken from the above calculations. Additionally, the loughran negative words file was used to evaluate the negative sentiment of the tweets. The calculation of weighted sentiment was done using loughran negative and tf-idf then further analysis was done.
2. **Textblob:** Textblob is one of the libraries which we came to know about through this course,so after hearing many good things about it such as part-of-speech tagging and sentiment analysis, we decided to give it a try to perform natural language processing tasks. So, we calculated the sentiment- polarity and subjectivity of each tweet of our dataset.
3. **Unsupervised sentiment analysis:** There are many machine learning approaches to handle sentiment analysis, but clustering is one of the unsupervised machine learning approaches that requires no labels while trying to group similar messages into a cluster. This cluster can be exploited

to find underlying sentiments toward an entity. K-means clustering along with the text blob was adopted to proceed in this analysis. Later, we plotted the elbow curve and on the basis of that curve number of clusters were selected for the analysis. Then based on the values of polarity and sentiment confidence, different clusters were plotted over the graph.

4. **Vader sentiment analysis:** VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of a sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative. We calculated the sentiment score of our data which includes positive, negative and neutral sentiment. Based on these sentiments, a pie-chart was created representing the percentage proportion of each sentiment.

4.6. Construction of models

After the completion of the sentimental analysis we proceed towards the construction of the models. Firstly, we created a correlation model using linear regression between public sentiment and stock market prices. For this, the weighted sentiment values calculated from the tf-idf and loughran negative file were used as public sentiment values(negative sentiment). Secondly, we created a correlation model using random forest regression. For this model, training and testing dataset was created and on these basis the predicted values of the stock market were evaluated. Later, the comparison was done between the actual and predicted stock market values.

4.7.Comparison of Models

In the last section, the comparison between the linear regression and random forest model was done based on the different parameters like accuracy, MSE, MAE and observations. Also, the actual values and predicted values of the stock market were compared. Further, the conclusion for the better performance of the models based on these parameters was made.

V. RESULTS AND DISCUSSION

5.1. Figures and Observations

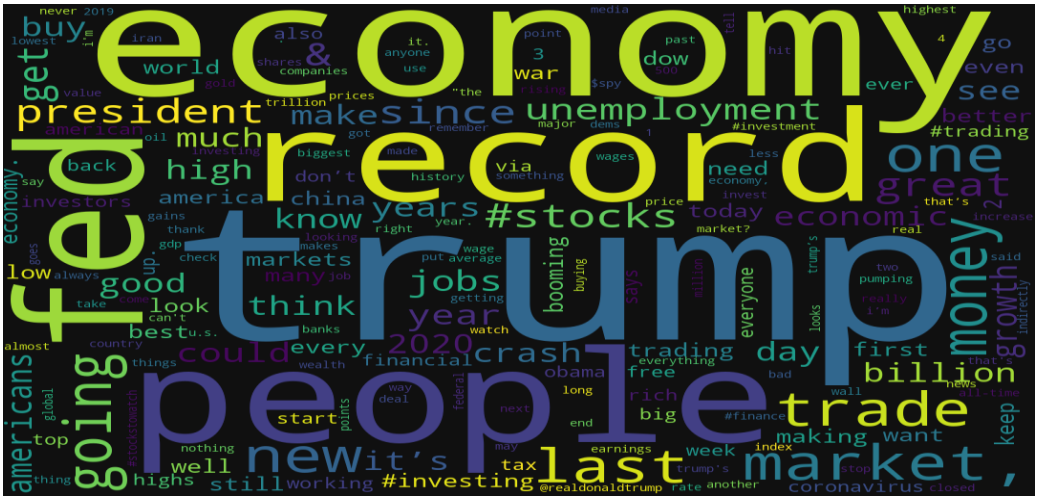

FIGURE 1. WORD CLOUD FOR JANUARY MONTH

Figure 1 represents the word cloud for the January month. From this image we can clearly see that the words "economy", "record", "people", "Trump" , " going" , "fed" are clearly visible, which shows these words are more influential in the January 2020. If we keenly observe this, we can also see words like money, high, billion, trade, booming are also visible. So, this reflects that the extracted tweets extracted for the January month are with positive words and positively influencing the stock market.

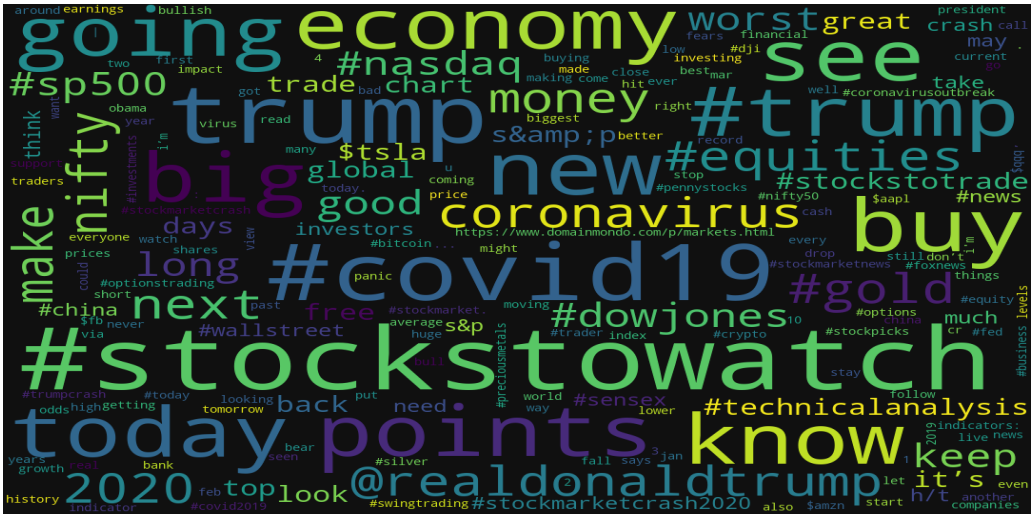

FIGURE 2. WORD CLOUD FOR FEBRUARY MONTH

Figure 2 represents the word cloud for the February month. From this image we can clearly see that the words "covid19", "stocks", "towatch", "Trump" , " points" , "today" are clearly visible, which shows these words are more influential in the February 2020. Here in this month the coivid19 was affecting human

activities round the world. So, according to the extracted tweets the stock market was also started affected by this pandemic.

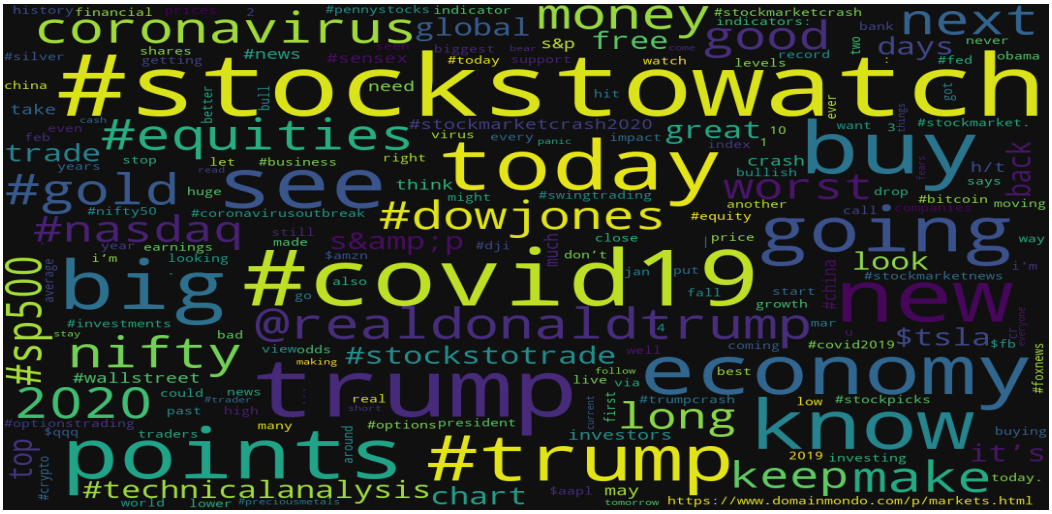

FIGURE 3. WORD CLOUD FOR MARCH MONTH

Figure 3 represents the word cloud for the March month. From this image we can clearly see that the words “covid19”, “stocks”, “towatch”, “Trump” , “ points” , “economy” are clearly visible, which shows these words are more influential in the March 2020. As the covid19 cases were increasing in this month, most frequent words related to this only. Here we can also see that the words like worst, crash, coronavirus were also visible which shows the stock market was affected by this pandemic by this time.

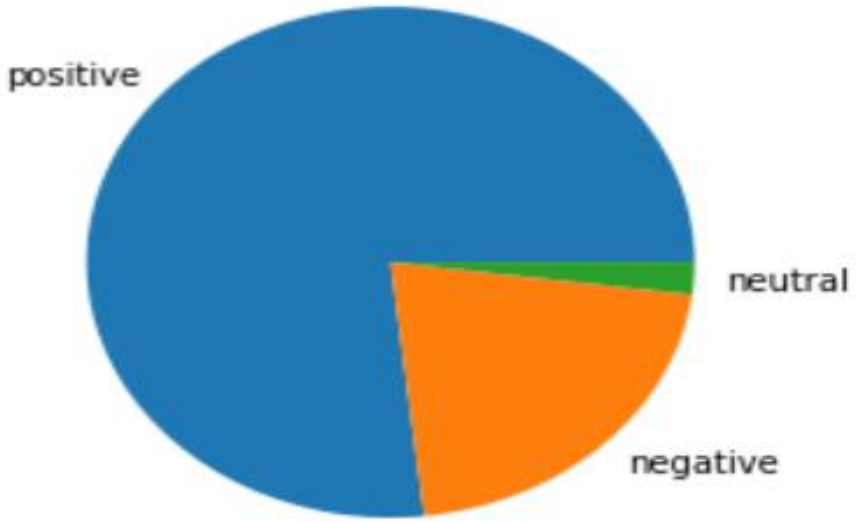

FIGURE 4. PIE CHART REPRESENTING PERCENTAGE OF POSITIVE, NEUTRAL AND NEGATIVE SENTIMENT

Figure 4 represents the pie chart showing the positive and negative sentiment percentage. These sentiment scores were calculated by the vader sentiment analysis. Here the positive sentiment was found to be 71 percent, the negative sentiment was to be 21 percent and the neutral sentiment was 2 percent. From this observation we can see there were more positive sentiments because for the first two months the stock market progress was good and later in February, the arrival of coronavirus may affect the public sentiment with negativity. So, we can say that Covid19 might have played a major role in affecting public sentiment with negativity and eventually the stock market.

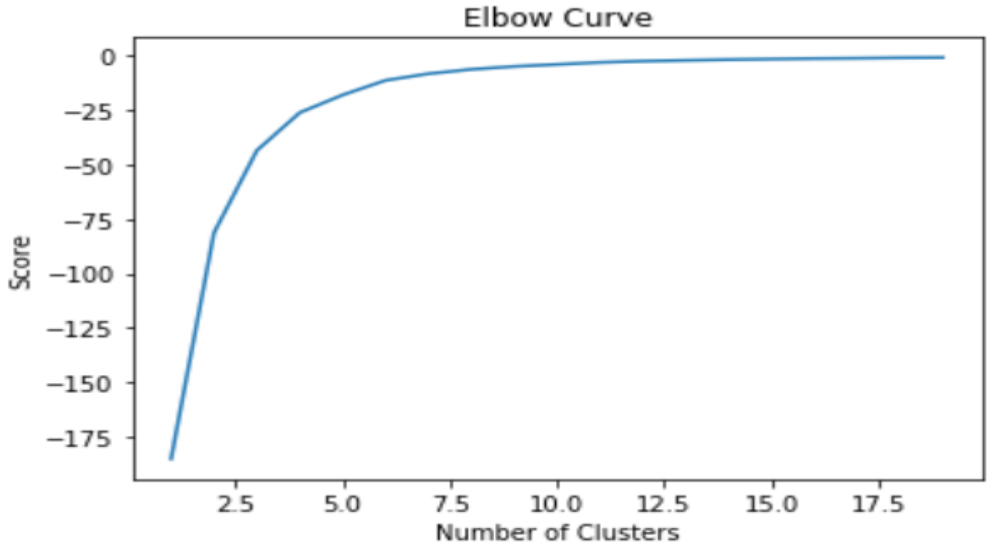

FIGURE 5. ELBOW CURVE REPRESENTING THE NUMBER OF CLUSTERS AND SCORE

Figure 5 represents the elbow curve with a number of clusters and scores as its x-axis and y-axis. This elbow curve is one of the parts of unsupervised sentiment analysis and was plotted to determine the number of clusters in the dataset. K-means clustering was adopted for evaluation and to get meaningful data about the cluster formation,and picked the elbow of the curve as the number of clusters to use. From this curve, we found that the number of clusters was five.

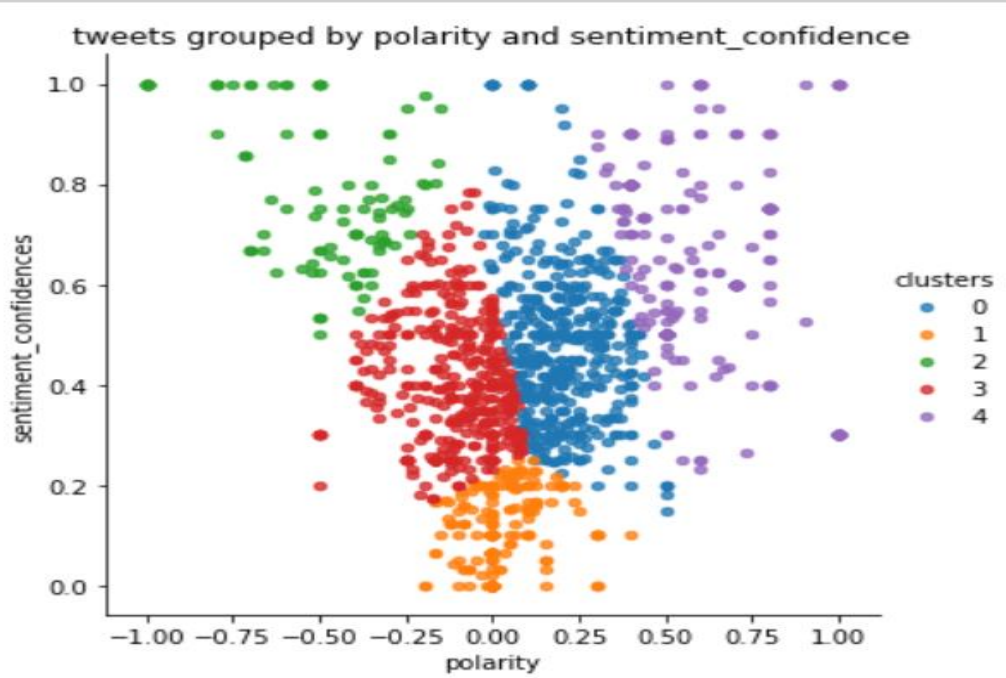

FIGURE 6. CLUSTERS BASED ON POLARITY AND SENTIMENT CONFIDENCE

Figure 6 represents the scatter plot for the clusters based on polarity and sentiment confidence. Here we can see that the dense clusters are between -0.25 to +0.25 polarity. Clusters beyond the -0.50 polarity with green color represent the negative sentiment of the population whereas clusters beyond the +0.50 polarity with purple color represent the positive sentiment of the population. The cluster grouping shown between green and purple clusters represents the neutral sentiment.

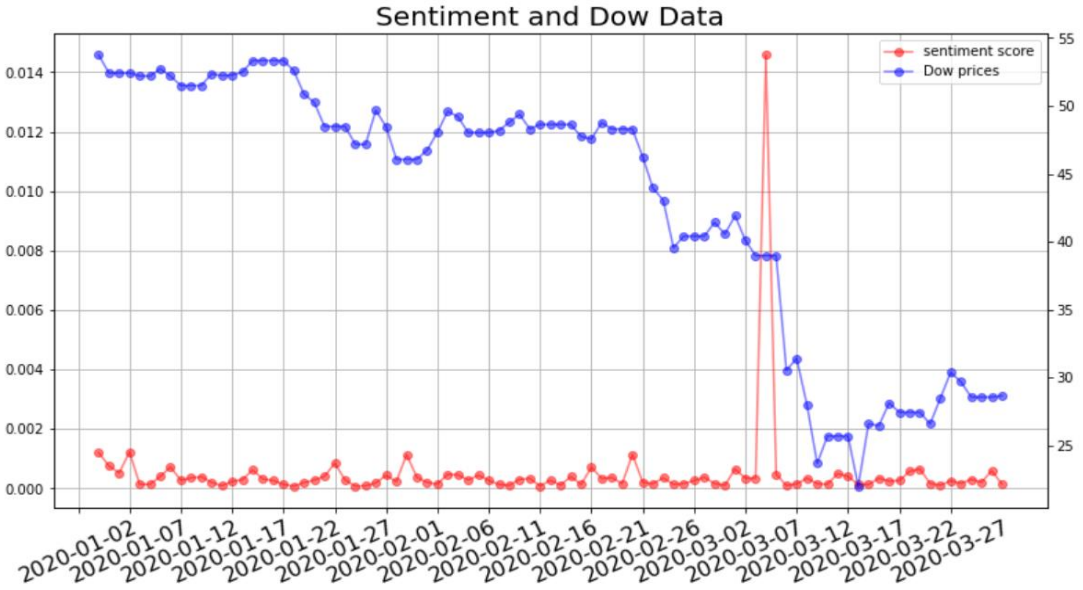

FIGURE 7.1. LINE GRAPH REPRESENTING THE COMPARISON BETWEEN THE SENTIMENT SCORE AND DOW VALUES

Figure 7.1 represents the line graph representing the comparison between the Dow values and sentiment score. In the above plot the dow data and the normalized negative score is plotted with date on x-axis.

It is clearly visible that there is a high negative score between the date range March 2nd 2020 and March 7th 2020 One of the sentiments with a high score.Then it can be observed a trend that the stock price has a sharp decline between March 7th 2020 to March 12th 2020.The First official death due to Covid-19 was declared on March 1st in Washington state,that might have impacted high negative sentiment score between 2nd of March to 7th of March.

the high negative score has a dominant effect over the graph scale so for better understanding and visualization this outlier was dropped.

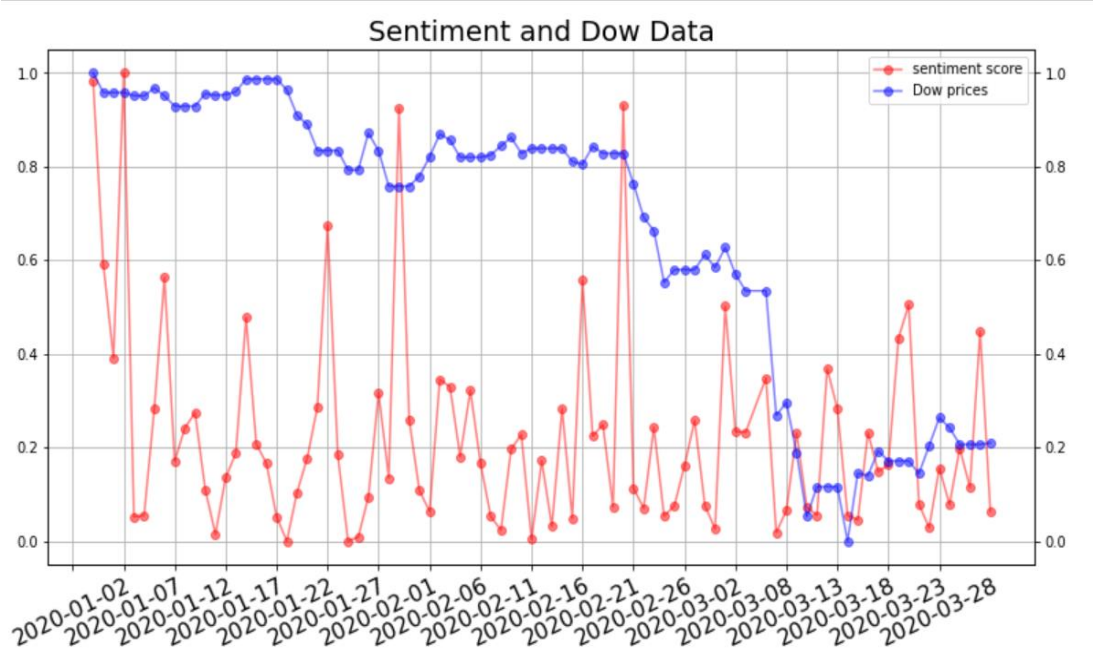

FIGURE 7.2. LINE GRAPH REPRESENTING THE COMPARISON BETWEEN THE SENTIMENT SCORE AND DOW VALUES WITHOUT AN OUTLIER

Figure 7.2 represents the line graph representing the comparison between the Dow values and sentiment score after dropping the outlier. In this figure, the Dow values and the normalized negative score is plotted with the date on X axis.It can be observed that when the negative score is high,there is a decline trend in DOW value not on the same day rather it can be visible in the next few days.

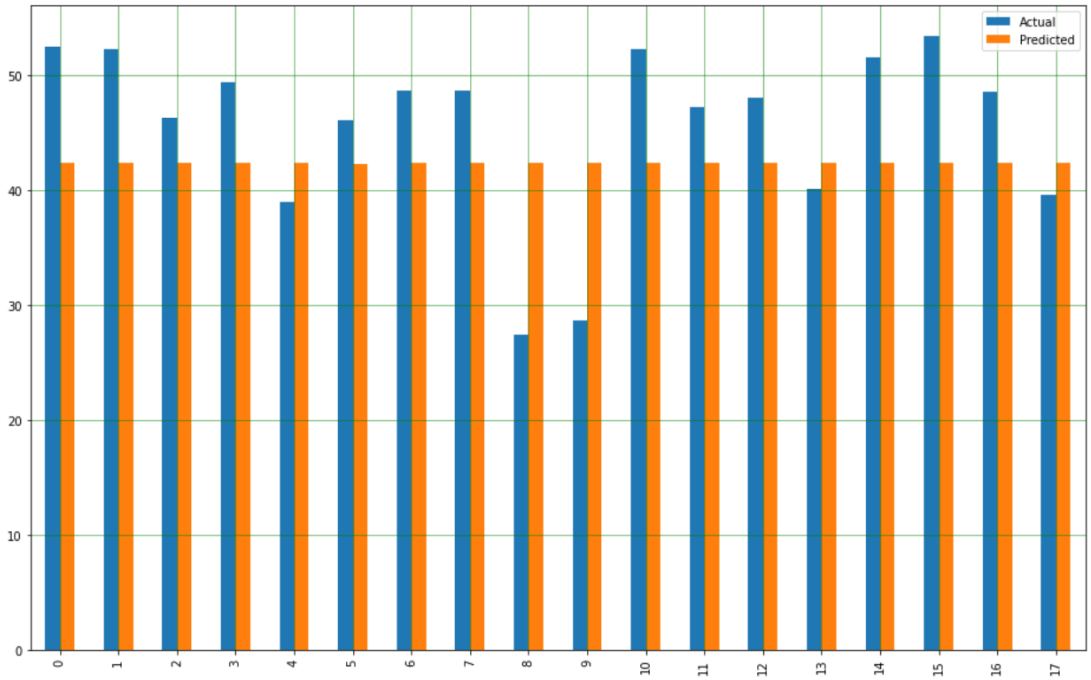

FIGURE 8. BAR GRAPH SHOWING ACTUAL AND PREDICTED PRICES BASED ON LINEAR REGRESSION MODEL

Figure 8 represents the actual and predicted Dow prices based on the linear regression model. Data set of 88 that contains the negative sentiment score and Dow Price records are randomly splitted into 80:20 train andTest Data.The above figure shows the actual and predicted Dow prices for 17 days.For Prediction of Test data the MAE is around 7.27.

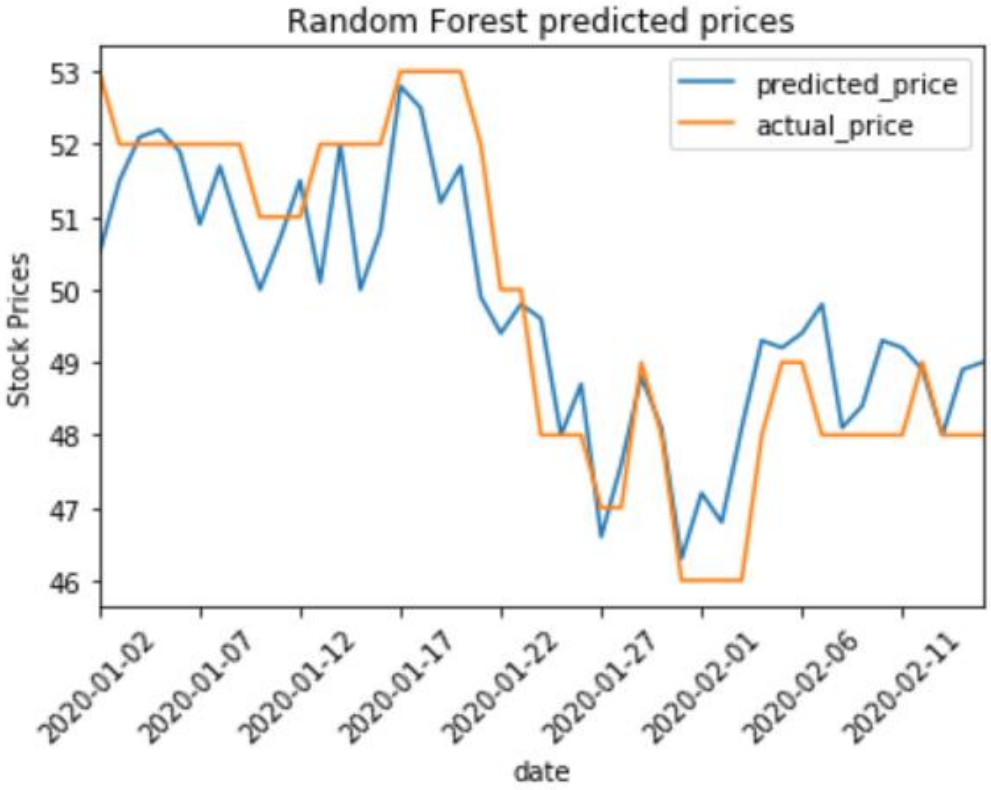

FIGURE 9. LINE GRAPH SHOWING ACTUAL AND PREDICTED PRICES OF TRAINING SET BASED ON RANDOM FOREST REGRESSION MODEL

Figure 9 represents the line graph representing the actual and predicted prices of training set based on the random forest regression model. For the training set we considered the sentiment values and stock prices from the start of January to mid-February and we observed that the predicted prices were close to the actual price in our training set. According to the data, the stock market prices were less fluctuating and were at the higher side and the same prediction was made by our sentiment analysis.

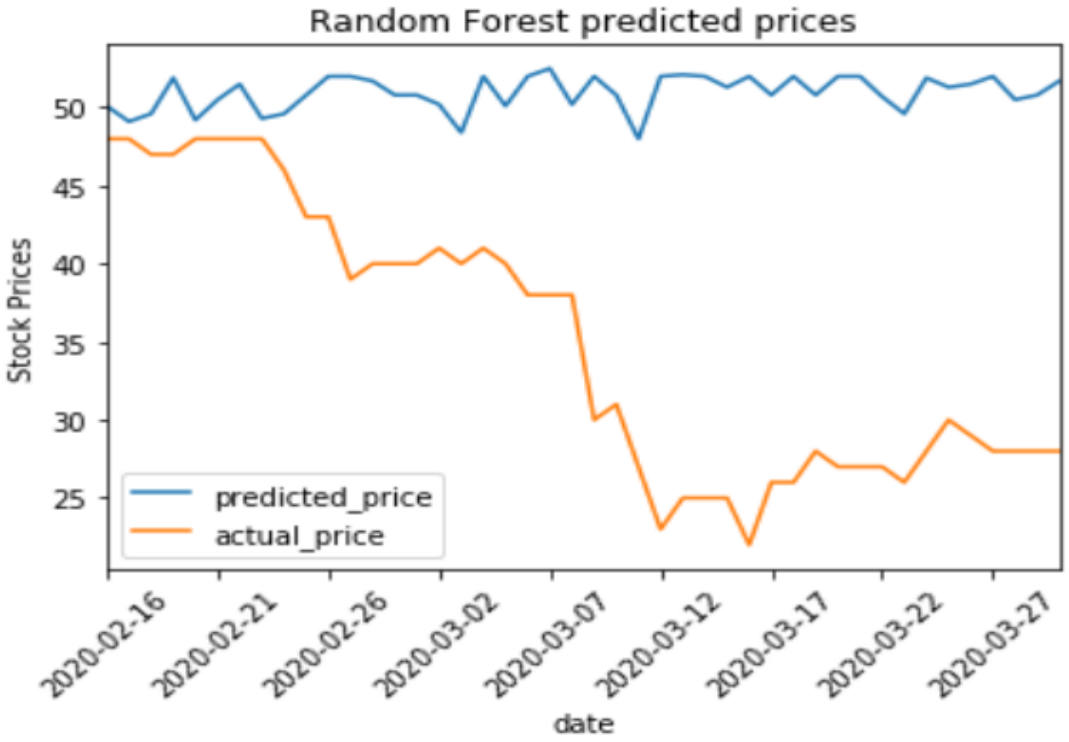

FIGURE 10. LINE GRAPH SHOWING ACTUAL AND PREDICTED PRICES OF TESTING SET BASED ON RANDOM FOREST REGRESSION MODEL

Figure 10 represents the line graph representing the actual and predicted prices of testing set based on the random forest regression model. For the testing set we considered the sentiment values and stock prices from mid-February to the end of March. By observing the above graph, we have seen that the predicted values were less fluctuating whereas the actual stock price is dropping rapidly. So, as per the training set provided, the stock market prices should not show high fluctuation and it should be at the high scale as according to January market prices. But due to the pandemic, there was a crash in the market at the end of February. So, from this graph it is clearly visible that there is a very huge difference between predicted and actual market prices. This may be because of the coronavirus outbreak all over the globe.

VI. EQUATIONS

For this project we used the below equations for the calculations:

6.1 Tf-idf calculation

For a term t in a document d, the weight Wt,d of term t in document d is given by:

Wt,d = TFt,d *log (N/DFt)

Where:

- TFt,d is the number of occurrences of t in document d.
- DFt is the number of documents containing the term t.
- N is the total number of documents in the corpus.

Once TF-IDF is calculated for each word in a sentence the first negative sentiment of the sentence is calculated using loghurn Negative word File .If any word of the of sentence of dataset is present in loghurn Sentiment file the Tf-IDF of that word is multiplied with 1 else it is multiplied with zero.

For example :

For I hate dog : negative sentiment of sentence is calculated by:

Tf-Idf of(I)*0+Tf-Idf of(hate) *1+Tf-idf of (dog)*0 as here only hate has negative sentiment

Once we calculated the negative sentiment score of the sentence, the score is normalized by dividing the score with the total number of words present in the sentence.

VII. CONCLUSION

We observed that the Mean absolute error,Mean square error and root mean square error for linear regression model was found to be 7.27,66.02 and 8.12 respectively, whereas for random forest regression model the values of mean absolute error and accuracy for training set is 0.9 and 98.17 and for testing set is 15.53 and 46.57. As we can clearly see that the mean absolute value for random forest regression model is high which means the average magnitude of the errors in a set of predictions is high. So, concludingly we can say that the linear regression model was better than the random forest regression model for our dataset. Therefore, Loughran sentiment with Tf-idf was found to be a better sentiment analysis technique than vader sentiment analysis for our dataset.

FUTURE ASPECTS

Finally, it's worth mentioning that our analysis doesn't take into account many factors. Firstly, our dataset doesn't really map the real public sentiment, it only considers twitter using, english speaking people. It's possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people's mood indeed affects their investment decisions, hence the correlation. But in that case, there's no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation - investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment. All these remain as areas of future research. The dow data and sentimental analysis can be considered as time series data and further analysis can be done.

Acknowledgment

This project would not have been possible without the guidance of Dr. Peng Xie, the instructor of this course (BAN675-Text Mining). Additionally, thanks to our family and friends for encouraging us for new research. Lastly, thanks to our team members for introducing and finalizing this area of research.

VIII. REFERENCES

Aggrawal, M., Oh, O. and Rao, H.R., "Community intelligence and social media services: a rumor theoretic analysis of tweets during social crises'", *MIS Quarterly,* 37(2),2013,407-426.

Alanyali, M., Susannah, H.M.,, and Preis,T., "Quantifying the relationship between financial news and the stock market", *Scientific reports,* 3:3578,2013.

Antenucci, et.al., "Using Social Media to Measure Labor Market Flows", *Nber,* 2014.

Bollen, J., Mao, H., and Zing, X.J., "Twitter mood predicts the stock market predictor".,*IEEE Computer,* 44(10), 2010, 91–94.

Bollen, J., and Pepe,A., "Modeling Public Mood and Emotion : Twitter Sentiment and SocioEconomic" Phenomena, 2011,450–453.

Hellström, T. and Holmström, K.,"Predicting the Stock Market," Technical Report Series IMATOM,1997-07,1998.

Kolchyna, et.al., "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination", 2015,32.

Kyong-jae, K., and Han, I., "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index", Expert Systems with Applications, 19(8), 2000,125-132.

Mittal, A. and Goel,A., “Stock prediction using twitter sentiment analysis”. 2012.

Pak, A. and Paroubek, P., “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, *European Language Resources Association (*ELRA), 2010.

Pandey, P., “Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)”,Analytics Vidya,September23,2018,https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f (accessed April 6, 2020).

Quah, T.S., and Srinivasan, B., “Improving Returns on Stock Investment through Neural Network Selection”, Expert Syst. Appl., 17, 1999, 295-301.

Ranco, et.al.,“The Effects of Twitter Sentiment on Stock Price Returns”, *PLoS ONE,* 10(9), 2015.

Raschka, S., *Python Machine Learning,* PACKT publishing Ltd., UK, 2015

Sarkar, T.,“Clustering metrics better than the elbow method”, Towards data science, September 6, 2019, https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6 (accessed April 13, 2020).

Tharsis,T.P.S., Kolchyna, O., and Aste,T., “Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry”,(i):19,2015.

Wojcik, R.,”Unsupervised Sentiment Analysis: How to extract sentiment from the data without any labels”,Towards data science, November, 26, 2019, https://towardsdatascience.com/unsupervised-sentiment-analysis-a38bf1906483 (Accessed April 15, 2020)

APPENDIX

File name	Content
README.md	Description about the project.
Text_Mining_Project_Master File.ipynb	Code for data extraction and data cleaning.
clean_tweet_data_master.xlsx	Final clean twitter dataset.
wordcloud.ipynb	Code for creating word cloud, in order to highlight the most frequently used keywords in the corpora.
proj_getDow.ipynb	Code for extracting DOW dataset from Yahoo Finance.
DOW.xlsx	DOW dataset.
DOW_StockPrediction_analysis.ipynb	Code for twitter sentiment analysis(VADER and Unsupervised), DOW stock prediction using Random Forest Regressor, and model accuracy.
ccdata.xlsx	Excel file after merging of tweets of particular date
TFIDF_Analysis_withSentiment.ipynb	Code for twitter sentiment analysis, DOW stock prediction using TFIDF, and model accuracy.

Github link for the jupyter notebook files used for the analysis:

https://github.com/maitreyee19/TextMining_project