Stock trend prediction based on social media articles

Abhinav Yalamaddi
abhi.yalamaddi@gmail.com

Satwik Arvapalli
satwik.arvapalli@gmail.com

***Abstract*—This paper presents a machine-learning approach to predicting stock trends using social media articles. The proposed method consists of three steps: data acquisition, data processing, and sentiment analysis. First, relevant social media articles are collected from different sources. Then, the acquired data is preprocessed by filtering out stop words, punctuation, and non-alphanumeric characters. Finally, sentiment analysis is conducted using the Classification algorithms to determine the sentiment of the article and the corresponding stock trend. The experimental results demonstrate that the proposed method yields a satisfactory prediction accuracy for stock trends.**

***Index Terms*—Classification, Prediction, Data Scrapping, Stock Trend.**

I. Introduction

The purpose of this paper is to discuss the use of machine learning algorithms for classifying tweets in order to predict stock prices. This paper focuses on the development of a model that can accurately classify tweets into positive and negative categories in order to predict the future trend of stocks. The goal of this project is to develop a model which can effectively analyze tweets and accurately predict stock prices.

In recent years, the use of social media, particularly Twitter, has become increasingly popular for stock market analysis and prediction. Twitter has become an important source of data for stock market analysis due to its high volume of data and its ability to capture sentiment. There have been several studies on the use of Twitter for stock market analysis and prediction. These studies have shown that Twitter data can be used to accurately predict stock prices.

However, in order to use Twitter data for stock market prediction, it is necessary to classify the tweets into positive, negative, and neutral categories. This is where machine learning algorithms come in. Machine learning algorithms are useful for classifying tweets into the correct categories. This can be done by feeding the tweets into a machine learning algorithm which is trained to categorize them.

The goal of this project is to develop a model that can accurately classify tweets into the correct categories so that we can predict whether a stock will go high or low only. The model will use machine learning algorithms to classify the tweets into the correct categories. The model will then use the classified tweets to predict stock prices.

The project will involve the development of a model that can accurately classify tweets into the correct categories. The model will be trained using a dataset of tweets and stock prices. The model will then be tested to evaluate its accuracy. The results of the tests will be used to determine the effectiveness of the model for stock market prediction.

The paper will discuss the development of the model, the evaluation of its accuracy, and the results of the tests. The report will also discuss the implications of the model for stock market prediction.

II. Dataset

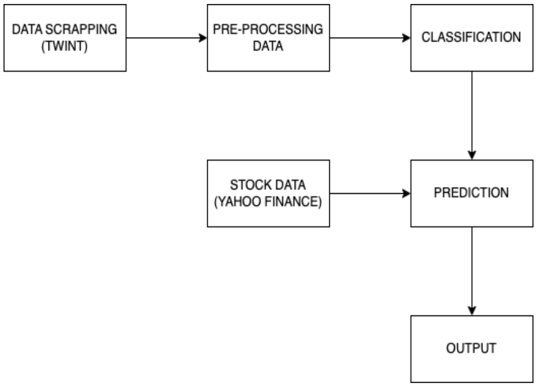
In this project, we used two main datasets:

1) The Twitter data set utilized in this paper includes publicly available tweets on Tesla, Inc. collected from Twitter.com between January 1, 2019 and July 14, 2019. The search term for tweets contains the following words: ”Tesla”, ”Elon Musk”, ”@Tesla”, ”@elonmusk”, ”$TSLA” and ”#Tesla”. The idea for this is that we want to capture a broader picture of sentiment in order to acquire a more overall attitude about the company rather than just tweets specifically linked to Tesla stock. The data includes the timestamp, username and tweet text for every tweet during that period.
2) The Tesla stock data is obtained from Yahoo Finance, and contains the date, opening, high, low, closing price and volume.

Aggregated values and sample data from Twitter		
Timestamp	Username	Tweet
2019-01-01 15:38:02	AccessWallST	Tesla (TSLA) Stock Sinks As Market Gains

Aggregated values and sample data from Tesla					
date	opening	high	low	closing	volume
2019-01-02	20.40	21.00	19.92	20.67	174879000

III. Proposed model



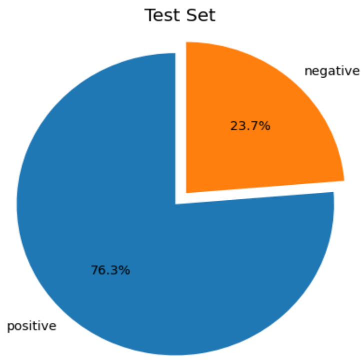


Fig. 1. Class Distribution in Test Data

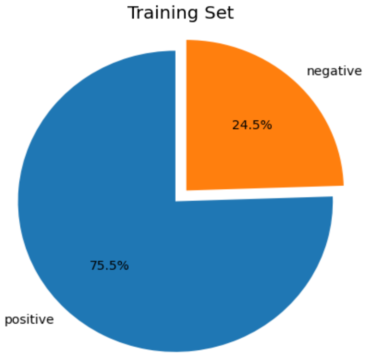


Fig. 2. Class Distribution in Train Data

A. Data Scrapping and Preprocessing

First, we collect data from Twitter (This dataset is collected using a custom code which uses a Twitter API). This social networking platform was chosen for its conciseness. After collecting data, we use a reduction method to clean it up by deleting spam, redundant, nonsensical, or irrelevant tweets. The following are also included in the preprocessing step:

- Tweets are converted into word tokens using nltk, which means the model evaluates two tokens/words at the same time. This indicates that if a tweet labels something as ”not good,” it will be regarded a negative remark rather than a positive one just because the word ”good” appears in it.
- Tags like ”@author”, URL’s are removed. These labels must be removed since they don’t provide any useful information for determining feelings.
- Stop words are eliminated. Stop words (such as an, is, are, the, etc.) are often seen in tweets and provide no useful data for ML classifiers.

We then created a Word-List of all the words appeared in positive and negative tweets respectively along with their frequencies. Used the Word-List generated as input of the model and classified the sentiment of the tweets by using models discussed in this paper.

IV. Model Learning and Prediction

We used multiple algorithms to build this model, including Naive Bayes, logistic regression, gradient boosting, AdaBoost, Decision trees and Random Forest. Each algorithm has its own strengths and weaknesses, and by combining them we were able to create a robust model that works well for our data. The different algorithms also allowed us to evaluate different aspects of the data and identify various patterns that can be used to improve the accuracy of the model.

1) Naive Bayes: We used a Naive Bayes algorithm to classify tweets. This algorithm is simple and efficient and is useful for classifying text data as it assumes that the occurrence of each word in a tweet is independent of the occurrence of other words, which simplifies the calculations. For our tweet classification, we used a bag-of-words approach which counts the frequency of words in the text, and then calculates the probability of the class given the words. We then used this probability to classify each tweet into one of the two classes, positive or negative. Our model was able to accurately classify tweets with an accuracy of 89.8%.

		Classifier Prediction	
		Positive	Negative
Actual	Positive	739	24
	Negative	78	159

2) Logistic Regression: We used a logistic regression algorithm to classify tweets as the algorithm was able to analyse the content of tweets and assign them to different categories based on the text. We used a set of labelled tweets to train the algorithm and then used it to classify new tweets. The algorithm provided accurate results and was easy to implement. Our model was able to accurately classify tweets with an accuracy of 76.3%.

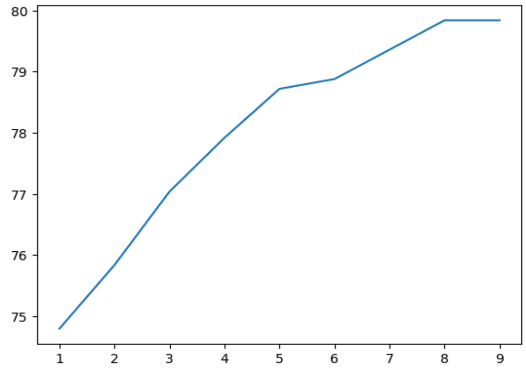
		Classifier Prediction	
		Positive	Negative
Actual	Positive	109	214
	Negative	10	971

3) AdaBoost: We used a gradient boosting algorithm as this approach combines multiple weak learners to create a strong model that can accurately classify tweets into positive and negative categories. It is an effective way to classify text data as it can learn from past mistakes and quickly identify important features. It also has the added benefit of being able to handle large datasets. We are able to accurately classify tweets with an accuracy of 81.12%.

		Classifier Prediction	
		Positive	Negative
Actual	Positive	144	179
	Negative	58	869

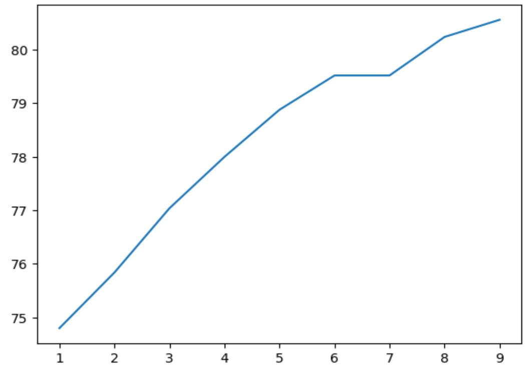
4) Decision Tree: The decision tree can then be trained on a data set of labeled tweets to identify which words or combinations of words indicate a positive or negative

sentiment. Once trained, the decision tree can be used to classify new tweets as either positive or negative. We implemented decision on the basis of both gini index and mutual information values.



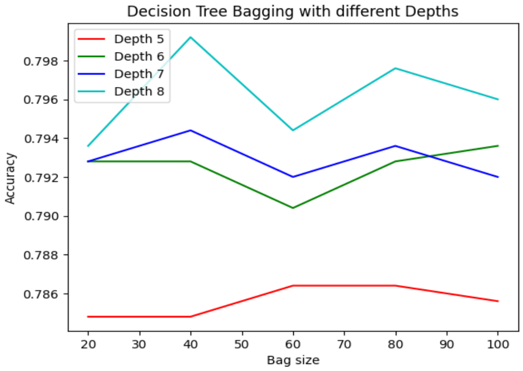
Accuracy vs Depths (Entropy)

As we can see in the above graph when accuracy was plotted against the depths of the decision tree; after depth 8 the accuracy got saturated.

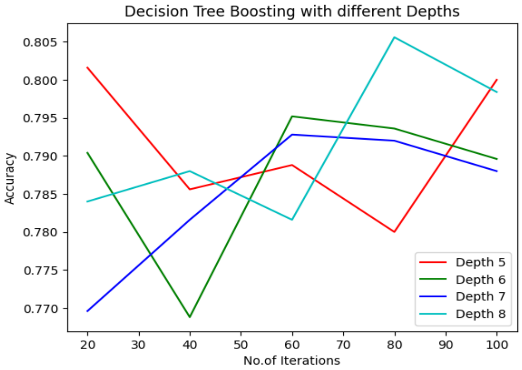


Accuracy vs Depths (Gini)

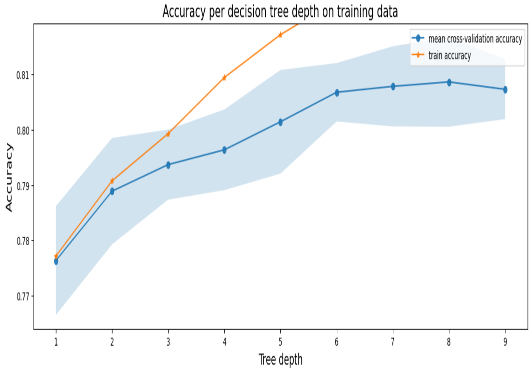
As we can see in the above graph when accuracy was plotted against the depths of the decision tree; even after depth 9 the accuracy not got saturated. We took a next step by doing bagging and boosting on multiple bag sizes and iterations for different depths.



Accuracy's of bagged decision trees were plotted against the bag size for each depth.



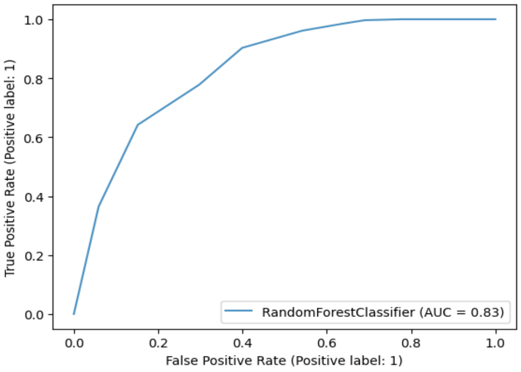
Accuracy's of boosted decision trees were plotted against the number of iterations for each depth.



Crossed validation was performed on the training data the mean cross validation accuracy was plotted along with the training accuracy against tree depths. We are able to accurately classify tweets with an accuracy of 81.12%

		Classifier Prediction	
		Positive	Negative
Actual	Positive	144	179
	Negative	58	869

5) Random Forest: We are able to accurately classify tweets with an accuracy of 82.64%.



Since random forest performed almost as efficient as Naive Bayes, we plotted the ROC curve as in figure.

		Classifier Prediction	
		Positive	Negative
Actual	Positive	144	209
	Negative	8	919

Comparing Model Accuracy’s implemented	
Model	Accuracy
Naive Bayes	89.8%
Random Forest	84.64%
Decision Tree	81.52%
AdaBoost	81.12%
Logistic Regression	76.3%

We can notice that the class imbalance was also main reason for relatively lesser accuracy’s.

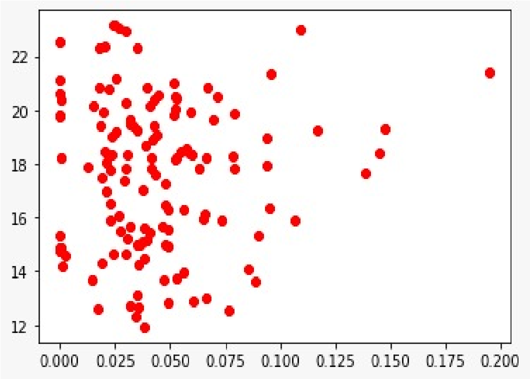


Fig. 3. Co-relation between tweets sentiments and closing stock price

To accomplish this, we take the average sentiment across all the tweets for each day and used the average sentiment to predict the change in stock trend. Therefore, if our hypothesis is correct, we would see the percent change in stock prices increase as the average sentiment increases and vice versa. If there is perfect correlation we would see a straight diagonal line. We also calculated the correlation coefficient to get a more objective measure of how the two variables correlate. The co relation coefficient calculated is 0.3 approx from that we can conclude that the stock prices are not highly co related with the sentiment probabilities but we can confirm that it cannot be neglected though.

V. Conclusion

Finally, Our analysis does not take into account many factors, such as the lack of relation to the real public sentiment due to the dataset only considering those who use Twitter and speak English. It is possible to observe a higher correlation if the actual mood is studied. It can be speculated that people’s moods indeed affect their investment decisions, thus leading to a correlation. Nonetheless, there is no direct correlation between those who invest in stocks and those who use Twitter more often, even though there is an indirect correlation - people’s investment decisions may be influenced by the moods of the people around them, i.e. the general public sentiment. All of these remain as areas of future research.

In this paper we are not directly able to predict the exact percentage raise and fall of the stock price, but we were able to predict whether the stock price might increase or decrease.

Acknowledgment

We would like to thank our professor, Dr. Sriraam Natarajan, for his guidance and support throughout the project duration. His expertise and advice have been instrumental in helping us to understand the nuances of our project and in producing this paper. We are immensely grateful for their encouragement and assistance.

References

[1] S. Coyne, P. Madiraju and J. Coelho, ”Forecasting Stock Prices Using Social Media Analysis,” 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2017, pp. 1031-1038, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.169.

[2] Mehta R, Mehta D, Chheda D, Shah C, Chawan PM. Sentiment analysis and influence tracking using twitter. Int J Adv Res Comput Sci Electron Eng. 2012;1(2):72–9.

[3] Skuza M, Romanowski A. Sentiment analysis of twitter data within big data distributed environment for stock prediction. Federated Conference on Computer Science and Information Systems. 2015; pp. 1349–1354.

[4] Talamás, Juan. (2021). Social media Effects on the market: Reddit Data analysis on Stocks. 10.13140/RG.2.2.24180.88960.

[5] Mankar, Tejas Hotchandani, Tushar Madhwani, Manish Chidrawar, Akshay C S, Lifna. (2018). Stock Market Prediction based on Social Sentiments using Machine Learning. 1-3. 10.1109/ICSCET.2018.8537242.