International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 11, Issue 11, November 2020, pp. 1123-1130, Article ID: IJARET_11_11_102
Available online at http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=11
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
DOI: 10.34218/IJARET.11.11.2020.102

USE SENTIMENT ANALYSIS TO PREDICT FUTURE PRICE MOVEMENT IN THE STOCK MARKET

Dr. K.S.M.V. Kumar
Professor, Department of Computer Science and Engineering,
Aditya College of Engineering, Madanapalle, Andhra Pradesh, India,

Dr. G Rajendra Kumar
Professor, Department of Computer Science and Engineering,
Sivani College of Engineering, Chilakapalem Jn., Srikakulam, Andhra Pradesh, India

J. Nageswara Rao
Sr. Assistant. Professor, Department of Computer Science and Engineering,
Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, Krishna Dt,
Andhra Pradesh, India

ABSTRACT

In recent years, the generation of user data has greatly increased. Due to the development of Internet-based applications, social media users have grown exponentially. Millions of users share their opinions every day. Therefore, social media has become a powerful communication tool and a rich source of opinion data. This data is usually used to influence a large number of people. Over the years, this process has gained momentum and led to the development of a new era of research, namely sentiment analysis. Sentiment analysis (also known as opinion mining) uses new technologies and algorithms to collect and analyze opinions about various products and services. The main goal of this article is to use sentiment analysis and machine learning to predict stock prices. This feature can help investors predict the stock market. It may not be possible to accurately predict the stock market based on historical price analysis alone. In order to improve the prediction, we perform sentiment analysis based on the opinions of different Twitter users and add sentiment scores and prices. Although forecasting the stock market is a tedious task, there are many forecasting methods. By using the tensor flow platform and Twitter API to get tweets, we have implemented a new sentiment analysis method. Our results show that this scheme has better accuracy than existing methods in predicting the stock market.

Keywords: Machine Learning, Prediction, Sentiment Analysis, Twitter API.

Cite this Article: K.S.M.V. Kumar, G Rajendra Kumar and J. Nageswara Rao, Use Sentiment Analysis to Predict Future Price Movement in the Stock Market, *International Journal of Advanced Research in Engineering and Technology*, 11(11), 2020, pp. 1123-1130.
http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=11

1. INTRODUCTION

Sentiment Analysis employs various techniques for collecting and analyzing opinions about products and services. In general, for decision making we consider opinions of different persons. This fundamental fact is also applicable for organizations. In olden days, there was no well-defined techniques for opinion mining. In those days when any person intends to make decision used to collect opinions manually form friends and relatives. Similarly when any organization wanted to make decision about production used to conduct surveys manually. With the assistance of surveys, organizations used to get opinions of various customers through which task of decision making was accomplished.

However, with the massive growth of social media contents and E-commerce in the past few years, the way of decision making has been totally changed. Nowadays, many applications and websites provide a platform for people to post their reviews about products online. Social media websites, blogs and discussion forums are the primary sources for sharing views and comments.

Furthermore, collecting related opinions from different websites and applications is a hectic task due to existence of large number of diverse websites, and each website may have a huge volume of opinionated text. Also, opinions are hidden in enormous forum posts and blogs. It is impractical to find relevant websites, extract related opinions, summarize and organize them into usable forms manually. It demands an automated opinion discovery and summarization system which is used for prediction. We used sentiment analysis in stock market prediction, as predicting the prices of the stock is very crucial. It is useful in deciding whether to buy the shares or sell them on the fly with profits. As the social media acts as a platform to view the reviews of the customers and apply strategic decisions whenever needed. We used Twitter as a source of the data by using the Twitter API we retrieved the tweets and trained different models of algorithms and compared the results on the tensorflow platform.

For sentiment analysis various approaches can be used which are listed in figure 1 (source: sciencedirect.com).

The following are the some applications of sentiment analysis in real world.

- Social Media Monitoring
- Customer Service
- Market Research
- Intelligent Transportation System
- Online advertising

Section 2 explains the research work done by various researchers in the direction of sentiment analysis.

2. RELATED WORK

Dev Shah et.al.[1] proposed a model in which they used moneycontrol.com to get the stock-specific news as the source and after they used a dataset containing the specific news articles

from the past six months relating to nifty and Sensex [1]. They developed a web scraper to handle the format of the article links using the moneycontrol.com API they got the articles and made a corpus of data. They used python to turn the corpus into numerical vectors for each article i.e. into n-grams model. Then they used stemming to reduce the words and scores are calculated based on the dictionary values. The decision of selling or buying is made relying on the score of the news article.

Bharadwaj et.al [2] proposed a method in which they fetched the live server data by using the python programming to perform sentiment analysis. By using the Beautiful soap they pulled the required data from the webpage and then saved the required information. Then the data is preprocessed for the feature selection and afterwards the sentiment analysis for the stock market is performed based on the variation of the predicted values.

GeolMittal et.al [3] proposed an algorithm in which the raw DJIA values are pre-processed and the tweets are analyzed by sentiment and classified into four classes depending on the mood. Then these moods and the DJIA values are given to model for training. In this way, future values are predicted by the system to make the decision of buying or selling.

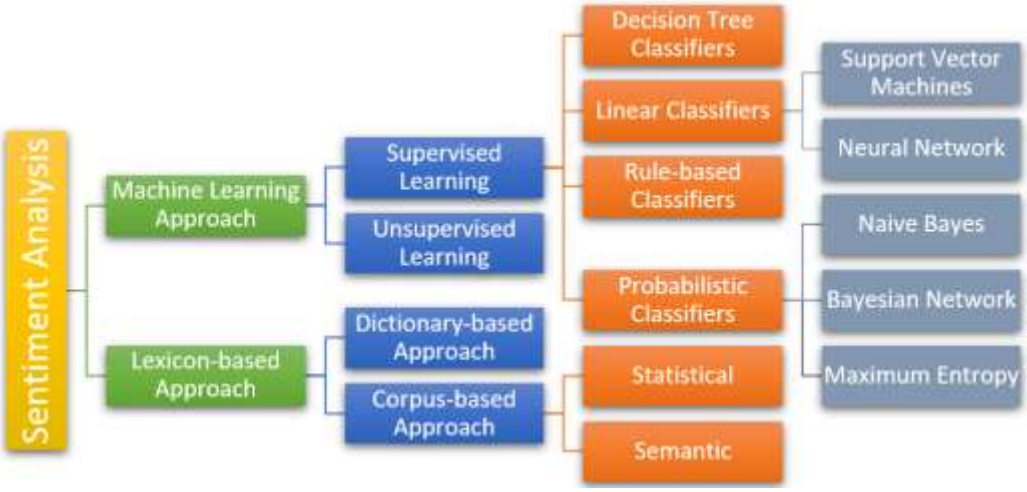

Figure 1 Approaches for Sentiment Analysis

Jain et.al [4] proposed a mechanism to get the sentiments of Indian people to the political issue. This method follows as collecting the raw tweets and then pre-processed and then the data is cleaned then some methods are applied to convert the data into training and testing data. Now the data is fed into the classifiers. This model used sentiwordnet dataset. In this model dataset is classified as positive, negative and neutral. Also accuracy values of different algorithms mainly KNN, random forest, and naive bayes are compared.

Jain et.al [5] proposed another good method in which they used the twitter API to get the data by using the API secret keys and access keys they obtained the tweets by signing as the developer. Then the data is pre-processed i.e. case conversion, stop words removal, punctuation removal, stemming, lemmatization and spelling correction are used. Once this feature extraction is done based on the frequency, finally, training and testing of different algorithms is done using the multinomial naive Bayes and decision tree.

Rajyalakshmi et.al [6] demonstrated the concept of sentiment analysis. This concept is explained step wise i.e. data collection, text preparation, sentiment detection, sentiment classification, and various approaches i.e. using the lexical based, machine learning and hybrid algorithms. They explained the merits and demerits of used algorithms. [6] Explained various dictionaries available for accomplishing sentiment analysis like SentiwordNet and LIWC etc.

Zhang et.al [7] performed the microblog sentiment analysis by considering the opinions of endowment insurance microblog as the study object for emotional analysis. They used the cloud collection to open the website then loop the page to crawl the data and then open another page to do the same to get the data until the end of blog. Then use an Emotional Dictionary to do the sentiment analysis. They processed the text and performed the emotional clustering analysis i.e. classifying the data based on the emotions of the text to get the polarity of the text. Based on this polarity the emotional analysis is done.

Ali Hasan et al. proposed a new model in which they gathered data from Twitter and then they translated Urdu to Hindi by using translator API. Afterwards the data storage is used for storing and preprocessing of the tweets i.e eliminations, removing URLs, special symbols and symbols elimination. Then the polarity of the words is calculated by using the text blob, sentiment, and W-WSD. After this step, validation is performed by using weka. This system employed the twitter dataset for the positive negative and neutral reviews. Based on twitter dataset, the training dataset is created and scored the tweets according to the machine. Then the classification model is used. The dataset is tested and classified the tweet as a positive or negative and finally positive or negative percent.

Imane EL Alaboui et.al [9] proposed a new and adaptable approach of sentiment analysis for analyzing big data. This model consists of building the annotated sentiment words basing on the context and selected set of the positive and negative hashtags. Classification is performed based on the annotated sentiment words and by balancing the new metrics and also by considering various classes like highly positive, moderately positive, lightly positive etc. And then the prediction is performed by combining the NLP and new metrics together.

3. PROPOSED METHOD

In proposed method first step is to collect the tweets and then set the prices according to the dates of the tweets. After this step, calculate the scores for the tweets and then decide the stock prices according to the prices and the sentiment scores by training the model.

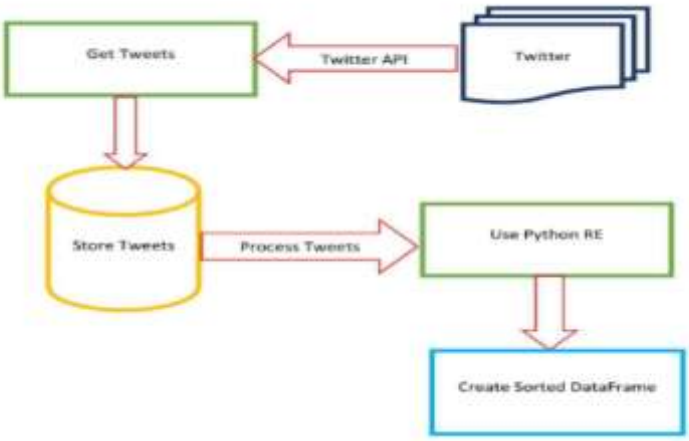

Figure 2 Flow Diagram for Retrieving and Storing of Tweets

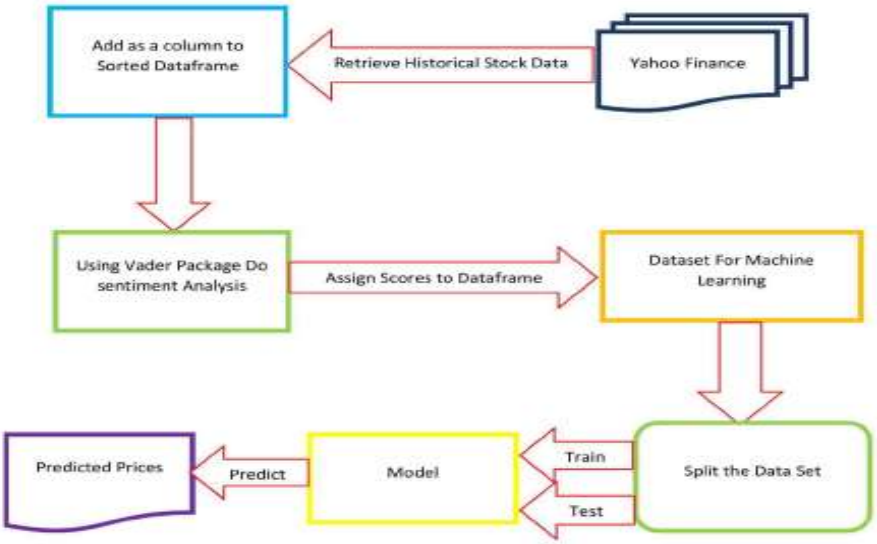

Figure 3 Processing of Tweets for Prediction

ALGORITHM

Step 1: Getting Tweets from Twitter

Step 1.1: Create a Twitter Account and Sign Up as App Developer
Step 1.2: Get the Secret Keys to connect to Twitter API
Step 1.3: Using Tweepy in Python Connect to API
Step 1.4: Create CSV of Tweets by retrieving them.

Step 2: Pre-processing of Tweets

Step 2.1: Remove Special Characters
Step 2.2: Creating a Date wise Tweets data frame and merge them

Step 3: Creating Data Frames for Use

Step 3.1: Getting the Historical Stock Data from Yahoo Finance CSV
Step 3.2: Assigning Prices for the respective Days in created Data Frame
Step 3.3: Substituting mean values where the price is not available for that day.
Step 3.4: Now create more columns in the data frame for negative and positive tweets.

Step 4: Perform Sentiment Analysis

Step 4.1: Import Vader package
Step 4.2: Calculate sentiment scores and assign to data frame

Step 5: Train Various Models and Predict Prices

Step 5.1: Using the Same dataset Divide into Training and Testing
Step 5.2: Train Models
Step 5.3: Get the Predicted Stock Prices

Getting the Tweets from Twitter

For this, we created a twitter account and signed up as the app developer. After signing in, twitter will provide us the consumer key and consumer secret keys with which we can connect

to the Twitter API. In python, there is a library called Tweepy with which we can use the keys to connect and retrieve the tweets by using the OAuth handler and by setting access with the keys provided. Then the name of the company from which we need to retrieve the tweets must be specified. Retrieved tweets are stored by creating a data frame using the pandas. The data frame consists of both tweets and date of the tweet. Finally, create a CSV file just to store the tweets before the pre-processing.

Pre-processing of Tweets

Pre-processing includes using the regular expressions library to just remove unwanted characters from the tweets and removing the punctuation marks and numbers and any unwanted spaces so its easy to stem words.so now pre-processed data is again stored into a data frame and ready for next step where the tweets are sorted according to the dates and made to come in a single line so we can compare the dates.

Creating Dataframes for use

Getting the historical data of the same stock from the Yahoo Finance is the next step. Yahoo. Finance website provides all the required stock prices for at least past year till date so we can get the closing price of the stock. So next step is just to read the CSV file into the Tensorflow platform by using csv lib in python and then storing in a data frame. Now according to the dates, we will just create a new data frame in which date tweets and prices will be stored based on a simple logic of string comparison in python we can achieved this.

Perform Sentiment Analysis

We imported VADER package from the nltk in python to Do the sentiment analysis which we give the scores to the tweets depending on the positive negative and neutral values of the tweets for a single day and then we created another four columns to the same data frame and stored the value and now the data frame is ready for the machine learning model training.

Machine Learning

We first took the analysis of the percent of the positive and negative tweets in the data we got and then we split the data frame by indexing into training and testing datasets. We also split the sentiment scores into training and testing. Now we were ready with datasets so we imported the sklearn package and used the random forest regressor to fit both the prices and tweets dataset and the sentiment scores dataset. And we predicted the prices with the test dataset. We also used the linear regression in the same way and also used MLPC classifier Out of which the Random Forest Performed Best with large dataset and with small dataset the linear regression worked well.

4. RESULTS

This proposed system is implemented by using python programming and tensorflow framework. We considered the date as additional metric for sentiment analysis. We collected tweets from twitter API and prices from yahoo finance for predicting prices. The sample dataset is shown in table 1. We trained the models and analyzed the accuracy in predicting prices. Table 2 and Table 3 illustrates the price prediction using Random forest and Linear Regression.

Table 1 Sample Dataset

	Date	Prices	Comp	Negative	Neutral	Positive
0	2019-10-25	1408	0.9977	0.038	0.8	0.162
1	2019-10-24	1408	0.9988	0.038	0.834	0.128
2	2019-10-23	1408	0.9989	0.038	0.882	0.08
3	2019-10-22	1414	-0.9999	0.095	0.88	0.025
4	2019-10-21	1408	0.999	0.029	0.884	0.087
5	2019-10-20	1408	0.9987	0.038	0.86	0.102
6	2019-10-19	1408	0.9997	0.03	0.856	0.114
7	2019-10-18	1416	1	0.026	0.858	0.116
8	2019-10-17	1396	0.9987	0.022	0.895	0.083

Table 2 Random Forest

Real Prices	Predicted Prices
1408	1408.1
1406	1407.2
1396	1404.35

Table 3. Linear Regression

Real Prices	Predicted Prices
1408	1407.43
1406	1407.05
1396	1406.88

Accuracy of price prediction using random forest, linear regression and neural networks is shown in Figure 4, Figure 5 and Figure 6.

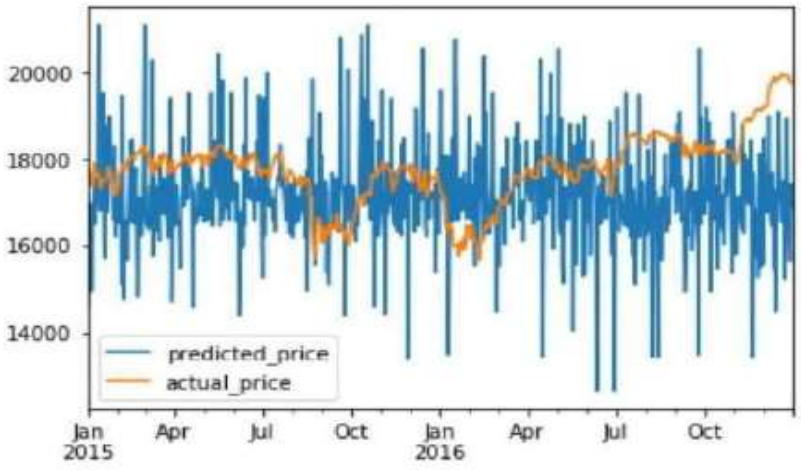

Figure 4 Random Forest with Large Dataset

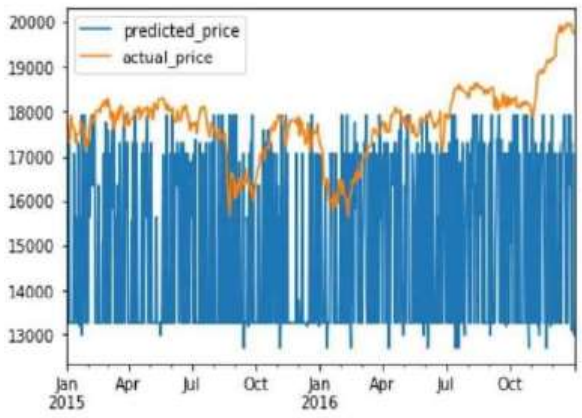

Figure 5 Linear Regression with Large Dataset **Figure 6** Neural Networks with Large Dataset

5. CONCLUSION

This work demonstrates the prediction of stock prices by using twitter API. To accomplish this task, we collected the tweets and created a data frame. By comparing the dates we gave the sentiment scores for the tweets. The prices are taken from yahoo finance and attached to the tweets. Then we trained the considered models and predicted the stock prices. Our experimental analysis illustrates that the random forest is best performing algorithm with larger Twitter dataset. Also our results shows that for smaller dataset the linear model performs well.

In the future we will improve the models by providing more data sources not restricted to twitter but also from other available sources so that the sentiment scores can help our system to improve and predict the prices more correctly.

REFERENCES

[1] Dev Shah, Haruna Isah and Farhana Zulkernine,"Predicting the Effects of News Sentiments on the Stock Market", cite as arXiv: 1812.04199,2018.

[2] Aditya Bhardwaja, Yogendra Narayanb, Vanrajc, Pawana and Maitreyee Duttaa, "Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty", 4th International Conference on Eco-friendly Computing and Communication Systems, 2015.

[3] Mittal and A. Goel, "Stock prediction using twitter sentiment analysis," Standford University, CS229 (2011http://cs229.stanford.edu/proj2011/GoelMittalStockMarketPredictionUsingTwitterSentimentAnalysis. pdf), vol. 15, 2012.

[4] P. Jain, "Application of Machine Learning Techniques Sentiment Analysis," 2016 2nd International Conference on Applied and Theoretical Computing and communication technology (iCATccT), IEEE, Bangalore, India, 2016.

[5] A. P. Jain, "Sentiments Analysis Of Twitter Data Using Data," in 2015 International Conference on Information Processing (ICIP),IEEE, pune, 2015.

[6] S.Rajalakshmi, "A Comprehensive Survey on Sentiment Analysis," Fourth International Conference on Signal Processing, Communication and Networking, 2017.

[7] Zhang, X., Fuehres, H., & Gloor, P. A, Predicting stock market indicators through twitter, ELSEVIER Procedia-Social and Behavioral Sciences, 2011, pp. 55-62.

[8] Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband,"Machine Learning Based Sentiment Analysis for Twitter Accounts", Mathematical and computer applications, 2018.

[9] Imane El Alaoui1,Youssef Gahi,Rochdi Messoussi1,Youness Chaabi1, Alexis Todoskoff and Abdessamad Kobi, "A novel adaptable approach for sentiment analysis on big social data", Big Data (Springer Open),2018.

[10] Sophia,Naman Awasthib and Vishal Sharmac,"Introduction to Machine Learning and Its Basic Application in Python" 10th International Conference on Digital Strategies for Organizational Success,2019.