2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) | 979-8-3503-4023-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACRS58579.2023.10404138

An Analysis on Sentiment Analysis and Stock Market Price Prediction

1st Manjusha Pandey
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
manjushafcs@kiit.ac.in

2nd Sinkon Nayak
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
sinkonnayak07@gmail.com

3rd Siddharth Swarup Rautaray
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
siddharthfcs@kiit.ac.in

Abstract—Sentiment analysis, otherwise called as opinion mining, is used for analyze the data to find out the opinion or attitude or feeling from them. These days this technique is used for anticipation of stock market. The idea is that by analyzing social media data, one can gauge the people sentiment around a particular stock or industry and use that information to make more informed investment decisions. This paper presents an analysis of sentiment for stock market prediction using Twitter data. The paper presents the potential for using sentiment analysis as a tool for predicting stock market performance. The outcome indicate that sentiment analysis can provide valuable insights into public sentiment around specific stocks or industries. However, the performance for stock anticipation depend on the quality of the data collected, the correctness of the sentiment analysis algorithms(SAA) used, and the ability to correlate the sentiment data with stock market performance.

Keywords—Sentiment Analysis, Predictive Analysis, Stock market anticipation.

I. INTRODUCTION

The stock market is a complex system that involves a range of factors that influence its performance, like financial data, economic indicators, and world events all contribute to the growth and tumble of stock prices[14]. However, it is often overlooked in public opinion. Due to the rise in the usage of social platforms, people share their thoughts and opinions and connect more as compared to the previous era [15]. People share their thoughts and opinions can spread rapidly across the internet, including their thoughts about specific stocks or industries [8].

In twitter every user is allowed to share their thoughts and opinions regarding various topics [12]. Twitter sentiment analysis is used to analyze the sentiment of tweets about specific stocks or industries[16]. By analyzing the sentiment of tweets in real-time, we can gain insights into public perception and use it to predict stock market movements[7]. And this area is gaining more interest [9]. Researchers and analysts have developed various models and algorithms to analyze the opinion of tweets and use it to predict stock prices which have attracted significant attention from researchers and analysts[13].

In this research work, we have tried to investigate the use of Twitter sentiment analysis in stock analysis and to develop a robust and accurate method for anticipating stock market performance using real-time sentiment analysis data. The research paper will present a detailed analysis of the tools and techniques used for Twitter sentiment analysis.

The ultimate goal is to provide investors and traders with a reliable tool that can help them make informed decisions about buying or selling stocks based on social media sentiment analysis.

II. LITERATURE SURVEY

Sentiment of Twitter data analysis and stock market analysis is a relatively new field, and there has been remarkable investigation in recent period of time to explore its potential applications and limitations. One important concept in this field is natural language processing (NLP), which is used to analyze and identify the tone and emotion behind the messages. Machine learning is also used to train algorithms to accurately classify tweets as optimistic, pessimistic, or indifferent.

V. S. Pagolu et al. analyze the twitter data of Microsoft by analyzing their sentiment and Supervised Machine Learning Algorithms (SMLA). They concluded that stock price and twitter sentiment are correlated [1]. wathi, T., et al. suggested a novel learning based algorithm with LSTM based sentiment analysis for stock anticipation using Twitter data [2]. Y. E. Cakra and B. Distiawan Trisedya, predict the Indonesian stock market by analyzing the sentiment and SMLAs and landed with a judgement that they are powerfully related to each other [3].

D. R. Pant, et al. used Recurrent Neural Network for the prediction of price of Bitcoin by analyzing Twitter sentiment [4]. X. Guo and J. Li, proposed a novel SMLA based on baseline correlation for the analysis of Twitter sentiment data to predict future stock price [5]. To anticipate the stock price R. Gupta and M. Chen, used StockTwits data which is a micro blogging platform to share sentiment related to stocks and financial marketing to predict using SMLA [6].

III. MACHINE LEARNING TECHNIQUE

Machine learning methods(MLM) are employed to examine the data and detect patterns or abnormalities in it. These algorithms can predict future trends, and provide suggestions for support. In this study, we have considered two classification model and three regression model, which are widely used for the anticipation of stock. Random forest(RF) and Gradient Boosting(GB) classifier provide better accuracy, can deal with overfitting issue also give better prediction result. Random forest regressor is used due to its robust nature, scalability and can capture nonlinear relationships. LSTM provide better predictive result in time series analysis and XGBoost provide higher accuracy, speed and scalability.

A. Random Forest Classifier

In MLM, one popular algorithm is the RF classifier. It's commonly used in applications such as image recognition, speech recognition, and recommendation systems. It blends numerous decision trees to make predictions, which increases its accuracy and reduces overfitting. This method combines the concepts of decision trees and ensemble methods. To build a random forest, we first create a group of decision trees using a technique called bootstrap aggregating (or bagging for short). For each tree, we randomly sample a subset of the training data with replacement (i.e., some instances may appear many times while others may not appear at all). The "random" in random forest comes from two sources of randomness: feature sampling and bootstrapping. In feature sampling, we randomly choose a subgroup of the features at each split point. This helps prevent overfitting and reduces the correlation between the trees. In bootstrapping, we randomly sample instances from the training set with replacement.

B. Gradient Boost Classifier

Gradient Boosting is a popular MLM that is utilized to create highly accurate predictions. It belongs to ensemble method because it blends or merges the weak classifiers to make a strong classifier model. It is accomplished by correcting the mistakes produced by the previous weak learners. After each iteration, the weights of the misclassified instances are increased to supply more value to those instances during the next iteration. The end output of the ensemble model is the weighted average of the anticipations made by all the weak learners[11]. This is called boosting process. A detailed explanation of how gradient boosting works is provided.

Ensemble learning is a technique in which various methods are aggregated to enhance the overall accuracy of the prediction. The primary objective of ensemble learning is that the combination of weaker models creates a strong model that is more accurate than the individual ones. Ensemble learning helps to lessen the variance, which reduces overfitting and improves the generalization performance.

C. LSTM Regressor

Long Short-Term Memory (LSTM) comes under Recurrent Neural Network (RNN) that was configured to address the vanishing gradient problem. It consists of multiple memory cells that can memorize data for a longer time period and make predictions based on long-term dependencies. To create this model, we added LSTM layers. The return_sequences parameter needs to be set to True every time we add a new LSTM layer, excluding the final layer. The input_shape is the number of time steps and the number of indicators.

After each LSTM layer, a Dropout layer is added to prevent overfitting. Dropout is a regularization method to deal with overfitting. The parameter passed to the Dropout layer is the fraction of nodes that will be dropped on each epoch, for this demo, we will use a dropout value of 0.2, which means that on each epoch we will randomly drop 20% of the units.

D. XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is a MLM known for its greater accuracy, speed, and scalability. It has an ability to model complex non-linear relationships between features and target variables. XGBoost uses an ensemble of decision trees to create a strong anticipation model. The architecture of XGBoost consists of an ensemble of decision trees.

The advantage of XGBoost over other boosting algorithms is its use of gradient-based optimization to minimize the loss function while building each tree. This approach enables XGBoost to converge faster and avoid getting stuck in local minima. Additionally, XGBoost includes L1 and L2 regularization techniques that help prevent overfitting of the model.

E. Random Forest Regressor

Random Forest Regression is an expansion of the Random Forest Classifier method. In regression tasks, the desire is to anticipate a continuous numerical value rather than a categorical label. The Random Forest Regression algorithm builds a collection of decision trees, where each tree is trained on a random subset of the training data.

IV. PROPOSED METHODOLOGY

Collection and selection data is a crucial step in conducting research. In this case, for analyzing the sentiment of twitter data for anticipating movement in stock price of Apple Inc. Data is used. Figure 1 demonstrate the workflow of the methodology followed.

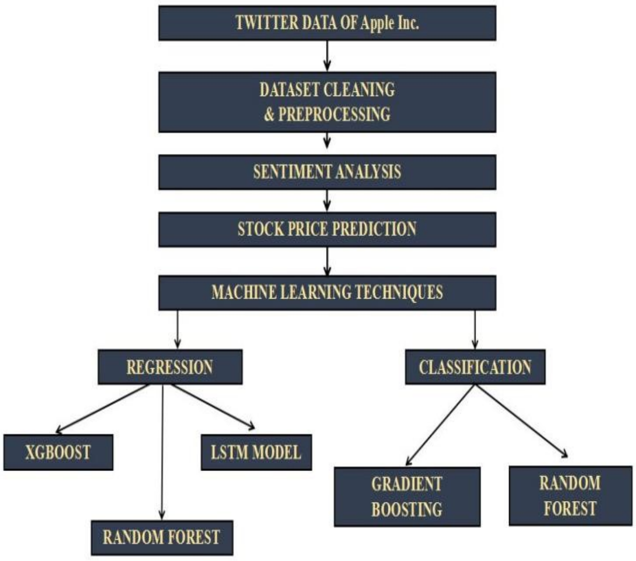

Fig. 1. Methodology

The introductory step is collection of data from Twitter of Apple Inc. company. We have taken the stock data of Apple Inc. Which is consist of tweets that mention the company's name or stock symbol like open, high, low, close, volume, Adj close, ts_polarity etc. The next step will be to preprocess the collected data. This will involve removing irrelevant information, such as URLs and hashtags, and cleaning up the text by removing stop words and punctuation marks. Once the data has been preprocessed, the sentiment of each tweet will be analyzed using MLM. The sentiment analysis(SA) will determine whether the tweet is positive, negative, or neutral. Finally, the sentiment scores will be used to anticipate the stock price. This will involve training a MLM that can anticipate stock prices based on the scores of sentiment. The success of this proposed approach is measured by the accuracy of the stock price predictions. The model's accuracy will be evaluated using various metrics.

V. RESULTS AND ANALYSIS

The research paper will analyze the outcome of the sentiment analysis and their correlation with stock market performance. To determine the interrelation between SA and stock price the statistical methods are used and analyzed.

TABLE I. RANDOM FOREST CLASSIFIER

	Precision	*Recall*	*F1-score*
0	0.48	0.29	0.36
1	0.55	0.73	0.63
Accuracy			0.53
Micro avg	0.52	0.51	0.50
Weighted avg	0.52	0.53	0.51

Table 1 is the depiction of performance of RF Classifier which denotes the accuracy of the model is 53%.

TABLE II. GRADIENT BOOST CLASSIFIER

	Precision	*Recall*	*F1-score*
0	0.53	0.33	0.41
1	0.57	0.75	0.65
Accuracy			0.56
Micro avg	0.55	0.54	0.53
Weighted avg	0.55	0.56	0.54

Table 2 is the depiction of performance of GB Classifier which denotes the accuracy of the model is 56%.

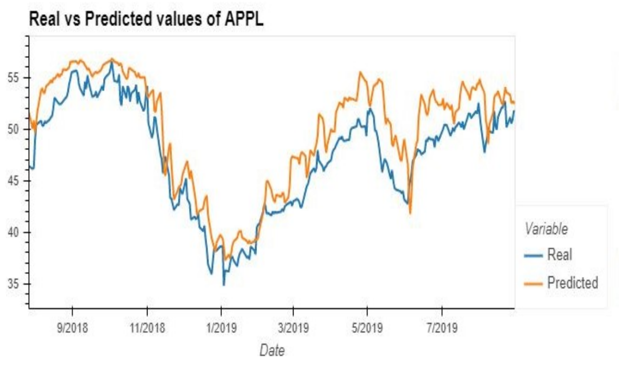

Fig. 2. LSTM

Figure 2 is the depiction of performance of LSTM which denotes the real vs predicted value of Apple Inc. data for the adj close column. The R-Squared value is 68.01% while RMSE value is 0.1334.

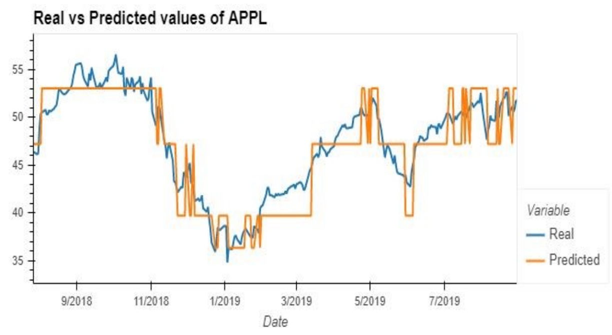

Fig. 3. Random Forest Regressor

Figure 3 is the depiction of performance of Random Forest Regressor, which denotes the real vs predicted value of Apple Inc. data for the adj close column. The R-Squared value is 82.27% while RMSE value is 0.0993.

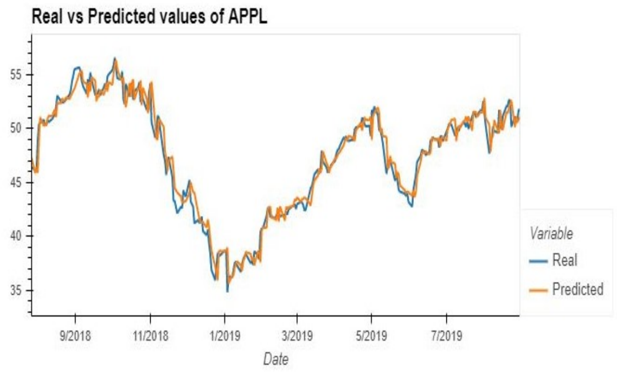

Fig. 4. XG Boost Regressor

Figure 6 is the depiction of performance of XG Boost Regressor, which denotes the real vs predicted value of Apple Inc. data for the adj close column. The R-Squared value is 95.92% while RMSE value is 0.0477.

VI. CONCLUSION AND FUTURE SCOPE

From the study of twitter data, this study concludes that the stock price and sentiment of twitter both are consonant with each other. Further, the regression methods perform better when compared to other classification methods. This study discover that the Twitter sentiment analysis can provide a valuable source of content for both investors and traders. By incorporating Twitter sentiment analysis into their investment strategies, investors can gain a finer market understanding, identify trends and patterns, and make more informed decisions about their investments. However, Twitter SA does not come without limitations. SA results may not always errorlessly indicate market sentiment, and there may be other factors that influence stock prices that are not reflected in Twitter data. Therefore, investors should use Twitter SA as one of several tools in their investment strategies.

REFERENCES

[1] V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 2016, pp. 1345-1350, doi: 10.1109/SCOPES.2016.7955659.

[2] wathi, T., Kasiviswanath, N. & Rao, A.A. An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. Appl Intell 52, 13675–13688 (2022).

[3] Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2015, pp. 147-154, doi: 10.1109/ICACSIS.2015.7415179.

[4] D. R. Pant, P. Neupane, A. Poudel, A. K. Pokhrel and B. K. Lama, "Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis," 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, Nepal, 2018, pp. 128-132, doi: 10.1109/CCCS.2018.8586824.

[5] X. Guo and J. Li, "A Novel Twitter Sentiment Analysis Model with Baseline Correlation for Financial Market Prediction with Improved Efficiency," 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 472-477, doi: 10.1109/SNAMS.2019.8931720.

[6] R. Gupta and M. Chen, "Sentiment Analysis for Stock Price Prediction," 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 2020, pp. 213-218, doi: 10.1109/MIPR49039.2020.00051.

[7] Sharma, V., Khemnar, R., Kumari, R., & Mohan, B. R. (2019, September). Time series with sentiment analysis for stock price prediction. In 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT) (pp. 178-181). IEEE.

[8] Chakraborty, P., Pria, U. S., Rony, M. R. A. H., & Majumdar, M. A. (2017, September). Predicting stock movement using sentiment analysis of Twitter feed. In 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT) (pp. 1-6). IEEE.

[9] Nayak, S., Pandey, M., & Rautaray, S. S. (2022). Reviews Based Sentiment Analysis for Optimizing Product Rating System. Available at SSRN 4121847.

[10] Pandey, M., & Rautaray, S. S. (Eds.). (2021). Machine Learning: Theoretical Foundations and Practical Applications (Vol. 87). Springer Nature.

[11] Gourisaria, M. K., Agrawal, R., Harshvardhan, G. M., Pandey, M., & Rautaray, S. S. (2021). Application of machine learning in industry 4.0. Machine learning: Theoretical foundations and practical applications, 57-87.

[12] Karimi, N., Dash, A., Rautaray, S. S., & Pandey, M. (2021). Customer profiling and retention using recommendation system and factor identification to predict customer churn in telecom industry. Machine Learning: Theoretical Foundations and Practical Applications, 155-172.

[13] Harshvardhan, G. M., Gourisaria, M. K., Sahu, A., Rautaray, S. S., & Pandey, M. (2021, March). Topic modelling Twitterati sentiments using Latent Dirichlet allocation during demonetization. In 2021 8th international conference on computing for sustainable global development (INDIACom) (pp. 811-815). IEEE.

[14] Kanaujia, P. K. M., Pandey, M., & Rautaray, S. S. (2017, February). Real time financial analysis using big data technologies. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 131-136). IEEE.

[15] Yadav, K., Rautaray, S. S., & Pandey, M. (2017). A Prototype for Sentiment Analysis Using Big Data Tools. In Computational Intelligence, Communications, and Business Analytics: First International Conference, CICBA 2017, Kolkata, India, March 24–25, 2017, Revised Selected Papers, Part I (pp. 103-117). Springer Singapore.

[16] Bhadra, K., Dash, A., Darshana, S., Pandey, M., Rautaray, S. S., & Barik, R. K. (2023, May). Twitter Sentiment Analysis of COVID-19 In India: VADER Perspective. In 2023 International Conference on Communication, Circuits, and Systems (IC3S) (pp. 1-6). IEEE.