# Stock prediction and Analysis using Twitter Sentiments: A Survey

[1]Hardik Lad, [2]Jairaj Saraf, [3]Siddhi Pawar, [4]Dr. Vidya Chitre

[1,2,3]UG Student, [4]Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India,

[1]hardik.lad773@gmail.com, [2]jairaj.saraf01@gmail.com, [3]siddharth0914@gmail.com,

[4]vidya.chitre@vit.edu.in

**Abstract - Stock market analysis and its prediction are one of the most popular research topics. To predict future stock performance based solely on historical stock market data has proven to be inadequate. This historical stock data when combined with public sentiments on the same stock or asset has proven to yield more accurate predictions. Tweets can be used to provide public sentiments on a particular asset when we consider them in aggregate. In this paper, we survey and analyze related works on this same concept and the aim is to develop a system that combines historical stock data and processed and labeled public sentiment scores and examine various effective machine learning techniques used for the time series model for providing accurate future stock predictions.**

*Keywords —sentiment analysis, stock market, accuracy, machine learning, regression, prediction, sentiments*

## I.     INTRODUCTION

In this modern era, the use of social media has reached unprecedented levels. A well-performed online deduction states that public sentiments and stock performance are correlated. It has been observed that sudden changes in stock markets movements have an effect on peoples' emotions and opinions on that particular stock and vice versa. When people are more happy, optimistic and in a good mood, they are more likely to make investments which in turn uplifts the stock market performance. The problem with this theory is, we cannot analyze each person's sentiments manually and quantify them. For this, we require a platform where we can observe and analyze publicly shared opinions about an asset or a stock, a platform like Twitter. Twitter is one of the biggest platforms to share your political, social, and economic views publicly. These publicly shared tweets can be easily mined using a number of open-sourced APIs and we can perform sentiment analysis on them. By mining these tweets and narrowing them down to tweets associated with a certain company and then by performing Sentiment analysis, we can see a pattern, correlation, and people's emotion associated with that certain company or asset.

In this present time, Stock market prediction is a well-versed topic. Researchers all around the world have put efforts into stock prediction, but it still remains a mammoth task. Conventional Stock prediction techniques involve fetching DJIA and stocks closing prices data and applying Machine Learning algorithms on it. Although this may have proved useful for a certain period, this method has certain drawbacks such as moderate accuracy rate and unsuccessful in predicting future stocks.

In various researches, it has been proven that news and Twitter sentiments play a pivotal role in shaping and predicting the future stock of a company. The fundamental idea behind this is to recognize emotional  patterns in these tweets and find a correlation between them and predict the future behavior of various stock prices.

## II.     LITERATURE SURVEY

[1] Pranjal Chakraborty, Ummay Sani Pria have predicted stock movement using sentiment analysis based on the Twitter feed. They have collected 1.5 million tweets through Sentiment140 API which contained 'stock market', 'StockTwits', 'AAPL' keywords from Jan. to Dec. 2016 for predicting stock movement. They also performed a random split over the dataset to divide the dataset into training and testing datasets. Later they used DJIA (Dow Jones Industrial Average), APPLE Inc. to closing stock index difference from Jan to Dec 2016. They tested on similar stock-related tweets on a different timeline to check how much they can predict the stock index. However, it cleared that extremely high and low difference is difficult to predict with Boosted Regression Tree.

[2] S. Naveen Balaji, P. Victer Paul, R. Saravanan have surveyedSentimental analysis based Stock Prediction using Big Data Analytics. They proposed a  new technique to recognize the hidden patterns, business information, user preference, etc. in accordance with Big Data is referred to as Big Data analytics. Big Data Analytics was classifiedinto 3 types i.e. Predictive, Prescriptive,  and  Descriptive. Predictive Analysis was used in various applications such as fraud detection, risk management, economic variation, etc. The accuracy obtained by them was 86.7% in predicting

stock market prices and a mean percentage average (MAPE) error was reduced by more than 6%. They concluded that predictive analysis will be the best analytical technique to predict stocks in advance.

[3] Venkata Sasank Paglou, Kamal Nayan Reddy Challa, Ganapati Panda used twitter for public sentiments sentiment and performed sentiment analysis on it for stock market movements. They found out that the main problem was the analyzers are trained with different corpus. So they developed their own sentiment analyzer. They classified tweets into negative, positive, and neutral sentiments. They annotated the polarities as 1 for positive, 0 for neutral, and 2 for negative emotions. They concluded that a strong correlation exists between the rise or fall in stock prices of the company. The main thing of their project was to develop a sentiment analyzer that judges the type of sentiment present in tweets. Former they claimed that positive emotions or sentiments of the public on Twitter about a company might reflect to stock price.

[4] A Sarkar, A K Sahoo, S Sah, C Pradhan in LSTMSA: A Novel Approach for Stock Market Prediction Using LSTM and Sentiment Analysis has stated that future variations of the stock market are independent of previous performance. They performed stock prediction for the Google stock using LSTM with and without taking news headline sentiments into account and found out that news sentiments are capable of explaining about 40% of the variances in the stock market. For preventing overfitting they attempted to set a dropout of 20% in LSTM by randomly dropping a number.

[5] In Stock Price Prediction Through the Sentimental Analysis of News Articles, Jaeyoon Kim, Jangwon Seo, Minhyeok Lee, Junhee Seok attempted to check the correlation between the positive index value of a news article and stock return value and check how viable it is for practical application. For that, they have used local Korean dictionaries, such as KOSAC for solely NLP-based stock price prediction. Due to difficulties in deriving high accuracy using Korean, they arrived at an approximate value of 0.3034 correlation, and thus they created an updated dictionary for future improvements.

[6] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C. Anastasiu, in Stock Price Prediction Using News Sentiment Analysis, used ARIMA, subtypes of RNN and Facebook Prophet algorithms and compared their accuracy. They collected stock data for S&P 500 companies for 5 years and also collected more than 250,000 news articles. The RNN algorithms proved more effective than ARIMA and FB Prophet, also they suggested that low or highly volatile stock prices do not respond well to the above-mentioned models.

[7] People with good or happy moods are more likely to increase their investments. The problem with this theory is that aggregating each persons' mood can be a trivial task.

Twitter provides a small population sample and also access to publicly posted tweets. Masoud Makrehchi, Sameena Shah, and Wenhui Liao [7] proposed a novel approach for a correlation between social media texts and related significant stock market movements in the same time frame. They extracted significant stock movements and collected appropriate pre, post, and contemporaneous text from social media sources (for example, tweets from Twitter). Subsequently, they assigned the respective label (positive or negative) for each tweet. They trained a model on this collected set and make predictions for labels of future tweets. They created successful trading strategies based on this system and find significant returns over other baseline methods. Their trading strategy was able to beat S&P 500 by about 20% returns in four months.

[8] Financial markets are ought to be driven by public sentiments. In recent years, Stock movement prediction has proved to become a very popular research topic, all credits to advancements in AI and ML. But it is still impossible to predict market movements with 100% accuracy. Yet this has not stopped researchers and traders to improve techniques for analysis and prediction of the same. Rubi Gupta and Min Chen [8] investigated the impact of sentiments extracted from a large-scale collection of tweets from StockTwits for five Companies (AAPL, AMZN, GE, MSFT, and TGT). The focus was to conduct sentiment analysis on StockTwits data and understand the impact of sentiments on stock price movements. They evaluated sentiment analysis using ML techniques i.e. Naive Bays, SVM, and logistic regression, and five featurization techniques (bag of words, bigram, trigram, TF-IDF, and LSA). The combination of logistic regression and TF-IDF achieved a reasonably high accuracy level, between 75% and 85%, for all five companies. A Time series model was used, and a comparison was done between using stock price as only input and using a combination of stock price and sentiments. The latter yields improved accuracy, with a magnitude of the improvement higher for companies with more sentiments data. This provided reasonable evidence that sentiments data has a positive impact on the accuracy of stock price change prediction.

[9] At some point in their lives, every person is ought to partake in financial investments. Sensex indices movement analysis has been a key research topic for more than a decade. Media, psychology, and sentiments play an important role in market trading. Sunil Kumar Khatri, Himanshu Singhal, and Prashant Johri [9] had analyzed the e-data collected from various sites with ANN. Posts/comments on social media by different people i.e. investors and domain experts can be analyzed to know the attitude towards any share or index values. This has paved the way towards many opportunities for research as well as mining social media for trade. They tried to reduce the errors in the prediction of the least possible values. Their paper targeted the index value of BSE which can also be extended to other equities. Further, this study also can be extended to

predict not just the direction of the market for a particular day but also to predict the closing value for the day.

[10] By virtue of the flourishing evolution in Artificial Intelligence, mainly in Deep Neural networks(DNNs) like Convolution Neural Networks and long short term memory, financial sector prediction and its analysis have become popular. Yichuan Xu and Vlado Kesel [10] experimented on the performance of DNN with a combined dataset consisting of tweet sentiments on the finance domain, stock history price, and stock price technical indicator. They found that the finance tweets that are posted from market closure till the market open in the next day have more predictive power on next day stock movement. This paper successfully studies the uses and applications of attention-based LSTM DNN in the analysis and prediction of future stocks. They also built a custom dataset consisting of stock history data, tweet sentiments, and technical indicators in the US stock market. Their paper tried to answer various aspects such as time sensitivity in finance tweets with respect to stock market prediction, period of the day wherein the finance tweet sentiments have more predictive power, and the possibility of having a common model for all the stocks.

### III.    DATASET OVERVIEW

| VARIABLE | DESCRIPTION |
| --- | --- |
| Date | Stock price date |
| Open | Open stock price value for the day |
| High | Highest stock price value of the day |
| Low | Lowest stock price value of the day |
| Close | Closing share price for the day |
| Volume | Number of shares traded for the day |
| Tweets | Short descriptive tweet |
| Compound | A compound overall sentiment value |
| Negative | Negative sentiment value |
| Neutral | Neutral sentiment value |
| Positive | Positive sentiment value |

### IV.    Block Diagram
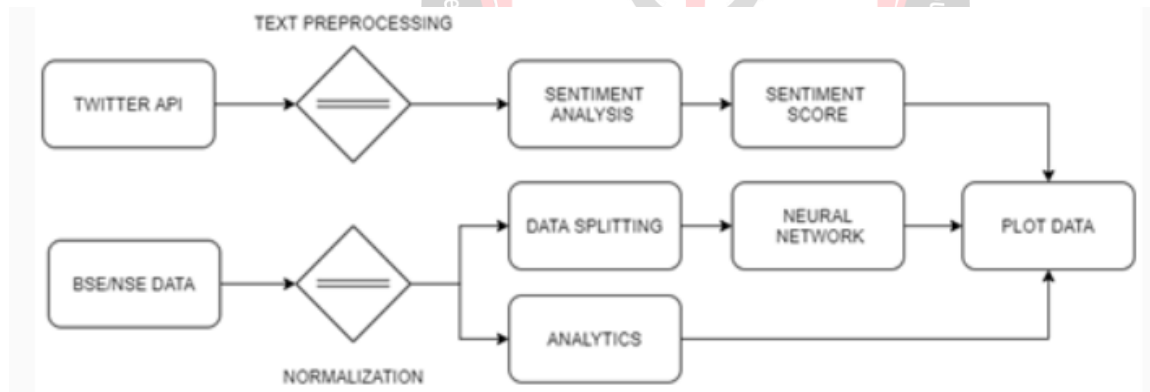


Fig 1. Block Diagram of the proposed methodology

### V.    METHODOLOGY

#### A.  Data pre-processing:

Preparing the raw data values in the dataset to be suitable for use with the models used, for example clubbing tweets of the same date together, averaging empty stock price data values due to a holiday, nullifying duplicates, and feature scaling like normalization and standardization for transforming data, etc.

#### B.  Data modeling:

To study differing outcomes of various models, models are chosen to experiment with results and find out the optimal one. Pre-processed data is shaped according to the needs of the model. Additional columns are introduced to represent essential variables. Sentiment Analysis APIs like Vader Sentiment Analyser's lexicon package can be used to perform sentiment analysis on the fetched tweets and give the necessary labels and polarities to each tweet. This data will be then combined to form a new dataset including both stock price values as well as Twitter sentiment scores for a particular day.

#### C.  Data prediction:

Suitable machine learning models can be trained on the pre-processed data. The models commonly experimented with are - Naive Bayes algorithm, LSTM, and the Random Forest regressor for time series models like Stock prediction having stock prices as well as tweet sentiments as an input. The dataset after performing feature scaling will be then split into training and testing sets, the training set being the one on which supervised learning will be implemented, and the testing set representing new data on which the model will predict necessary results.

### D. Data visualization:

The tweets fetched are then visualized into a pie chart form. This visualization helps us in identifying the positive and negative tweet percentages. Also, the end results including the actual and predicted values can be plotted in a time series graph to compare both the values, and the accuracy and errors of the model can be identified for testing the relevance of the particular model and as a basis for future stock movement predictions.

## VI. CONCLUSION

After analyzing and surveying all the papers and understanding all the author's points of view on market movement, we can safely conclude that news and Twitter sentiments do hold a significant effect on stock market movement. Also, the greater the number of tweets and Stock market data on a particular company mined, the more accurate will be the future stock prediction. Furthermore, we cannot narrow down a particular algorithm as the best used for stock predictions because different stocks may have different accuracies using a particular algorithm. After surveying the papers we observed that Random forest regressor, LSTM, SVM, and Naive Bayes were the commonly used algorithms for the time series model. Furthermore, we came to the conclusion that the use of stock price data along with twitter sentiments in a particular model yields better accuracy than conventional stock prediction methods.

## REFERENCES

[1]     P. Chakraborty, U. S. Pria, M. R. A. H. Rony and M. A. Majumdar, "Predicting stock movement using sentiment analysis of Twitter feed," 2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT), 2017, pp. 1-6, doi: 10.1109/ICIEV.2017.8338584.

[2]     S. N. Balaji, P. V. Paul and R. Saravanan, "Survey on sentiment analysis based stock prediction using big data analytics," 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), 2017, pp. 1-5, doi: 10.1109/IPACT.2017.8244943.

[3]     Pagolu, Sasank & Reddy, Kamal & Panda, Ganapati & Majhi, Babita. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 1345-1350. 10.1109/SCOPES.2016.7955659.

[4]     Sarkar, Abhijoy & Sahoo, Abhaya & Sah, Sourav & Pradhan, Chittaranjan. (2020). LSTMSA: A Novel Approach for Stock Market Prediction Using LSTM and Sentiment Analysis. 1-6. 10.1109/ICCSEA49143.2020.9132928.

[5]     J. Kim, J. Seo, M. Lee and J. Seok, "Stock Price Prediction Through the Sentimental Analysis of News Articles," 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), 2019, pp. 700-702, doi: 10.1109/ICUFN.2019.8806182.

[6]     S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[7]     M. Makrehchi, S. Shah and W. Liao, "Stock Prediction Using Event-Based Sentiment Analysis," 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013, pp. 337-342, doi: 10.1109/WI-IAT.2013.48.

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

[9]     S. K. Khatri, H. Singhal and P. Johri, "Sentiment analysis to predict Bombay stock exchange using artificial neural network," Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, 2014, pp. 1-5, doi: 10.1109/ICRITO.2014.7014714.

[10]     Y. Xu and V. Keselj, "Stock Prediction using Deep Learning and Sentiment Analysis," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 5573-5580, doi: 10.1109/BigData47090.2019.9006342.