STOCK MARKET PREDICTION USING NEWS ARTICLES

[1]B. R. Ritesh, [2]Chethan R., [3]Harsh S. Jani, [4]Sheetal V. A.
[1]Student, Undergraduate, [2]Student, Undergraduate, [3]Student, [4] Undergraduate, Assistant Professor
[1]Department of Computer Science,
[1]BMS College of Engineering, Bangalore, India

Abstract*—Stock market prediction involves the prediction of change of stock prices. It has its applications in finance and determining the state of the economy. In our work, we aim to demonstrate a way by which we can relatively accurately predict whether a stock's price will increase or decrease on a day to day basis. This report describes our approach to make predictions from newspaper articles. We concentrate on two aspects: (1) Daily prediction of a stock's price; and (2) A system which provides a relatively high return on investment using our prediction of stocks. This technique can be used to determine the impact of natural language when it comes to the stock market.

***Index Terms*—NLP, Finance, Machine Learning**

I. INTRODUCTION

The Stock Market is always evolving and it is important to keep up with the latest trends. Day trading in stocks is risky, more so if you are untrained. Trading with stocks is not an easy process because a stock's value undergoes changes continuously, maybe changing every second. 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.

Primarily, only numerical data was used in the prediction of changes in stock prices. This was highly inaccurate and of extremely low technical quality. A more accurate way to go about this is through the use of fundamental analysis [1] i.e. to factor in the real economy into the predictions. Efficient Market Hypothesis has a strong correlation with the publication of news articles. As more financial data is available to people, at a much faster pace, it is plausible to utilize it into fundamental analysis for the prediction of changes in stock prices.

Various methodologies have been implemented to utilize these articles. The most prominent among these being sentiment analysis of newspaper articles for underlying correlations in the behaviour of stock prices[2]. These predictions have further use in simulation of trading systems to try and obtain higher returns on investments compared to the traditional methods which have been in use for the past few decades. A 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.

Most of these analyses however, have been implemented over a variety of datasets as well as different stock exchanges leading to some difficulty in comparison between them. Thus, a comprehensive analysis is required to compare these methods and determine which learning models work better.

II. LITERATURE REVIEW

The indicators of changes or trends in stock prices include textual data comprising of newspaper articles and numerical data comprising of previous stock prices. Over the past years, the indicators from numerical data have been extensively analyzed to develop more accurate means of predicting the movement of stock prices. The indicators from textual data have been comparatively underutilized.

Gidofalvi, in 2001, proposed the idea of using textual indicators (newspaper articles) to predict the change in the price of a stock. He assigned three movement classes- up, down and unchanged. Each movement class indicated the prediction of the change in a stock's value; whether it would go up or down or remain unchanged. To avoid ambiguity, he only included articles which were published within trading hours. He combined it with the volatility of a stock, its β value, and proved a strong correlation between a newspaper article and the stock price behaviour encompassing a 20 minute window before and after the publication of the article[3].

He used a Naive Bayesian text classifier to extract the indicators. Given a news article, d, the probability that a movement class c follows is given by:

$$P(M = c \mid d) = \frac{P(d|M=c)P(M=c)}{P(d)} \qquad (1)$$

Although he achieved significant classification results for alignments of articles 20 minutes prior and 20 minutes post the change in a stock price, the predictive power of the classifier was relatively low. One significant reason for the low predictive power of the classifier stems from the fact that the □-values depicting volatility of a stock[4], which have been used to determine the movement of a stock, do not in all cases cases model the relative movement of a stock correctly.

Kari Lee and Ryan Timmons in 2007 trained predictors to simulate stock trading using a *Bag of Words* algorithm and a *Maximum Entropy* algorithm. They used an algorithm to compute relevance for each word in training. The articles used were the New York Times articles obtained from the English Gigaword Corpus.

The evaluation was based on the results of two trading systems:

1. Simulated trading based on news articles predictions where each day a news article referenced at least one company in question, the "money" for that day was divided among the stocks referenced.
2. Baseline trading – all the stocks on the list received an equal share at the beginning of the month, held for the entire period, and sold at the end regardless of any change in price

The *Bag of Words* approach gave an average of 2.05% return per test month compared to the 0.615% return using the *Baseline* trading approach. The *Maximum Entropy* approach outperformed the *Bag of Words* approach. It averaged a 2.77% return per test month. Variability

in the individual test months was seen as well with the baseline approach often surpassing the simulated trading. The *Bag of Words* approach is much faster to run but takes significantly more memory to implement, thus constraining the amount of data that can be trained at once.

Although some measure of success was seen, it is not very likely to transfer over to real world success[5]. The articles were not specifically financial news articles. The data contained a large factor of noise. No segregation was made to articles which reported past happenings or simply mentioned the stock name, without having any financial impact. The author proposed training the paragraph weights for the *Bag of Words* approach, but lacked the programming time to do so.

Qicheng Ma in 2008 classified the stock movement prediction as a NLP classification task. They used news articles from the Wall Street Journal and the Reuters Financial corpus as the data for their predictions.

A Naive Bayes classifier and a Maximum Entropy classifier was being used. For every word occurred in d, where d is restricted to a paragraph, the probability of a conditional class P(c|d) is given as:

$$P(c|d) \sim \prod_{w \in d} P(w|c)P(c) \quad (2)$$

The MaxEnt classifier labelled each word in a relevant paragraph and then the weighted averages or the plurality across all the words was taken to determine the class, i.e. each paragraph was classified as positive, negative or neutral.

Cross-validation style simulations were ran. Out of the three years of data, 1 month of stock figures and articles were put aside for testing while the remaining were trained. This was done for every month in the three years of data. Intrinsic evaluation of the classification task was done as well as the extrinsic evaluation of calculating the average monthly return.

The MaxEnt classifier significantly outperformed the Naive Bayes classifier with a 5.2% average return compared to a 2.6% average monthly return[7]. The Naive Bayes classifier performed worse than the baseline in these simulations. It produced the same average return per month, but had a much higher volatility.

Although statistically significant results were achieved, the implementation was done only for the Microsoft (MSFT) stock and not for general applicability[6]. There also exists the deviation produced by using data from articles which are chronologically ahead of the month used for testing. Ideally, only past data should be used for the training.

Kalyani Joshi, Prof. Bharathi H. N and Prof. Jyothi Rao also did a prediction of stock movement prediction using news articles. Different techniques such as Naive Bayes Classifier, Random Forest algorithm, sentiment analysis and SVM were being used.

A lot of preprocessing was also being done.
Tokenisation of the document to operate on a word level, creating a stop-word list to clean the data, TF-IDF(Term frequency - Inverse Document frequency), Stemming algorithm and polarity detection algorithms were being used to get the text data in a more structured form.

Once the preprocessing process was completed, a sentiment detection algorithm was being used. An algorithm to calculate the sentiment score of a document was developed. to reduce the complexity, a transformation from full text version of the document to document vector was made.

The test data was utilized by three classification algorithms- Naive Bayes Classifier, Support Vector Machine and Random Forest Classifier. All three classifiers were implemented for various degrees of cross validation, data splitting as well as including new test data. Predictably, the Naive Bayes method obtained the weakest degree of classification accuracy at 75% accuracy on new data. The Random Forest method performed slightly better at 80%. Support Vector Machine proved to be the most accurate classifier for new test data, achieving 90%.

One of the most prominent takeaways from this scientific paper was the hit rate achieved. A statistically significant hit rate was obtained by each of the classifiers. Most of it can be chalked up to the pre-processing of the data. Of course, only the data of Apple Inc. was taken for this experiment.

Table 1 Review of methodologies used

Evaluation	**Reference**	**Model**
Movement Prediction	Gidofalvi	Naive Bayes
Movement Prediction Trading Simulation	Kari Lee, Ryan Timmons	Bag of Words Maximum Entropy
Movement Prediction Trading Simulation	Qicheng Ma	Naive Bayes Maximum Entropy
Movement Prediction	Kalyani Joshi et al.	Naive Bayes SVM Random Forest

III. CONCLUSION

Considering the benefits offered by natural language processing in sentiment analysis, we try to use the same for stock price estimation. We also observed the fact that Naive Bayes classifier does not produce much success or benefit when compared to the other algorithms being used. We also observed another important fact that preprocessing of unstructured text data from the news articles will definitely enhance a given classifier and give that extra improvement required in the performance of a model. This paper has surveyed the various techniques for stock market prediction such as Data Mining, Naive Bayes Classification, Maximum Entropy Model, Bag of Words, Random Forests, Markov Processes and Support Vector Machines to forecast the market prediction of stocks.

IV. ACKNOWLEDGEMENT

We are grateful to BMS College of Engineering for having provided us with the facilities needed for the successful completion of this Survey paper. The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

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).