Twitter sentiment analysis for stock prediction

Sanjam Singh*, Amandeep Kaur**

*(Bachelor of Engineering in Information Technolgy, Chandigarh University (CU), Mohali, Punjab, India
[1]18BET1102@cuchd.in)

**(Assistant Professor in Computer Science & Engineering, Chandigarh University (CU), Mohali, Punjab, India
[2]amandeep.e9596@cumail.in)

ABSTRACT

The stock market can be influenced and moved by a variety of variables, including external and internal influences. To overcome this challenge, a variety of data mining approaches are commonly used. Because of supply and demand fluctuations, stock prices fluctuate every second. Machine learning, on the other hand, will provide a more accurate, precise, and sensible process for resolving stock and market price concerns. ML algorithms have been employed to improve new ways of building simulation models that can estimate stock market movements and whether they will gain or lose. Support vector machines (SVM), Naive Bayes regression, and other approaches were used in several sentiment analysis research. The amount of training data available determines the efficiency of machine learning algorithms. In this research, It is shown that monitoring Twitter tweets to predict stock prices is profitable. To begin, we use Sentiment 140 Twitter data to train multiple algorithms. Since this ranked best in emotional analysis, we used SVM to determine the average mood of tweets for each trading day (0.82 accuracies). We then used an emotional analysis of a year's worth of tweets that included the phrases "stock market," "StockTwits," and "AAPL" to predict AAPL stock prices and DJIA index values. Two models, Boosted Regression Trees and Multilayer Perceptron Neural Networks, were employed to calculate the closing price difference between AAPL and the DJIA. When it comes to predicting stock prices, neural networks outperform traditional models by a significant range.

Keywords— Tweets, ML, SVM, Twitter, Sentiment Analysis, Stock Prediction

1. INTRODUCTION

A huge volume of data is exchanged online through various social media platforms in today's society. Data permeates almost every part of existence. Twitter is a microblogging service with millions of users and millions of tweets every day. Because each tweet is limited to 280 characters, Twitter generates more than 70 billion characters per day. Though each tweet may be insignificant, we may extract useful information about public opinion and sentiment evaluations about certain topics.

One of the most active and efficient forms of business is the stock market, often known as the stock exchange [1]. To make money while limiting risk, small businesses, investors, and banks must employ a challenging technique. According to the EMH, stock market prices are significantly affected by new evidence and resemble a random walk pattern [2]. Although this theory was developed and accepted by the scientific community as a need for managing markets in general, some individuals

have endeavoured to extract changes in the way stock markets function due to emerging factors [3][4].

This research used the Twitter API which provides a streaming API, to analyze financial data, and the data is continually returned. Each piece of data obtained represents the user's current condition or attitude toward a certain subject as shown in Fig1. A Twitter account and basic HTTP authentication are required to view this [5][6].

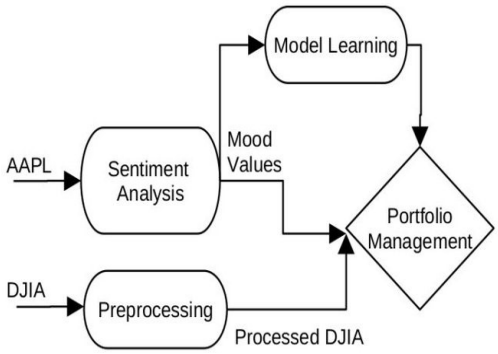

Fig 1. Flowchart.

2. LITERATURE SURVEY

Predictive analysis is a massive and demanding topic of a survey that needs the formulation of credible estimations. The development of prediction models for the global SM has advanced significantly in recent years. Traditional techniques and approaches, such as time series prediction analysis, are used in literature reviews. There are also several ML algorithms used. We derived some

information on SM analysis methodologies from the literature.

The paper [7] discussed SM moves that have a considerable influence on the market value of particular businesses due to domestic and global considerations. The research looked at three different features of Brazilian social media behaviour on Twitter: (A) total number of Tweet emotions; (B) Tweet sentiments with likes; and (C) Balanced sentiments should be acknowledged. In their endeavour to create SA in Portuguese, they employed directly the Multilayer Perceptron methodology.

The author [8] presented news SA, The stock price prediction was made by a company that analyzes stock market volatility. They also recommended that SA be used to rate articles by concatenating positive, negative, and neutral rating strings into single concatenated strings. Any stock market predictions classification model analyzes the SA's output.

In [9] the author developed TCS market price prediction based upon the following factors: start, major, minor, finish, and quantity. The research examined the effect of linear, polynomial, and radial base functions on regression models based on the optimism estimates of either the projected outcomes. With a confidence score of 0.97, the linear regression approach outperformed other strategies. Stock values change dramatically as the global market economy grows. Due to past trends and prior stock values, stock prices are difficult to predict even with expertise.

The proposed [10] method presents Word2vec and N-gram are two distinct textual representations for assessing public attitudes in tweets. The author analyzed the interaction between stock market performance and sentiments expressed in tweets using sentiment analysis and trained ML algorithms on Twitter tweets. Keywords like $MSFT, #Microsoft, and others can be used to retrieve data from Microsoft's Twitter API.

The strategy [11] developed A product's sentiment analysis is based on collecting tweets about the brand and classifying them as positive or negative sentiment. This analysis presents a proposed methodology for clustering tweets and categorizing them using data sources and data-driven training methods. For this research, 1200 tweets were reviewed and analyzed for the brand "Apple." SVM, CART, Random Forest, and Logistic regression will be compared to the suggested model. A confusion matrix can be used to compare expected and actual values.

3. METHODOLOGY

3.1. DATA

The first step is to download and extract tweets from Twitter. Once the consumer key and access token have been set up, this happens. When tweets are obtained from Twitter, special characters are removed. The tweets are then shown in a data frame along with their dates [12].

Since they are "A" and "B," the tweets are labelled "1" and "0." A random split is used to divide the dataset into a training dataset and a testing dataset. The training dataset has 50,000 tweets, but the testing dataset has just 13,500 [13].

Table 1. Data distributions.

	Training Data	Testing Data
A	27,492	12,953
B	25,132	14,032

Between November 2021 and March 2022, tweets including the terms "stock market," "StockTwits," and "AAPL" were gathered to determine stock movement, as shown in Table 1 [14].

3.2. PROCESSING OF DATA

The data on stock prices collected is only partial due to weekends and public holidays when the stock market is closed. A simple method is used to estimate the missing data. Stock data follows a concave function in the large majority of cases. So, if the stock value on one day is x and the value on the next day is y, the stock value on the next day is y[9]. Abbreviations, emoticons, and other data such as photos and URLs are common in tweets. As a result, tweets are pre-processed to accurately represent popular moods. We used three filtering steps: tokenization, stop-word removal, and special character removal using regex matching to pre-process tweets [15].

1) Processing: Based on the quantity of available area, tweets are split down into specific words, and non-essential symbols like emojis are removed. A list of all the variables in the equation that have been removed is created. Make a list of every tweet's concrete terms.

2) Get removal of stop words: These are words that don't communicate any feeling. Terms like a, is, the, with, and others are deleted from the list of words when a tweet is split.

3) Regex matching for special character removal is a Python utility for detecting URLs and replacing them with the word URL [16].

3.3. SENTIMENT ANALYSIS

Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), Boosted Tree (BT), and Random Forests (RF) models were used to calculate the sentiment analysis. The sentiment value for the test data tweets was then projected using all of the models [17]. The formulas were used to determine this. given below in equations (1) and (2) respectively,

$$Precision = \frac{TP}{(TP+FP)} \quad (1)$$

$$Recall = TP(TP + FN) \quad (2)$$

And found a difference of 0.01 accuracy, as shown in Table 2 [13][15].

Table 2. Results of Sentiment Analysis

	Accuracy	F1 Score	Precision	Recall
LR	0.81	0.81	0.81	0.81
SVM	0.82	0.82	0.82	0.82
DT	0.73	0.73	0.73	0.73
BT	0.70	0.70	0.70	0.70
RF	0.69	0.69	0.69	0.69

4. STOCK MOVEMENT PREDICTION

The Support Vector Machine was used to extract the expected emotional values from tweets connected to the various keywords in our investigation. We also created a new dataset based on the daily average emotional value of these tweets. To assess DJIA closing difference values, the phrases "stock market" and "StockTwits" were chosen, while to project AA closing difference values, the term "AAPL" was implemented [18].

The next day's stock price variations were predicted by the method. The training data set's average sentiment ratings are based on a day's worth of tweets, with the closing price difference between that day and the next being the same. As a result, based on the emotional impact of today's tweets, we can predict how much the stock market will succeed or fail the next day. To put it differently, to estimate the current day's stock value, we'll need the prior day's average net cost [19].

Although the Boosted Regression Tree model is trained on average sentiment values, the tweets obtained for testing were recognized using SVM. Our Boosted Regression Tree model used these average marginal values to calculate the stock difference for the next day. The average sentiment values of tweets including the terms "stock market" and "StockTwits" are trained using the DJIA's closing price difference, but not those of tweets featuring the term "AAPL".

During the testing period, which lasted from November 2021 to March 2022, we plotted both real and expected stock variances. In addition, the MAE and RMSE between real and projected stock variations were calculated using the Boosted Tree model and the MLP regression model, as shown in Table 3 [20].

Table 3. Prediction Results of Stock

Tweets with #	Boosted Tree	MLP regression
stock market	86.17	65.97
StockTwits	79.11	75.71
AAPL	1.41	0.96

5. CONCLUSION & FUTURE WORK

The essential sentiment analysis is obtained with the help of Twitter APIs. There are already a variety of techniques for constructing stock models, which we will leave for later. Some of them include creating a business model by grouping companies based on their business, accounting for the negative impact on a company's stock price due to news about other companies in the same industry, and examining more general industry and global news that could indicate market stabilization [2][7].

We managed this by collecting a set of stock-related tweets and using SVM to get the average sentiment value. The training set was obtained by combining those tweets with today's and tomorrow's DJIA or AAPL closing stock index adjustments. Then we examined similar market-related tweets to determine if the stock of the index could be accurately estimated by algorithms [10][17].

Moreover, the neural network can outperform the boosted regression tree. For all three datasets containing the terms "stock market," "StockTwits," and "AAPL," the Multilayer Perceptron Neural Network model had reduced MAE and RMSE

compared to the Boosted Regression Tree model. Our results also reveal that identifying extremely significant rather than too many variances in stock indexes using a boosted regression tree is challenging. Even on those days, our models performed well on the data set provided [19][20].

Acknowledgements

An acknowledgement section may be presented after the conclusion if desired.

REFERENCES

[1] A. Nayak, M. M. M. Pai, and R. M. Pai, "Prediction Models for Indian Stock Market," *Procedia Comput. Sci.*, vol. 89, pp. 441–449, 2016, DOI: 10.1016/j.procs.2016.06.096.

[2] A. Srivastava, V. Singh, and G. S. Drall, "Sentiment analysis of Twitter data: A hybrid approach," *Int. J. Healthc. Inf. Syst. Informatics*, vol. 14, no. 2, pp. 1–16, 2019, doi: 10.4018/IJHISI.2019040101.

[3] S. V. Kolasani and R. Assaf, "Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks," *J. Data Anal. Inf. Process.*, vol. 08, no. 04, pp. 309–319, 2020, doi: 10.4236/jdaip.2020.84018.

[4] J. Bai *et al.*, "Object Detection in Large-Scale Remote-Sensing Images Based on Time-Frequency Analysis and Feature Optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, doi: 10.1109/TGRS.2021.3119344.

[5] S. Pacharkar, P. Kulkarni, Y. Mishra, A. Jagadambe, and S. G. Shaikh, "Predicting Stock Market Investment Using Sentiment Analysis," *Int. J. Adv. Res. Comput. Commun. Eng. ISO*, vol. 3297, pp. 109–114, 2007, doi: 10.17148/IJARCCE.2018.7321.

[6] S. Selvin, R. Vinayakumar, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, no. January 2018, pp. 1643–1647, 2017, doi: 10.1109/ICACCI.2017.8126078.

[7] A. E. O. Carosia, G. P. Coelho, and A. E. A. Silva, "Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media," *Appl. Artif. Intell.*, vol. 34, no. 1, pp. 1–19, 2020, doi: 10.1080/08839514.2019.1673037.

[8] P. G. S. Mate, "Issn No : 1006-7930 STOCK PREDICTION THROUGH NEWS SENTIMENT ANALYSIS .," vol. XI, no. Viii, pp. 36–40, 2019.

[9] D. Bhuriya, G. Kaushal, A. Sharma, and U. Singh, "Stock market predication using a linear regression," *Proc. Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2017*, vol. 2017-Janua, pp. 510–513, 2017, doi: 10.1109/ICECA.2017.8212716.

[10] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," *Int. Conf. Signal Process. Commun. Power Embed. Syst. SCOPES 2016 - Proc.*, pp. 1345–1350, 2017, doi: 10.1109/SCOPES.2016.7955659.

[11] R. Soni and K. J. Mathai, "Improved Twitter Sentiment Prediction through Cluster-then-Predict Model," vol. 4, no. 4, pp. 559–563, 2015, [Online]. Available: http://arxiv.org/abs/1509.02437.

[12] Y. Lu and J. Chen, "Public opinion analysis of microblog content," *ICISA 2014 - 2014 5th Int. Conf. Inf. Sci. Appl.*, pp. 1–5, 2014, doi: 10.1109/ICISA.2014.6847451.

[13] G. Krishna, "System," *2018 Int. Conf. Adv. Comput. Commun. Informatics*, pp. 1981–1985, 2018.

[14] J. Liu, Z. Lu, and W. Du, "Combining enterprise knowledge graph and news sentiment analysis for stock price volatility prediction," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2019-Janua, pp. 1247–1255, 2019.

[15] X. Wang and X. Luo, "Sentimental space based analysis of user personalized sentiments," *Proc. - 2013 9th Int. Conf. Semant. Knowl. Grids, SKG 2013*, pp. 151–156, 2013, doi: 10.1109/SKG.2013.16.

[16] Q. Li, L. L. Jiang, P. Li, and H. Chen, "Tensor-based learning for predicting stock movements," *Proc. Natl. Conf. Artif. Intell.*, vol. 3, no. 2004, pp. 1784–1790, 2015.

[17] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," 2004, [Online]. Available: http://arxiv.org/abs/cs/0409058.

[18] Twitter Business Basics (n.d.). https://business.twitter.com/en/basics.html

[19] Agarwal, A., Xie, B., Vovsha, I.., Rambow, O. and Passonneau, R. (2011) "Sentiment Analysis of Twitter Data." *Proceedings of the Workshop on Languages in Social Media LSM'11, Stroudsburg, PA,* June 2011, 30-38.

[20] Ghiassi, M., J. Skinner, and D. Zimbra. "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network", *Expert Systems with Applications*, 2013.