Stock Market Prediction using Machine Learning

Rohit B R[#1], Rajeeva Shreedhara Bhat[#2], Abhishek Manohar[#3], Mamatha K R[#4]
[1,2,3]Student, [4]Assistant Professor,
Information Science and Engineering,
B M S College Of Engineering,
Bengaluru, India

ABSTRACT
A stock (also known as equity) is a security that represents the ownership of a fraction of a corporation. This entitles the owner of the stock to a proportion of the corporation's assets and profits equal to how much stock they own. Units of stock are called "shares."

In today's finance world stock trading is one of the most significant exercises. Stock market prediction is a demonstration of attempting to decide the future estimation of the stock price and other monetary characteristics related to stock trading. This paper clarifies the forecast of a stock making use of Machine Learning. The technical and central or the time series analysis is utilized by the a large portion of the stockbrokers while making the stock forecasts.

Primarily, only numerical data was being used for stock prediction. A much better and accurate way is to use of fundamental analysis i.e. to factor in the real economy into the predictions. Using news articles to predict the stock movement on a daily basis is much more reliable.

Naive Bayes, Random Forests, Perceptrons and Linear Support Vector Machines are the Machine Learning methods which have been used for prediction. Linear support vector machines are highly regarded as one of the better text classification machine learning techniques.

Keywords — *stock market, machine learning, predictive analysis, NLP*

I. INTRODUCTION

Stock Market is continuously evolving and it is necessary to keep up with the latest trends. Trading with stocks is not easy as its value keeps on changing continuously. A stock that might have gone up in green the whole day might suddenly end up going down and finish red at the end of the day. In such situations, one wants to know whether they should buy a stock or sell one. Stock analysis increases the probability of your call being right.

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information.

The efficient market hypothesis posits that stock prices are a function of information and rational expectations, and that newly revealed information about a company's prospects is almost immediately reflected in the current stock price. This would imply that all publicly known information about a company, which obviously includes its price history, would already be reflected in the current price of the stock. Accordingly, changes in the stock price reflect release of new information, changes in the market generally, or random movements around the value that reflects the existing information set.

II. LITERATURE SURVEY

In [1], the immediate impacts of news stories on the stocks based on the Efficient Markets Hypothesis are analysed. Data and text mining techniques are used.
[2] tests whether the Efficient Market Hypothesis still holds in its context. Sentiment analysis is made on various news articles.
Gidofalvi[3] implemented a text classifier based on Naive Bayes using the financial news to track changes in the stock prices and concluded that there is a strong correlation between news articles and varying trend in stock prices before and after the articles were published.
The work in [4] is focused on the relationship between the news articles (breaking news) and stock prices.
Kari Lee and Ryan Timmons[5] in 2007 trained predictors to simulate stock trading using a Bag of

Words algorithm and a Maximum Entropy algorithm. An algorithm to determine the relevance of each word in training was used. Lot of importance is given to the prediction of stocks. The price of a stock is determined by the investors.

Qicheng Ma[6] used a naive bayes classifier and the maximum entropy classifier to predict movement of stocks and found out that maximum entropy approach gave better results compared to naive bayes model.

[7] considers non quantifiable data such as financial news articles of a company and determines the trend in future stock price using news sentiment classification. [8] deals with the basic concepts with respect to statistical theory of learning and generalization.

[9] explores the use of Support Vector Machines (SVMs) for learning text classifiers from examples. It analyses the particular properties of learning with text data.[10] is used to obtain various fanatical news articles. Local and global market news is obtained from [11].

[12] Zipline is a Pythonic algorithmic trading library. It is an event-driven system for back testing.

III. ABOUT THE DATASET

Primarily, only numerical data was being used for stock prediction. This was highly inaccurate and of low technical quality. A more better and accurate way is to use of fundamental analysis i.e. to factor in the real economy into the predictions. Using news articles to predict the stock movement on a daily basis is much more reliable. Vast majority of the experiments have shown positive results which strongly suggest that newspaper articles can have a contributing factor in predicting the changes to a stock value.

A Python module was used to extract the links of a webpage which contains the articles for that particular company. Once the link extraction was completed, text extraction was done for each news article using the text extractor python module. The last step in data extraction is the writing of text to CSVs for usage.

The text extracted in the text extraction step is totally raw. Hence, cleaning of the obtained data was required. All the string replacements were applied to remove the unnecessary data. Also, selective advertisements were being removed. After cleaning up the data, the data is stored in CSV format. The reason for using the .csv format is that it is easy to write to and read from as well as easy to manipulate.

IV. HIGH LEVEL DESIGN

The Architecture of the system for training and testing the model is shown in the figure 1.

It begins with the input as search terms for a specific company whose stocks will be used for trading simulation. It cleans up the financial news article and generates a sentiment score for it. This score is used in the prediction of movement of the stock. The Reinforcement learning system uses this to decide on a trading strategy and maximize return on investment.

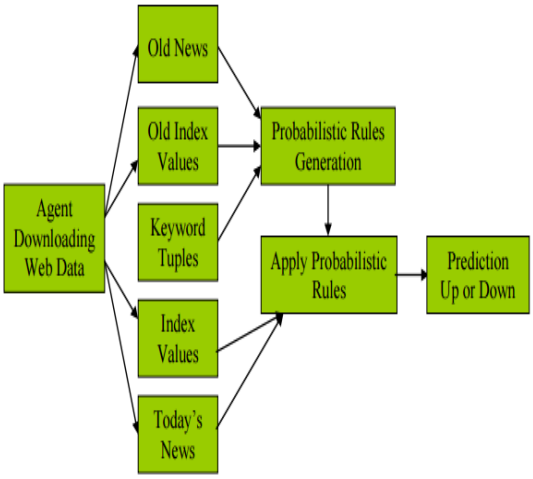

Fig 1:Proposed Design

V. METHODOLOGY

The stimulation of the trading system begins with searching for news paper articles about a specific company whose stocks will be considered for the prediction. Various news paper articles about the company are obtained as shown in figure 2.

A stemming algorithm is applied to the financial newspaper articles . Stemming helps in cleaning up the data and making it easier for sentiment analysis by segregating words with extremely similar meanings such as discourage, discouraging, etc.

The sentiment of the newspaper article is analyzed on the cleaned data. The TF-IDF-Term frequency, inverse document frequency is used. In this method, the more frequent a particular word is in multiple documents relating to the same company, the less weightage it is given.

This is particularly helpful in removing the impact of extremely common words such as ‘the’, ‘of’, and any other low information providing words. The output of this sub-system is a sentiment score for the related company.

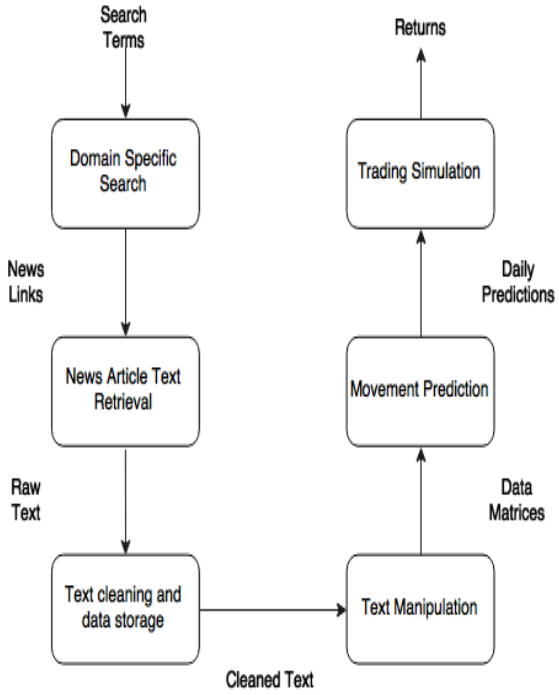

Fig 2: Actual Design

The Trading System Simulation uses the previously obtained sentiment score in deciding whether to hold the stock, short it, or buy it. The system will be compared against a baseline method of buy-and-hold as well as certain other classifiers such as Naive Bayes. Only the closing stock values will be considered.

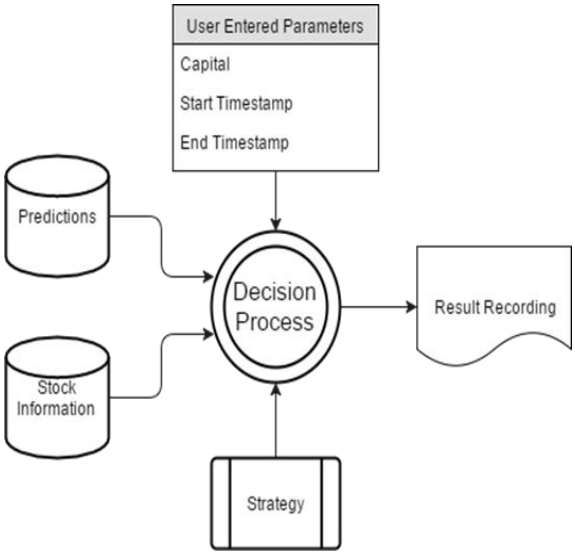

Fig 3: Block diagram of the Trading Simulation System

The above block diagram shows the major components of the trading simulation system. The roles of the six different components are:

Predictions: A container of the predicts for the security for each day of trading which will occur.
Stock Information: A container containing the details of the stock values for each day of trading.
User Entered Parameters: A container for user entered parameters such as the capital invested, as w ell as the timestamps for trading simulation.
Strategy: The trading strategy to be employed while running the trading simulation.
Decision Process: The decision process makes use of the above four to buy or sell stocks.
Result Recording: All the results from the above are logged and graphs are generated from these to be displayed on the web dashboard.

VI. RESULTS

Once the testing was performed with respect to the four classifiers used, it was found that Linear SVM usually gave the best accuracy. It was observed that the accuracy increased with the increase in training to testing ratio. The accuracies of the 4 different classifiers with different training and testing ratios for all the companies were tabulated.

The most prominent learners were the SVM and the Random Forest classifiers. Naive Bayes performed consistently subpar as compared to Linear SVM and Random Forest. The perceptron classifier was extremely erratic in nature and gave varying accuracies.

For Microsoft and Tesla, the accuracies on training ratios above 60% drastically dropped. The stocks of Amazon and Google gave the best results, accuracy-wise.

Table 1: A comparison between the average accuracies of all stocks per classifier

Classifiers	50%	60%	70%	80%
Naive Bayes	58.06862	58.49402	57.02342	63.26038
Perceptron	54.48778	56.39096	54.89534	52.83884
RandomForest	59.32456	60.47632	60.32588	62.6246
LinearSVM	60.09164	61.6288	61.45784	67.27508

Upon 80% training to testing ratio for Google, 98% accuracy on predictions was achieved. The overall accuracy of the entire system, encompassing all methods and ratios came to 60.00% with Linear SVM performing the best. A graph depicting the accuracies of the different classifiers was plotted as shown in figure 4.

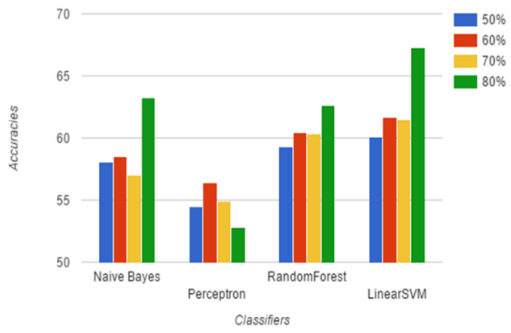

Fig 4: Comparison Of Accuracies

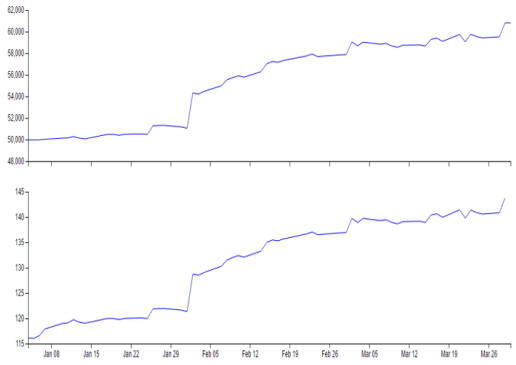

Fig 5: Comparison between portfolio value and value of stock

VII. CONCLUSION

Prediction of stock movements based on online news articles were made. Based on the results and observations, it was found that there existed a relation between news articles and stock movement. This agrees with the opinion that one of the major contributing factors in stock movement is the public perception of the company.

VIII. REFERENCES

[1] Gabriel Pui Cheong Fung, Jeffry Xu Yu, Hongian Lu. "*The Predicting Power Of Textual Information On Financial Markets*".

[2] Chen, Jerry and Aaron Chai, Madhav Goel, Donavan Lieu, Faazilah Mohamed, David Nahm, Bonnie Wu, Predicting Stock Prices From News Articles.

[3] Gidofalvi, Gyozo, "*Using news articles to predict stock price movements*." Department of Computer Science and Engineering, University of California, San Diego.

[4] Aase, Kim-Georg, "*Text mining of news articles for stock price predictions*." (2011).

[5] Timmons, Ryan and Kari Lee, "*Predicting the stock market with news articles*." CS224N Final Report (2007).

[6] Ma, Qicheng, "*Stock price prediction using news articles*." CS224N, Final Report (2008).

[7] Joshi Kalyani, H. N. Bharathi and Rao, Jyothi, "*Stock trend prediction using news sentiment analysis*." (2016).

[8] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[9] Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Cornell, Thorsten Joachims, 1998

[10] NASDAQ: Stock Exchange, URL: www.nasdaq.com [Last Accessed: May, 2020]

[11] Google Finance, URL: finance.google.com [Last Accessed: May, 2020]

[12] Zipline: Trading Simulation Library, URL: www.zipline.io [Last Accessed: April, 2020]