Journal of Money and Economy
Vol. 15, No. 3, Summer 2020
pp. 235-251
DOI: 10.29252/jme.15.3.251

Original Research Article

Forecasting Stock Price Movements Based on Opinion Mining and Sentiment Analysis: An Application of Support Vector Machine and Twitter Data

Babak Sohrabi[*] **Ahmad Khalili Jafarabad**[†]
Ardalan Hadizadeh[‡]

Received: 29 Feb 2020 Approved: 6 Feb 2021

Today, social media networks are fast and dynamic communication intermediaries that are vital business tools, as well. This study aims to examine the views of those who are involved in Facebook stocks to understand the pattern and opinion about the intended future stock price. Yet another goal of this paper is to create a more accurate forecasting pattern compared to the previous ones. Two datasets are used in this paper; the first contains 1.6 million tweets that have already been emotionally tagged, and the second has all the tweets about Facebook stock in eighty days. We conclude that positive news about a company excites people to have definite opinions about it, which results in encouraging them to buy or keep that specific stock. Also, some news can hurt users' views as most of the time, things get more complicated, and uncertainties make it harder to forecast the direction of stock movement. By using text mining and python programming language, we could create a system to be operable in those situations.

Keywords: Social Networking, Stock Prediction, Group Emotion, Collective Emotion, Sentiment Analysis, Opinion Mining, Neural Network**.**
JEL Classification*:* C53, M15

1 Introduction

Social networks are a means of communication people mainly use to share their information. However, communication is not the only application of these networks; instead, data extracted from these platforms could play a

[*] Faculty of Management, University of Tehran, Tehran, Iran; bsohrabi@ut.ac.ir (Corresponding Author)
[†] Faculty of Management, University of Tehran, Tehran, Iran; akhalili@ut.ac.ir
[‡] Faculty of Management, University of Tehran, Tehran, Iran; ardalan.hadizadeh@ut.ac.ir

significant role in different businesses. One of the first theories about the stock price prediction was the efficient-market hypothesis (EMH), alternatively known as the efficient market theory, a hypothesis indicating that share prices reflect all information. According to the EMH, stocks always trade at their fair value on exchanges, making it impossible for investors to purchase undervalued stocks or sell them for inflated prices (Timmermann & Granger, 2004), (Marwala & Hurwitz, 2017). Therefore, it should be impossible to outperform the overall market through expert stock selection or market timing, and the only way an investor can obtain higher returns is by purchasing riskier investments (Fama et al., 1969). In other words, the efficient market is the one that adapts quickly to new information, and the efficient market hypothesis indicates that most of the time, the stock market price is random and cannot be predicted.

But behavioral economics believes that positive and negative views of a community about a specific stock can have a significant effect on its price. One of the most popular resources are the news headlines, but the critical question is whether the headlines are good enough or not? It should be noted that the headlines are not necessarily the people's opinion. Different people can have different understandings of a specific headline, and also these headlines can be manipulated by the governments and the journalists (Q. Li et al., 2014).

In this study, we are going to perform fundamental analysis on the Facebook stock. This paper examines the opinions of those involved with Facebook stocks so that we can summarize their mindsets to predict the general behavior of this community of individuals so that we can collectively consider possible price movements. As such, the attempt will be to predict the Facebook stock price and create a more accurate pattern than the previous ones.

Our goal is to create a system that can predict the Facebook stock price in advance by analyzing the user's comments on Twitter. However, to use the boundless social network data, an appropriate algorithm is required to understand the most precise state of people's emotions and examine the existing sentiment with stock changes as well as find the right relationship between them. As such, they are selected by their accuracy and suitability; hence, our algorithm is based on a word-to-vector conversion. Under this process, it first converts each word into a vector and then puts them into a multidimensional space, then measures the distances between them using support vector machines. After transforming the text into numbers, it can predict the usage of a word. Also, after training a classifier, it will be able to

tag every sentences feed into it emotionally (Zainuddin & Selamat, 2014), (D. & Gore, 2016).

Using the Bernoulli classifier and an emotionally tagged dataset will enable us to create the classifier mentioned above which helps learn the feelings within a tweet by observing so many tagged tweets (Juan & Vidal, 2002). In this way, business owners can know the opinion of the customers and the collective vision of people about the value and status of their businesses. They can also serve customers better by understanding significant problems related to a product or service and identifying critical issues that can cause significant damage to the business.

The outline of the paper is as follows. Section 2 discusses the previous works on this subject and also talks about different tools that were made to analyze text. Section 3 discusses the method that we made by studying the previous methods and talks about the neural network algorithm that we use. Section 4 is where we talk about our findings and the way we use them to predict the price movements. Section 5 is where we talk about the limitations and our conclusions.

2 Theoretical Framework and Research Background

Coming to the point that sentiments can be analyzed in the texts, the results acquired during the study led us to many possibilities. During the course, various forums were surveyed, including the ever-expanding Internet data that could not be possible without a mechanism due to its length. For that matter, the sentiment analysis got prominence to classify the existing data even in high volumes. The following are some of the methods applied by previous researchers to analyze people's sentiments about stocks.

Schumaker, Zhang, Huang & Chen (2012) developed a system called AZFinText using a combination of text analysis and stock changes as well as basing them on sentiment and economic change process analyses. This system focused on subjective texts rather than the objective ones, but the result was very significant. 53.5 % of the cases with positive comments saw a price drop, and 52% of the cases with negative opinions showed a rise in stock prices. It concluded that it was good to sell upon finding positive comments and buy in the case of negative ones. Of course, this study lacked sufficient accuracy, and one could only make correct predictions with a probability close to 50%. One thing which should not be kept secret is that in social networks, one can also see the views and opinions of other users about a particular text, but in this kind of system, there was no attention to the user's feedback.

In 2014, to understand people's opinions around a particular news headline, their writings on different social media were collected and applying an algorithm called Bag of Words (Wallach, 2006), (Y. Zhang et al., 2010), researchers tried to understand the emotion of each written words. In Bag of Words algorithm, the researchers were more concerned about things like the meaning of each word and how many times it was used. According to Tetlock (2008), people tend to be optimistic or pessimistic about a news topic, which leads to increased skepticism about a stock market change. Still, this forecast is not sufficiently accurate with a mere headline review. The next step in this kind of work is using psychological dictionaries. Harvard-IV-4 Psychological Dictionary is a matrix that holds essential words with their tagged emotions as well as it keeps each word's strength. As we can see in different papers, words do not have the same effect and power. For example, bad and terrible are not the same! As we move on, we get to know Lexicon, a language's inventory of lexemes or lexical items, or word forms. Lexemes are not atomic elements but contain both phonological and morphological components. When describing the Lexicon, a reductionist approach is used, which tries to remain general while using minimal description. To specify the size of a lexicon, lexemes are grouped into lemmas, which are generated by inflectional morphology. In other words, lemmas are represented in dictionaries by headwords which list the citation and any irregular forms, since these must be learned to use the words correctly. Lexemes coming out of a word by derivational morphology are considered new lemmas. The Lexicon is also organized according to open and closed categories where the first categories, such as determiners or pronouns, rarely give new lexemes, and their functions are primarily syntactic. In contrast, the open categories, such as nouns and verbs, have highly active generation mechanisms, and their lexemes are more semantic in nature.

While previous researches did not pay enough attention to the influence of social media on different aspects of social life, with the expansion of the data engagement system, it gradually emerged as a good source for understanding people's feelings about any subject or commodity. It should be noted that the current state of society influence the stock market and what environment could better than social networks to access it (Patil & Atique, 2015).

The use of word analysis, or Lexicon, is a key to sentiment analysis, and its implementation helps carry out the analysis effectively and obtain acceptable results. This method can be used in two ways: analyzing each situation through experts or machine automatically. For instance, researchers came up with an automated system that could extract attributes from data of

Wall Street Journal and create a learning system by tagging them, then compared the existing attributes with new ones by dividing them into positive or negative categories.

Initially, using Lexicon to check for new words followed the simplest case or its closest synonym and family. Over time, more sophisticated algorithms were added to accommodate the appropriate tag so that all words could be reviewed in one paragraph, and the most repeated tag would determine the emotion of the text (Mohammad et al., 2009).

In a 2017 study, it was found that investors were sub-magnified by technical and non-technical analysis, even though technical investors acted on a general basis with graphs and numbers. But they did not necessarily help them influence society.

And compared to a person who invested non-technically but had more followers in virtual networks, it had a greater impact on connecting with other users in virtual space, and even more for the right investment. So, we conclude that people's profiles are also very important, and more followers mean a higher impact in this area (Oliveira et al., 2017).

Given the importance of predicting stock prices, many efforts have been made over time to examine the relationship between investor's sentiment and price changes. Still, the results have never been the same and a view has been developed that, contrary to all the results obtained so far, the relationship between these two may not be linear.

Therefore, using Granger's causality study from a non-linear perspective, it is true that the most important information obtained from the stock market is its daily price. Still, it is not the only factor affecting the future price, so the data created on social networking sites to explore other dimensions (You et al., 2017). We also want to design a system that examines both the vocabulary and the whole sentence together to make a system that understands emotions within social media activities.

3 Research Methods and Variables

As of the classification of research tools, the present study is based on descriptive and correlation methods.

Due to Twitter's restriction on access to outdated data and third-party apps, we preferred NASDAQ databases, which were ready to use and thus made some data mining attempts on them. Among the advantages of this data is a brief report on top tweets of the day and users' preferences and interests in them. The daily stock price data is also taken from Yahoo Finance using the Facebook stock trademark.

In this study, we needed two statistical samples, the first a large dataset containing a variety of tweets, each with an emotional tag. In other words, it required a dataset of tweets that had previously been analyzed sentimentally by a trusted human consciousness or a machine trained to do so, like the one we are trying to achieve. The next step is gathering tweets about the company in our eighty-day timeframe.

We used the Multiple Linear Regression to numerically predict the stock movement. Then, we tried to predict the next day's stock price by considering our factors with the following formula. These factors include the number of positive and negative tweets, the power of each tweet by analyzing the word strength and popular retweets of each day.

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \;\; i = 1, \ldots, n$$

This method evaluates variables in an environment similar to reality but does not identify the cause and effect relationships. In the multiple linear regression method, the parameters of a linear model are estimated using a target function, and the value of the variables. So, if we have n observations of the independent variable X, which has p dimensions, and we want to establish a linear relationship with the response variable y, we can use the above linear regression model.

Since the independent variable X has p dimensions, we have replaced its value in each dimension with a one-dimensional independent variable. The index i also shows the number of observations. Finally, ε is the error of the regression model. But we do not need to compute all of that without the help of our computers. A linear regression model in two dimensions is a straight line; in three dimensions, it is a plane, and in more than three dimensions, a hyperplane. We use the regression to predict the next day's price. To do that, we are going to use Python to do that for us using the Scikit-Learn library, which is very easy to use. You just have to enter your data and give it a few samples about our numbers, which means it can understand the value of a variable using previous examples and the values of the current variables that you feed into it. We will use variables such as stock day price, number of positive tweets, number of negative tweets, and coefficients of influence (ranging from -10 to +10), and we will get the predicted price as a result. The closer the price to the actual price of the next day, the better the system is designed.

To get the best results, we have to review our input data and reduce its problems to minimize the noise within the selected text. In this section, we

want to remove duplicates, errors, incorrect characters, and emojis. We used the Beautiful soup and Re library in Python's programming language as well as the Notepad ++ software features to clear the data and get it ready to be used.

In the training phase, we separate each tag from the corresponding tweet and store it in two arrays and then train the system using Gensim, NLTK, Pandas, and Sklearn libraries in a Python programming language. In the Gensim library, we specify the vector of each tweet using the Word2Vec algorithm and the CBOW (support vector machine-based) method.

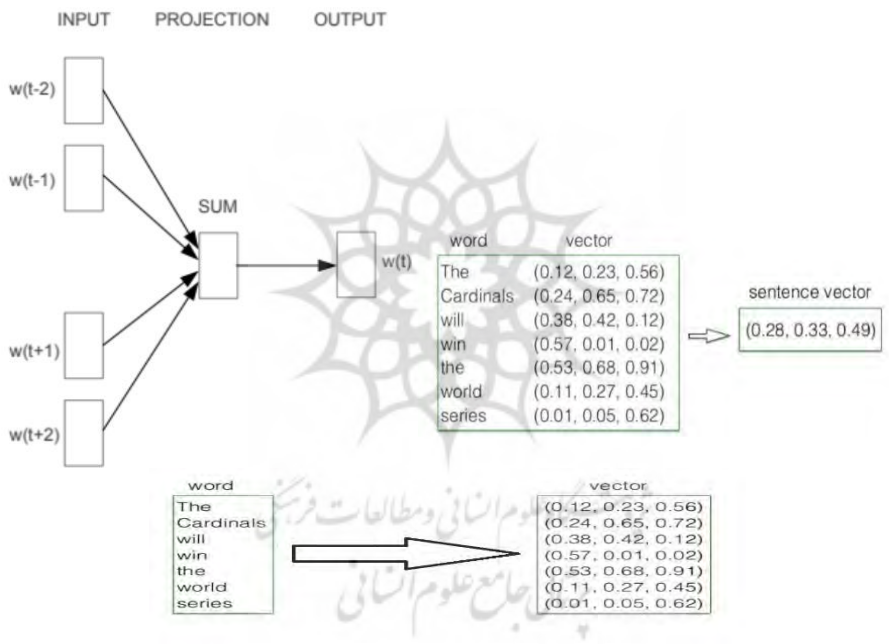

Figure 1. How to Assign Vector to Words and Sentences.

This model was created by Google in 2013 and is a predictive deep learning-based model to compute and generate high quality, distributed and continuous dense vector representations of words, which capture contextual and semantic similarity. Essentially these algorithms create their own vocabulary set that can predict a missing word in a text document. By using this ability, it can make a precise vector out of a sentence. Usually, we can specify the size of the word embedding vectors, and the total number of

vectors is essentially the size of the vocabulary. It makes a lower-dimensional model than the traditional Bag of Words. With this model, for instance, we can guess a word by having other words around it, and also, we can understand its feelings using our train data set and vectorized sentences.

After preprocessing the data and converting the words and sentences into vectors, we train the classification system using the Bernoulli algorithm to classify vectors by presenting them with aesthetic tags. Bernoulli, therefore, can be the most powerful method for predicting binary data.

Using a ready-made set of high-powered vocabulary words, they are examined daily, and their impact on prediction is significant. The reason for using this method is the lack of a semantically neutral sentence identifier, and almost the majority of such sentences fall into the positive semantic category. According to the theories associated with the sent message type, it has a high impact on its acceptance rate and the use of more robust vocabulary can have a more significant impact on the recipients and by using this section any positive or negative tweets is given a specific weight value, and finally concludes the whole day by comparing the weighted value of the positive or negative sum.

So, in this step, we also divide the number of powerful words based on their positive or negative sentiments, intending to specify this number for each day.

This time, we use the lexicon method to understand the emotions in the text. For that matter, first, we need a dictionary containing a number of commonly used words with positive and negative semantics. The Harvard-IV4 Dictionary has been chosen thus, which contains 1,045 highly influential words with a positive semantic impact. Also, its negative vocabulary has 1160 words, which have a greater impact on the recipients. Therefore, the number of these words is examined each day and multiplied by the number of tweets generated. For example, there are 250 positive and 100 negative tweets for a day, but a combination of their highly impacted words probably makes up 800 positives versus 200 negative words, so the calculation is as follows:

Power of positive words = Positive words divided by the sum of powerful words

In the example above, this number will be 0.2 and the power of negative vocabulary at 0.8.

Now we are doing the weighting done in number:

Positive score = 250 * 0.2 = 50
Negative score = 100 * 0.8 = 80
250 => 50 100 => 80
250-100=150 => 50-80=-30

So, in the first place, the review of the day was considered as a positive one, but with respect to the strength of the emotion within the words, it affects our findings, and the emotional burden of the day is going to change.

4 Research Results

Our first finding following a sentiment analysis of the tweets from NASDAQ dataset with $FB tag showed a number of positive and negative daily tweets. The second finding was the highest daily retweets contained in the data set itself, followed by the sentiment analysis based on the number of likes, comments, and retweets (-10 for the most popular negative tweets and 10 for the most popular positive ones). We also maintained daily trading volumes to examine the impact of users' comments on the stock market's volume. Then, we predicted the values using a regression formula.

When we compared the regression results with the real price of the following day, it was found that in 91% of the cases, the difference of predicted price with the actual one was less than one dollar. This difference in 46% of the cases was below 50 cents, as indicated in the following table.

Table 1
Numbers obtained by sentiment analysis of tweets along with predicted prices

Date	Close	Volume	Positive	Negative	Power Nodes	Prediction
2-Apr-2016	114.3	NA	181	30	-3	113.18
3-Apr-2016	113.4	NA	246	58	-4	112.32
4-Apr-2016	112.6	48487900	789	267	-1	112.7
5-Apr-2016	112.2	22962400	630	181	5	112.8
6-Apr-2016	113.7	20814600	760	270	-1	113.3
7-Apr-2016	113.6	20714500	985	233	-10	111
8-Apr-2016	110.6	48497800	1200	279	-4	109.7
9-Apr-2016	109.8	NA	330	83	1	109.44
10-Apr-2016	109.4	NA	349	68	-2	109.1
11-Apr-2016	109	39762300	1095	298	6	110.1
12-Apr-2016	110.6	26248100	1434	354	1	110.5
13-Apr-2016	110.7	88227400	1883	721	1	110.6
14-Apr-2016	110.8	28473300	1176	278	-5	110
15-Apr-2016	109.6	20922800	954	345	2	109.9
16-Apr-2016	110	NA	274	68	1	110

Date	Close	Volume	Positive	Negative	Power Nodes	Prediction
17-Apr-2016	110.2	NA	306	49	1	110.4
18-Apr-2016	110.4	21092700	766	202	8	111.8
19-Apr-2016	112.3	30210500	893	222	0	112.2
20-Apr-2016	112.4	21027900	744	152	5	113
21-Apr-2016	113.4	20875200	702	211	-8	111.8
22-Apr-2016	110.6	38458200	806	206	-1	110.17
23-Apr-2016	110.3	NA	295	62	1	110.11
24-Apr-2016	110.2	NA	301	60	-1	110.3
25-Apr-2016	110.1	21017900	668	197	-7	109
26-Apr-2016	108.8	22521500	1006	139	5	109.7
27-Apr-2016	108.9	52213100	2906	421	10	111.4
28-Apr-2016	116.7	87110100	3198	758	4	117.5
29-Apr-2016	117.6	37140600	1134	293	3	117.4
30-Apr-2016	118.1	NA	429	84	1	117.4
1-May-2016	118.3	NA	345	66	1	117.6
2-May-2016	118.6	28095200	798	224	-6	116.7
3-May-2016	117.4	24117500	704	169	3	117.1
4-May-2016	118.1	23448400	797	144	-1	117
5-May-2016	117.8	22056700	836	151	9	118.5
6-May-2016	119.5	26216200	643	223	-1	118.5
7-May-2016	119.4	NA	221	83	-0.5	118.3
8-May-2016	119.3	NA	206	57	-0.5	118.3
9-May-2016	119.2	21138100	800	198	4	119.5
10-May-2016	120.5	23220000	720	250	-5	119
11-May-2016	119.5	22038400	884	162	4	119.23
12-May-2016	120.3	22035500	582	125	-2	118.9
13-May-2016	119.8	18124300	732	167	-3	118.5
14-May-2016	119.2	NA	294	85	-1.5	118
15-May-2016	119	NA	301	71	-1.5	117.7
16-May-2016	118.7	31247800	795	193	-3	117.2
17-May-2016	117.3	21328600	739	145	1	116.7
18-May-2016	117.7	21642300	752	164	-4	116.3
19-May-2016	116.8	20544100	694	150	3	116.5
20-May-2016	117.3	18944800	501	124	1	115.8
21-May-2016	116.7	NA	283	64	-2	115.58
22-May-2016	116.3	NA	319	91	-2	115.3
23-May-2016	116	20441000	697	189	9	116.8
24-May-2016	117.7	20183600	993	200	7	118.3
25-May-2016	117.9	20019800	693	123	8	118.4
26-May-2016	119.5	18481300	691	160	-1	118.4
27-May-2016	119.4	13464400	642	169	-2	118
28-May-2016	119.1	NA	246	78	-1	117.9
29-May-2016	119	NA	214	83	-0.5	117.9
30-May-2016	118.9	NA	292	46	-0.5	117.8
31-May-2016	118.8	23547600	613	145	0.1	117.9
1-Jun-2016	118.8	15029500	705	137	1	118
2-Jun-2016	118.9	13228300	529	111	2	117.07

Date	Close	Volume	Positive	Negative	Power Nodes	Prediction
3-Jun-2016	118.5	14135100	641	118	1	117.76
4-Jun-2016	118.6	NA	294	67	1	117.8
5-Jun-2016	118.7	NA	316	49	-0.5	117.9
6-Jun-2016	118.8	12744700	765	162	-5	117.08
7-Jun-2016	117.8	17103000	177	27	3	117.4
8-Jun-2016	118.4	14505600	584	131	1	117.7
9-Jun-2016	118.6	13859200	999	343	-5	116.78
10-Jun-2016	116.6	18510800	220	80	-8	113.8
11-Jun-2016	115.6	NA	435	70	-3	113.8
12-Jun-2016	114.5	NA	300	100	-3	13.4
13-Jun-2016	113.9	31718200	584	131	5	114.3
14-Jun-2016	114.9	17618500	673	147	6	115.55

As we can see in the table, the first column indicates the survey date, while the second is the stock's closure price.

We applied the following method to find the number of days when the market was inactive:

In order to specify the price for a day, we must add the last available price to the first market price and then divide it by two to obtain the missing price. This method has been applied in all previous researches in this field.

Example: If the last available closure price (10th price) for June is $116.6, and the first available one is $113.9 (June 13th), then the first missing day's price (i.e., June 11th) will be:

Price on June 11th = (116.6+113.9)/2=115.25

The next part is the volume of transactions that have been made throughout the day, which refers to buying or selling stocks in the last twenty-four hours. It should be noted that this column is going to remain empty on the days the market is closed, and hence, we use NA marker (NA = Not Available) to show that the amount of transactions is unknown.

The next column is the number of unique tweets with a positive sentiment, which is the output of our program, and that follows with a column listing the number of negative tweets.

The "Powerful Nodes" column is where we put the score of most popular tweets for each day. The last column gives the price obtained from the regression results.

Last but not least, we affect the words power in each day using the Harvard-IV dictionary. Here is a detailed preview of the table that also includes our predictions about the direction of stock price movements.

Table 2
Predictions about direction of stock price movement

Date	Unique Tweets Direction	Unique Anxiety	Powerful Direction	Nodes	Result	Authenticity
2-Apr-2016	N	-	N		-	T
3-Apr-2016	N	-	N		-	T
4-Apr-2016	N	A	N		-	T
5-Apr-2016	N	-	P		NA	NA
6-Apr-2016	N	A	N		-	T
7-Apr-2016	U	A	N		-	T
8-Apr-2016	U	A	N		-	T
9-Apr-2016	U	-	P		+	F
10-Apr-2016	U	-	N		-	T
11-Apr-2016	P	A	P		+	T
12-Apr-2016	P	A	P		+	T
13-Apr-2016	N	A	P		+	T
14-Apr-2016	U	A	N		-	T
15-Apr-2016	N	A	P		+	T
16-Apr-2016	U	-	P		+	T
17-Apr-2016	P	-	P		+	T
18-Apr-2016	P	-	P		+	T
19-Apr-2016	P	A	P		+	T
20-Apr-2016	P	-	P		+	T
21-Apr-2016	N	-	N		-	T
22-Apr-2016	N	A	N		-	T
23-Apr-2016	U	-	P		+	F
24-Apr-2016	U	-	N		-	T
25-Apr-2016	N	-	N		-	T
26-Apr-2016	P	A	P		+	T
27-Apr-2016	P	A	P		+	T
28-Apr-2016	P	A	P		+	T
29-Apr-2016	U	A	P		+	T
30-Apr-2016	P	-	P		+	T
1-May-2016	P	-	P		+	T
2-May-2016	N	-	N		-	T
3-May-2016	P	-	P		+	T
4-May-2016	P	-	N		NA	NA
5-May-2016	P	-	P		+	T
6-May-2016	N	-	N		-	T
7-May-2016	N	-	N		-	T
8-May-2016	N	-	N		-	T
9-May-2016	U	-	P		-	T
10-May-2016	N	-	N		-	T
11-May-2016	P	-	P		+	T
12-May-2016	N	-	N		-	T
13-May-2016	N	-	N		-	T
14-May-2016	N	-	N		-	T
15-May-2016	N	-	N		-	T
16-May-2016	N	-	N		-	T
17-May-2016	P	-	P		+	T
18-May-2016	P	-	N		NA	NA
19-May-2016	P	-	P		+	T

Date	Unique Tweets Direction	Unique Anxiety	Powerful Direction	Nodes	Result	Authenticity
20-May-2016	N	-	P		NA	NA
21-May-2016	P	-	N		NA	NA
22-May-2016	N	-	N		-	T
23-May-2016	N	-	P		NA	NA
24-May-2016	P	A	P		+	T
25-May-2016	P	-	P		+	T
26-May-2016	U	-	N		-	T
27-May-2016	N	-	N		-	T
28-May-2016	N	-	N		-	T
29-May-2016	N	-	N		-	T
30-May-2016	P	-	N		NA	NA
31-May-2016	U	-	P		+	F
1-Jun-2016	P	-	P		+	T
2-Jun-2016	N	-	P		NA	NA
3-Jun-2016	P	-	P		+	T
4-Jun-2016	U	-	P		+	T
5-Jun-2016	P	-	N		NA	NA
6-Jun-2016	P	-	N		NA	NA
7-Jun-2016	P	-	P		+	T
8-Jun-2016	U	-	P		+	T
9-Jun-2016	N	A	N		-	T
10-Jun-2016	N	-	N		-	T
11-Jun-2016	P	-	N		NA	NA
12-Jun-2016	N	-	N		-	T
13-Jun-2016	U	-	P		+	T
14-Jun-2016	P	-	P		+	T

In the above table, the "Unique Tweets Direction" is the prediction of the stock price movement using unique tweets, which is weighted by the Harvard-IV dictionary. This column has three states where considering the value obtained after taking into account powerful word coefficients. We found that neutral conditions are created in many cases and, therefore, we added the unbiased state using the "U" marker. This condition happens when the positive and negative effects are close to each other. The other states are the same as before, "P" for a positive load, and "N" for a negative load. "NA" conditions are still used for the situations when our system cannot predict the case.

After examining the emotional burden of each day, it should be checked whether this market is anxious or under normal circumstances. To find anxiety among market users, we have to check the number of unique tweets. If their number exceeds a thousand tweets a day, that indicates an unusual occurrence and a matter of concern. In other words, in such a scenario, it can be said that people are anxious, and they are tweeting to get some comfort or share their worries with others, or they are seeking the approval of their opinions from

other people. In such a case, the priority is given to prediction results of retweets that are our powerful Nodes mindset (Brand, 1990). Here, users are confused about the situation and look for a credible source to guide and interpret the situation for them. So who can be better than intellectual leaders popular for their good predictions or their thoughtful comments about the stock market? They also have some virtual power, especially with regard to their followers or other users' engagement in their social media activities (likes, comments, and retweets). Social media, in such situations, gives the users a sense of belonging to a group and also most of the time, reduce their anxiety.

The next column is about the daily top tweets where prediction is clear if both directions (unique and retweets) are the same. However, if the directions are not the same, and we see anxiety in tweets, the direction will be the popular thoughts of people, which are the powerful Nodes tweets (popular retweets).

Finally, the last column indicates the results and measures. A conclusion is not possible if both are in normal mode and opposite directions, and hence, the results are labeled as "NA". If the predicted direction is the same as that of the next day's real direction, we call it a success and mark it as a true prediction with "T". However, if it does not match the real direction, the prediction is liable to be false and is marked with "F".

By examining the results, we analyzed the performance of the proposed model. In 81.3% of cases, the directions were predicted correctly. It is noteworthy that the system failed to predict only three out of fifteen cases. In other words, the twelve cases indicated the "NA" situation. That means, there is only a 4.68% error in forecasting the momentum of stock prices.

5 Conclusion

First, let us take a look at the limitations we have encountered. Our primitive limitation is the data itself as fetching them from twitter API is not completely free or easy, and somehow, it is unstructured (JSON files). So it takes too much time and effort to make that data clean, structured, and usable. If we want to use third-party apps to mine data, our most important problem is the app's membership price, which is mostly high. Due to data limitations, it was not logical to perform our analysis in long-term periods. The next problem is the process of fetching, preprocessing, and analyzing data, which are costly and time-consuming, so we just analyzed one single company's stock movements. To get more consistent results, we recommend that it is better to analyze a few more companies' stock movements. Due to the high load of data (1.6 million tweets to train the prediction system), the way of coding on this

project is critical. The codes must be developed in a way that can be run in a short amount of time.

Our last problem was the lack of profiling. We did not have enough information about the users and only judged the power of each tweet by their acceptance from the rest of the users. It is highly recommended that in future researches, an optimal way to be found to save users profiles. Each tweet can get another field of information about its impact depending on whom that posts it (make the researcher add another weighted field about the power of tweet's writer and the number of its likes and shares).

In this study, we used an emotionally tagged dataset using the Python programming language and word-to-vector algorithm. From this perspective, we had enormous sources of data to analyze in an efficient way to minimize the time and other resources they consume. Creating a system like this helps companies at macro and micro levels to be able to predict important things and take advantage of them.

As the study indicated, the price prediction accuracy was 82% using regression. It also predicted the price movement a day before with 81% accuracy. After analyzing the results, it could be concluded that the error rate in the model was below 5%, which was not by chance and hence, be reliable. Another thing that was observed when there was anxiety in the market was often a dramatic increase in the number of unique tweets. It can be argued that this increase in activities on twitter is the result of users' anxiety and the likelihood that shareholders will buy or sell their stock. As the results indicate, in a 3-day of anxiety leads to a noticeable increase in stock trades. To prove this part, there is a need to analyze stock movements and activities of a lot more companies on social media.

The most important finding of the current research is that the power of each tweet is not equal to others, which means one user's tweet might be more effective depending on its views, likes, comments, and the number of followers. The effect of these tweets can be too much as they might lead others to change their minds in an uncertain circumstance. In other words, these people with high impact can be intellectual leaders of a virtual society. So, when things get unclear and doubtful, people do not know what to do, and they need guidance. In these situations, they will rely on the mentioned powerful tweets and try to go with the wave. That is how we can predict the movement direction in uncertain situations.

If possible, we will be more accurate if you can rank them by reviewing people's profiles and the number of tweets viewed. It makes the power of each

node different depending on its popularity. The long-term follow-up should also be considered if possible.

References

Brand, A. (1990). *The Force of Reason: An Introduction to Habermas' Theory of Communicative Action*. 1(1990), 1–152.

Bholane Savita, D., & Gore, D. (2016). Sentiment Analysis on Twitter Data Using Support Vector Machine. *International Journal of Computer Science Trends and Technology (IJCST)*, 4(3), 831–837

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The Adjustment of Stock Prices to New Information. *International Economic Review*, 10(1), 1–21.

Juan, A., & Vidal, E. (2002). On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12), 2705–2710.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The Effect of News and Public Mood on Stock Movements. *Information Sciences*, 278, 826–840.

Marwala, T., & Hurwitz, E. (2017). Efficient Market Hypothesis. In T. Marwala & E. Hurwitz (Eds.), *Artificial Intelligence and Economic Theory: Skynet in the Market* (pp. 101–110). Springer International Publishing AG.

Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating High-Coverage Semantic Orientation Lexicons from Overtly Marked Words and a Thesaurus. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, August, 599–608.

Oliveira, N., Cortez, P., & Areal, N. (2017). The Impact of Microblogging Data for Stock Market Prediction: Using Twitter to Predict Returns, Volatility, Trading Volume, and Survey Sentiment Indices. *Expert Systems with Applications*, 73, 125–144.

Patil, H. P., & Atique, M. (2015). Sentiment Analysis for Social Media: A Survey. *2015 2nd International Conference on Information Science and Security (ICISS)*, 1–4.

Schumaker, R. P., Zhang, Y., Huang, C. N., & Chen, H. (2012). Evaluating Sentiment in Financial News Articles. *Decision Support Systems*, 53(3), 458–464.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3), 1139–1168.

Timmermann, A., & Granger, C. W. J. (2004). Efficient Market Hypothesis and Forecasting. *International Journal of Forecasting*, 20(1), 15–27.

Wallach, H. M. (2006). Topic Modeling: Beyond Bag-of-Words. *Proceedings of the 23rd International Conference on Machine Learning*, 977–984.

You, W., Guo, Y., Zhu, H., & Tang, Y. (2017). Oil Price Shocks, Economic Policy Uncertainty, and Industry Stock Returns in China: Asymmetric Effects with Quantile Regression. *Energy Economics*, 68, 1–18.

Zainuddin, N., & Selamat, A. (2014). Sentiment Analysis Using Support Vector Machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 333–337.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). Understanding Bag-Of-Words Model: A Statistical Framework. *International Journal of Machine Learning and Cybernetics*, 1(1), 43–52.