*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**


# MEASURING THE SOCIAL MEDIA SENTIMENT INFLUENCE ON STOCK MARKET PRICE PREDICTION

Kostubh Tomar
M.Tech Scholar,
Department of Computer Science and Engineering,
Compucom Institute of Technology and Management

Abhishek Sharma,
Assistant Professor,
Department of Computer Science and Engineering,
Compucom Institute of Technology and Management

Dr. Akash Saxena
Professor
Department of Computer Science and Engineering,
Compucom Institute of Technology and Management

*Abstract*—**The machine learning is become more and more popular for understanding the patterns and information in various complex applications. In this work we are utilizing the machine learning techniques for understanding the patterns of stock market price. The stock market price prediction is one of the most fluctuating in nature. That incorporates the different influencing factors which make it much dynamic to understand. Therefore in this presented work for understanding the trends of the stock market data we have use the social media data analysis techniques for more accurate data prediction. The proposed model includes the natural language processing (NLP) techniques and machine learning algorithms for first computing the orientation of the sentiments of the social media. Then the common regression method has been employed over the yahoo financial data. The financial data prediction will be used with the social media sentiment orientation and the word sentiment scores for correcting the error in predictions. The implementation of the proposed model has been utilized the JAVA technology and API implementation for demonstrating the effectiveness of the prepared model. The comparative performance study among two other similar variants of stock market price prediction models demonstrates the superiority of the proposed model. The model will be evaluated and compared in terms of accuracy, error rate, and memory and time consumption. The model provides up to 5-8% improvement into the accurate prediction.**

*Keywords*—**social media sentiment, stock market price, prediction, market sentiments, performance improvements, predictive solution design**

## I. INTRODUCTION

Machine learning techniques are enabling us to analyze the data and obtain the essential patterns for the applications. These patterns are used in applications for prediction, classification, clustering, and relationship building. The prediction and classification is the task of supervised learning algorithms. The supervised learning techniques are utilizing the historical pre-identified data to learn the data patterns [1]. Additionally after learning the learning algorithm performs classification and/or prediction. However, the prediction is a task of calculating the continuous variable, on the other hand the classification task involve the prediction of a specific variable based on which the algorithm had learnt [2].

In this work, the ML technique has used for prediction based techniques design. The proposed predictive technique is demonstrating how ML algorithms will be used for stock market close price prediction. The stock market price prediction is one of the most complicated domains due to influence of various other real world scenarios. Additionally it is also fluctuating with the parameters which are outside of the market scope. Therefore, the prediction of stock market price is a complicated task. In order to demonstrate the predictive technique we have collected the data form YQL (yahoo query language) and used with the machine Learning techniques for accurate prediction. This study provides the

*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**

detailed steps and formulation strategy to accomplish the required accurate prediction.

## II. PROPOSED WORK

The proposed work is motivated by the predictive data analysis algorithms of the machine learning in order to understand the highly fluctuating data patterns. In this context the stock market price data has been considered for experimentation and system design. This chapter involves the detailed design and algorithm description of the proposed model.

### A. System Overview

The financial stock market price prediction is one of the complex natures of task which is never being accurate. The complexity involves various direct and indirect features which will suddenly impacting the performance of the stock market price. The political, social, financial, natural events are majorly influencing the stock market price. Additionally this event are suddenly fluctuating the existing patterns of the stock market prices. Therefore the analysis of the additional influencing parameters is an essential subject of study in predictive data analysis. On the other hand the social media is become more popular among the people. The social media is a place where anyone can post the information. Therefore it is become one of the popular data sources which are used for sensing the stock market influence in different real world events.

Therefore, in this presented work we are proposing a new method for improving the stock market price prediction techniques which will use the only financial data as well as financial and social media sentiment indicators. In this context the proposed work involved the method of social media text data analysis technique and the continuous data prediction techniques. The proposed model first extracts the social media data as well as the Google search data for collecting the information about the company's stock market price. Additionally the NLP based models are used for recovering the sentiment score of the collected information. The sentiment score has used as financial influence indicator. Additionally the predictive algorithm has involved for prediction of the initial stock market price based on trend. Both the consequences are combined for making final prediction. This section provides the basic overview of the proposed data modeling. Additionally the next section describes the proposed methodology for conducting the required prediction.

### B. Proposed Methodology

The architecture of the proposed working model is demonstrated in figure 2.1. Additionally their components details are discussed in this section.

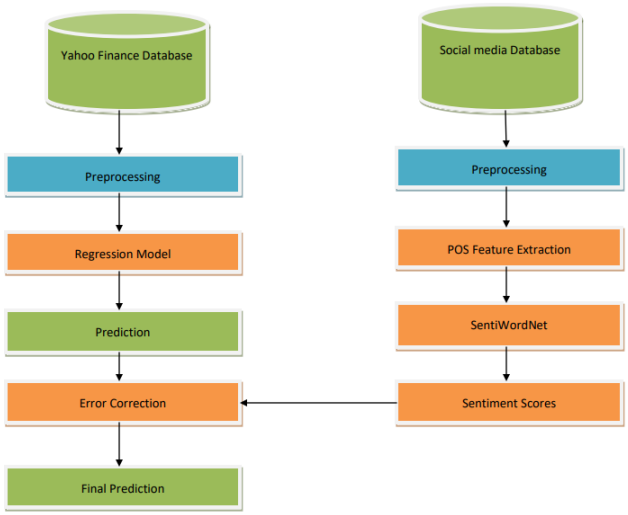


Figure 2.1 proposed social media based stock market

The proposed work is motivated to enhance the technique of stock market price prediction. Therefore an improved model has been proposed for implementation. The following components are included in this work:

**Yahoo financial data base:** the Yahoo is the one of the open source data base which will provide the financial data for different company stock prices. The Yahoo provides the API for making the query to the database and extracts the data for the target company. The API has utilizes the company code, start date and then end date for extracting the stock market price data from the database.

**Preprocessing:** the data obtained from the online source can be noisy in nature, thus it is needed to be clean the data for utilizing with the proposed model for prediction. In this context, the missing values in the extracted data has been calculated and removed from the database.

**Regression model:** the regression is one of the classical and essential techniques which are used for making prediction of the continuous values. The regression models will usages the line equation for calculating the relationship between two instances of data. Additionally based on the calculated relationship the model has predicted the next instance of the prediction. In order to understand the relationship among two instances of data line equation will be used.

**Prediction:** the prediction of the values is made on the basis of the line equation. The line equation has been defined using the following:

$$P = M * X + E$$

Where the P is prediction of the values, M is known as the slop and the E is the error which is need to be adjusted after each prediction.

*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**

**Social media database:** in this presented work we have use the twitter social media data for obtaining the information about the stock market price. Thus we have use the company names as search query into the twitter and then obtain the relevant twits form the social media data. the extracted twits has been utilized with the different tools and techniques for producing the required patterns as outcome.

**Pre-processing:** the text pre-processing is a technique, by which we can filter the text data. Additionally the text data is transformed into such format by which the data will support by the learning algorithm. Therefore in this work we first we remove the stop words from the extracted news. In next process we have removed the special characters form the text data. Finally, the data is transformed in vector to be used in further processing.

**POS tagging:** the Part of Speech (POS) tagging is the process of obtaining the POS tags from the input sentences. The aim is to measure the structural information of the text data based on NLP Stanford parser. The parsed sentences are then classified using the simple text classification technique based on k-nearest neighbor (KNN) classifier. The orientation of the model is predicted in terms of positive (+), negative (-) and neutral (+). That orientation has been used to decide which operation will be used for error correction in further error correction phase based equations.

**Sentiwordnet:** the sentiwordnet is a open source API which is used to compute the sentiment score in terms of positive and negative as well as the their values are also captured using this API. Thus this API has used here for generating the score for the words extracted from the social media text. The scores are further used for improving the performance of the predictions based on previously predicted financial data.

**Sentiment score:** the senti-word-net is an API which consumes the data in terms of text words and produces the sentiment score of the input words. Let we have extracted N words for each news instance which is consolidated and a final score for each news instance has been computed using the following formula:

$$S_i = \frac{1}{N}\sum_{i=1}^{N} W_i$$

The sentiment score has then scaled between 0-1 scale. Thus we have use the min max normalization for scaling of the sentiment scores. The following formula is used for the required task.

$$SS_i = \frac{S_i - \min(S)}{\max(S) - \min(S)}$$

Table 2.1 sentiment score table

| Score | Label |
|---|---|
| 0.0-0.2 | Low |
| 0.2-0.4 | Mid |
| 0.4-0.6 | Normal |
| 0.6-0.8 | High |
| 0.8-1.0 | Over |

The above given table is an outlook table which is used to map the sentiment score of the predictions for correcting the error.

**Error correction:** the error correction in initial prediction is one of the essential step which will help to correct the prediction of the proposed model. The following equation will be used for performing the error correction in the previously predicted value P.

$$fP = P \pm E \pm SS_i$$

Where, P is the predicted score of stock market price, E is the error in prediction, and $SS_i$ is the sentiment score of the social media text based news.

**Final prediction:** the error corrected predictive outcome has been considered as the final prediction of the proposed data model. Finally we have computed the performance of the implemented model to demonstrate how effectively the model helps to understand the variations in the stock market price patterns.
This section demonstrates how the proposed work computes the predictions of the stock market price based on financial data analysis and social media news analysis. That model helps to improve the model accuracy of the predictions.

### C. Proposed Algorithm

The proposed algorithm for making the prediction is discussed in the following table 2.2.

Table 2.2 proposed algorithm

| | |
|---|---|
| Input: social media database D, financial database F, company name and code C<br>Output: predicted price fP | |
| Process: | |
| 1. | $T = \text{Twitter.Search}(C)$ |
| 2. | $F = \text{Yahoo.search}(C)$ |
| 3. | $Pr = \text{preprocessText}(T)$ |
| 4. | $Fr = \text{preprocessData}(F)$ |
| 5. | $P = \text{LR.predict}(Fr)$ |
| 6. | $Tag = \text{POS.Tag}(Pr)$ |
| 7. | $O = \text{KNN.classify}(Tag)$ |
| 8. | $SS = \text{sentiwordNetScore}(Pr)$ |
| 9. | if($o == \text{positive} \mid \mid O == \text{neutral}$) |
| a. | $fP = P \pm E + SS_i$ |
| 10. | else |
| a. | $fP = P \pm E - SS_i$ |
| 11. | End if |

*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**

| 12. Return fP |
|---|

## III. RESULTS ANALYSIS

This chapter provides the comparative performance study of two three different experimental scenarios. The scenarios of experiments are given as:

1. A model setup with simple prediction algorithm
2. A model setup with only two sentiment indicators
3. A model setup with five sentiment score grading

Additionally the following performance parameters are used for demonstrating the performance.

### A. Accuracy

The accuracy of the implemented models is measured using the following formula:

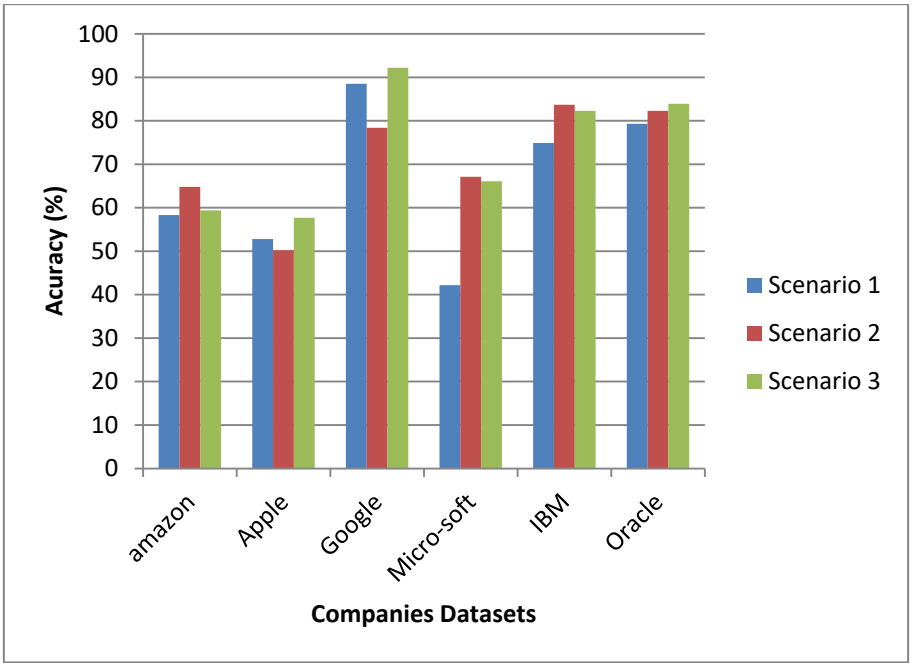


Figure 3.1 Accuracy of the models

$$\text{Accuracy } (\%) = \frac{\|\text{Actual} - \text{pridected}\|}{\text{Actual}} \text{X100}$$

The accuracy of the implemented three experimental scenarios is demonstrated in figure 3.1. The experiments with the different companies stock market price is demonstrate using YQL database and tweeter based social media data has been used. The performance of the models demonstrates the model with the sentiment indicators can improve the prediction accuracy. Additionally the accuracy becomes more accurate when we increasing the grading of sentiments. The X axis of the diagram demonstrates the company's stock price used for experiment and Y axis demonstrates the prediction accuracy.

Table 3.1 Accuracy of the models

| Companies | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| Amazon | 58.3 | 64.8 | 59.4 |
| Apple | 52.8 | 50.2 | 57.7 |
| Google | 88.5 | 78.4 | 92.2 |
| Micro-soft | 42.2 | 67.1 | 66.1 |
| IBM | 74.9 | 83.7 | 82.3 |
| Oracle | 79.3 | 82.3 | 83.9 |

### B. Error rate

The error rate demonstrates how far we are from the actual prediction of the stock market price. The error rate of the proposed experimental analysis is measured using the following equation.

$$\text{Error rate} = 100 - \text{Accuracy}$$

*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**

The error rate of models is demonstrated in the figure 3.2 and table 3.2. Both the representations containing the percentage error rate of the models.

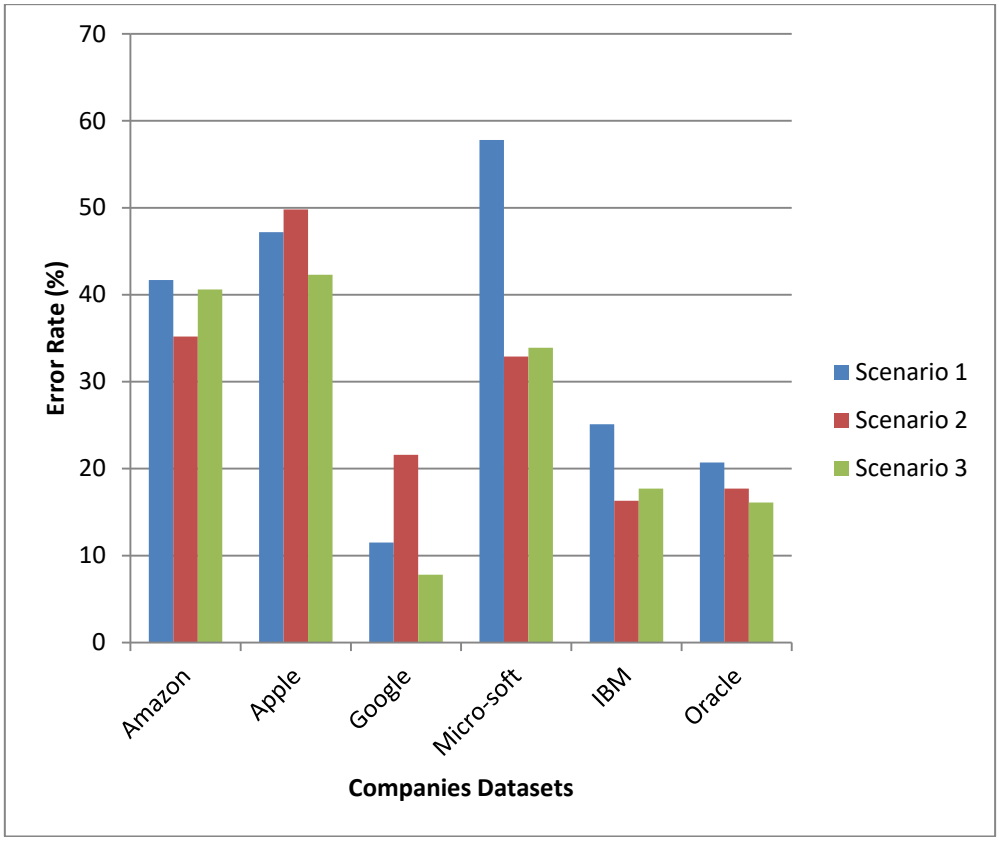


Figure 3.2 Error Rate in (%)

Table 3.2 error rate

| Companies | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| Amazon | 41.7 | 35.2 | 40.6 |
| Apple | 47.2 | 49.8 | 42.3 |
| Google | 11.5 | 21.6 | 7.8 |
| Micro-soft | 57.8 | 32.9 | 33.9 |
| IBM | 25.1 | 16.3 | 17.7 |
| Oracle | 20.7 | 17.7 | 16.1 |

In figure 3.2 the X axis contains the companies datasets used for experiment and Y axis shows the performance in terms of percentage (%). According to the obtained results with the different set of company's dataset we found that the proposed model provides more accurate results as compared to the other two implemented models for comparison.

**C. Prediction Delay (Time)**

The amount of time the predictive algorithms are required to predict the final outcome of the stock market price is known as the prediction delay. That can be measured using the following formula:

$$\text{delay} = \text{Algorithm End time} - \text{start time}$$

*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**

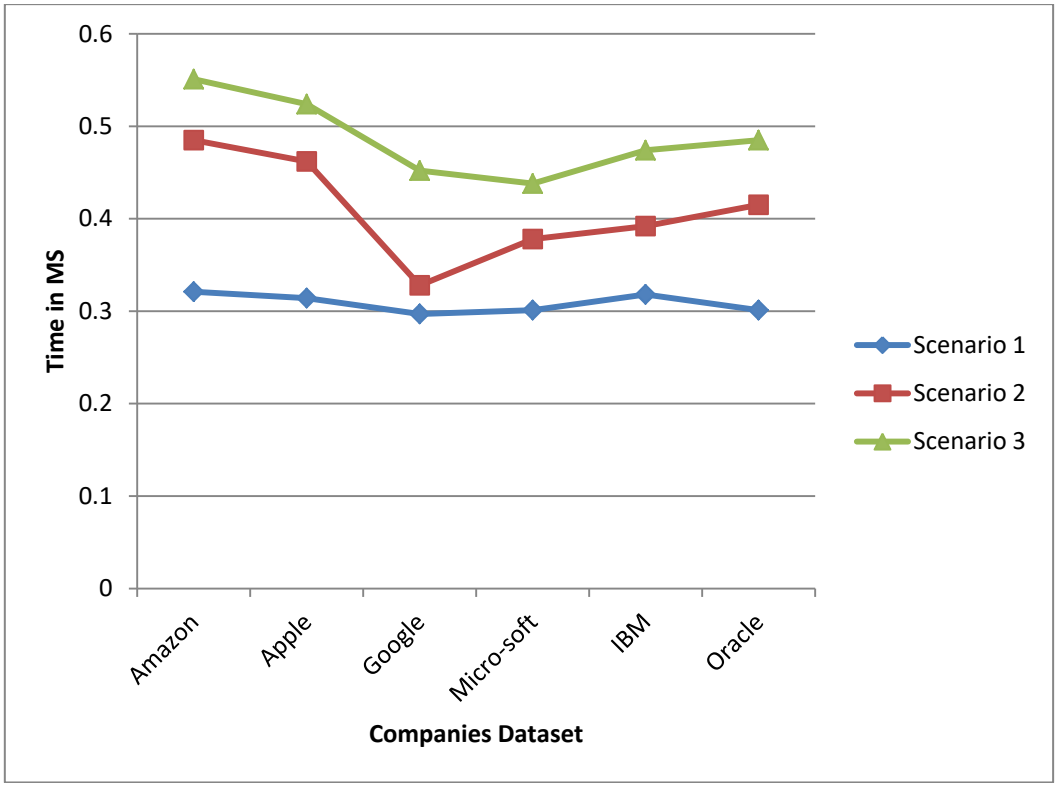


Figure 3.3 time consumption in prediction

The proposed work is measuring the time consumption of the models for predicting the next value in stock market price which is discussed here as the delay for approach. The X axis shows the different dataset utilized in the experiment and Y axis shows the prediction time in terms of milliseconds (MS). According to the experimentation of the proposed method which is demonstrated in scenarios 3 is time complex as compared to other two implemented scenarios. The proposed model is time consuming because the model compute the predicted values in two times additionally involve computation of social media sentiment score using text mining . That makes it fewer expensive in terms of time.

Table 3.3 Time consumed

| Companies | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| Amazon | 0.321 | 0.485 | 0.551 |
| Apple | 0.314 | 0.462 | 0.524 |
| Google | 0.297 | 0.328 | 0.452 |
| Micro-soft | 0.301 | 0.378 | 0.438 |
| IBM | 0.318 | 0.392 | 0.474 |
| Oracle | 0.301 | 0.415 | 0.485 |

**D. Memory**

The memory consumption is the amount of main memory which is engaged by the implemented process or algorithm. The memory used by the models has been measured using the following equation:

$$\text{memory used} = \text{total assigned} - \text{free memory}$$

*International Journal of Engineering Applied Sciences and Technology, 2024*
**Vol. 9, Issue 06, ISSN No. 2455-2143, Pages 158-166**
**Published Online October 2024 in IJEAST (http://www.ijeast.com)**

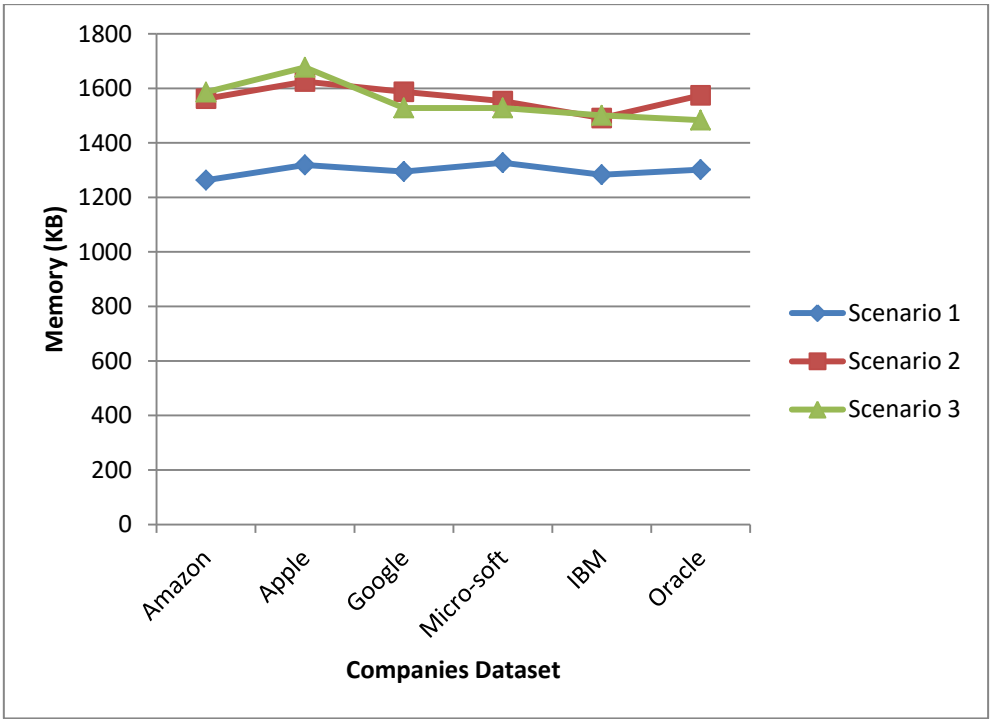


Figure 3.4 memory used

The memory consumption of the models is demonstrated in figure 3.4 an table 3.4.

Table 3.4 memory used

| Companies | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| Amazon | 1263 | 1562 | 1586 |
| Apple | 1319 | 1625 | 1677 |
| Google | 1295 | 1587 | 1528 |
| Micro-soft | 1327 | 1553 | 1528 |
| IBM | 1283 | 1491 | 1501 |
| Oracle | 1302 | 1574 | 1483 |

The memory usages of the algorithm is demonstrated in Y axis and X axis contains the memory used. The memory used is measured here in terms of kilobytes (KB). According to the obtained results the variants of predictive algorithm based sentiment analysis is consuming more memory as compared to the simple predictive model. In addition the model that is utilizing the sentiment score and their grading demonstrating higher memory consumption.

## IV. CONCLUSIONS

In this chapter we are demonstrating conclusion of the entire efforts carried out. The conclusion has been made on the basis of experimental analysis and the observation made based on literature collected for study. Additionally the possible future extension of the proposed work has been provided.

### A. Conclusion

The stock market is a classical point of attraction among data scientist and investors. The machine learning and data mining techniques bases a number of models are available in this area of study. The nature of market is making is complex to make predictions. The stock market is a very dynamic in nature and influencing with very real world events. The political, natural and accidental events are impact the prices of stock market price. Some events are positively impact and some of them are negative. Basically these are market sentiment which influencing the market prices. Now in these days the social media platform has become a very popular and low cost source of data. Therefore, social media sentiment analysis is essential for understanding the sentiments of stock market too. In this context the proposed work provides a machine learning based study to demonstrate the influence of the sentiment in stock market price prediction.

The proposed model consumes text mining and predictive data analysis technique for demonstrating how the social media sentiments can influence the stock market price prediction. The technique utilizes the stock market price prediction with sentiment indicators too make second and final prediction. The social media based score is divided into five categories which help to improve the predictive performance beyond the methods which are utilizing the text sentiment indicator based stock market price prediction techniques. The implementation of the proposed sentiment analysis based stock market price prediction technique has been performed using JAVA technology and for experimental data analysis yahoo finance API is used. Finally the performance of the proposed model is measured in terms of relevant parameters. The summary of proposed model is described in table 4.1.

Table 4.1 performance summary of model

| Parameters | Scenario 1 | Scenario 2 | Scenario 3 |
|---|---|---|---|
| Accuracy | Low | Mid | High |
| Error rate | High | Mid | Low |
| Time | Low | Mid | High |
| Memory | Low | Mid | High |

According to the obtained results the proposed model we found the that the proposed model based on the methods which are usages only positive and negative indication can be improved more when we categorize the sentiments scores in more smaller scale. The proposed model contributes to demonstrating the impact of different events in the local and global scale can impact the performance of social media performance.

**B. Future Work**

The proposed work is aimed to improve the accuracy of the existing concept of sentiment based stock market price prediction, which is successfully accomplished in this study. During the study we have also identified some extensions of the proposed model in this current work. The following research plans we have to extend the current work.

1. The proposed model currently demonstrates the online and offline social media data for simulation of the proposed concept. In near future we planned to demonstrate the model with real time price prediction.
2. The effect of more small scale sentiment division is need to be studied and how we distinguish the more accurate price.
3. Need to be implementing more light or efficient computational methods which reduce the delay in training and testing time and can learn and predict accurately with less fewer amount of data..

REFERENCES

[1] A. S. Albahri, R. A. Hamid, J. k. Alwan, Z. T. Al-qays, A. A. Zaidan, B. B. Zaidan, A. O. S. Albahri, A. H. AlAmoodi, J. M. Khlaf, E. M. Almahdi, E. Thabet, S. M. Hadi, K. I. Mohammed, M. A. Alsalem, J. R. Al-Obaidi, H. T. Madhloom, “Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review”, Journal of Medical Systems (2020) 44: 122

[2] Q. Bi, H. Yan, C. Chen, Q. Su, “An Integrated Machine Learning Framework for Stock Price Prediction”, CCIR 2020, LNCS 12285, pp. 99–110, 2020.

[3] M. Vicari, M. Gaspari, “Analysis of news sentiments using natural language processing and deep learning”, AI & SOCIETY (2021) 36:931–937, https://doi.org/10.1007/s00146-020-01111-x

[4] D. Shah, H. Isah, F. Zulkernine, “Stock Market Analysis: A Review and Taxonomy of Prediction Techniques”, Int. J. Financial Stud. 2019, 7, 26; doi:10.3390/ijfs7020026

[5] K. K. Nivethithaa, Dr. S. Vijayalakshmi, “Survey on Data Mining Techniques, Process and Algorithms”, Journal of Physics: Conference Series 1947 (2021) 012052, IOP Publishing, doi:10.1088/1742-6596/1947/1/012052

[6] https://www.softwaretestinghelp.com/data-mining-process/

[7] J. M. David, K. Balakrishnan, “Significance of Classification Techniques in Prediction of Learning Disabilities”, https://arxiv.org/ftp/arxiv/papers/1011/1011.0628.pdf

[8] F. Alam, S. Pachauri, “Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA”, Advances in Computational Sciences and Technology, ISSN 0973-6107 Volume 10, Number 6 (2017) pp. 1731-1743, © Research India Publications

[9] S. L. Pandhripande and Aasheesh Dixit, “Prediction of 2 Scrip Listed in NSE using Artificial Neural Network”, International Journal of Computer Applications (IJCA), Volume 134, No.2, January 2016.

[10] Dr. B. Srinivasan, K. Pavya, “A Study on Data Mining Prediction Techniques in Healthcare Sector”, International Research Journal of Engineering and

Technology (IRJET), PP. 552-556, Volume 3, Mar-2016

[11] Vipin Kumar, Joydeep Ghosh and • David J. Hand, "Top 10 algorithms in data mining", Knowledge and Information System, PP. 1–37, (2008).

[12] Vapnik V (1995), the nature of statistical learning theory. Springer, New York.

[13] G. Kavitha, A. Udhaya Kumar, D. Nagarajan, "Stock Market Trend Analysis Using Hidden Markov Models", available online: https://arxiv.org/ftp/arxiv/papers/1311/1311.4771.pdf.

[14] H. Yang, L. Chan, I. King, "Support Vector Machine Regression for Volatile Stock Market Prediction", Intelligent Data Engineering and Automated Learning IDEAL, PP. 391- 396, Springer-Verlag Berlin Heidelberg 2002

[15] Investopedia US, "A Division of Value Click", Inc., "Investopedia" "http://www.investopedia.com /articles/05/032905.asp",05 March 2013.

[16] Mrs. K. S. Mahajan, R. V. Kulkarni, "A Review: Application of Data Mining Tools for Stock Market", International Journal Computer Technology & Applications, Volume 4, PP. 19-27, 2013.

[17] S. Simon, A. Raoot, "Accuracy Driven Artificial Neural Networks in Stock Market Prediction", International Journal on Soft Computing (IJSC), Vol.3, No.2, May 2012.

[18] D. V. Setty, T. M. Rangaswamy, K. N. Subramanya, "A Review on Data Mining Applications to the Performance of Stock Marketing", international Journal of Computer Applications (IJCA), Volume 1, No. 3, PP. 24-34, 2010.

[19] S. Tulankar , Dr R. Athale, S. Bhujbal, "Sentiment Analysis of Equities using Data Mining Techniques and Visualizing the Trends", IJCSI International Journal of Computer Science Issues, Vol. 10, No 2, July 2013.

[20] M. Vijh, D. Chandola, V. A. Tikkiwal, A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques", Procedia Computer Science 167 (2020) 599–606

[21] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis", VOLUME 8, 2020

[22] A. Moghar, M. Hamiche, "Stock Market Prediction Using LSTM Recurrent Neural Network", Procedia Computer Science 170 (2020) 1168–1173

[23] I. K. Nti, A. F. Adekoya, B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock‐market prediction", J Big Data (2020) 7:20, https://doi.org/10.1186/s40537-020-00299-5

[24] J. Liu, H. Lin, L. Yang, B. Xu, D. Wen, "Multi-Element Hierarchical Attention Capsule Network for Stock Prediction", IEEE Access, VOLUME 8, 2020

[25] M. McCoy, S. Rahimi, "Prediction of Highly Volatile Cryptocurrency Prices Using Social Media", International Journal of Computational Intelligence and Applications Vol. 19, No. 4 (2020) 2050025 (28 pages)

[26] S. Bouktif, A. Fiaz, M. Awad, "Augmented Textual Features-Based Stock Market Prediction", IEEE Access, VOLUME 8, 2020

[27] F. G. D. C. Ferreira, A. H. Gandomi, R. T. N. Cardoso, "Artificial Intelligence Applied to Stock Market Trading: A Review", IEEE access, VOLUME 9, 2021

[28] I. K. Nti, A. F. Adekoya, B. A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions", Artifcial Intelligence Review, https://doi.org/10.1007/s10462-019-09754-z

[29] D. V. Cruz, V. F. Cortez, A. L. Chau, R. S. Almazán, "Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID‐19 Periods", Cognitive Computation, https://doi.org/10.1007/s12559-021-09819-8