International Journal of Statistics and Applied Mathematics

**ISSN:** 2456-1452
Maths 2024; SP-9(1): 90-93

https://www.mathsjournal.com
Received: 15-10-2023
Accepted: 28-11-2023

**Pal Deka**
Assistant Professor, Department of Agricultural Statistics, Biswanath College of Agriculture, Biswanath Chariali, Assam, India

**Manash Pratim Barman**
Associate Professor, Department of Statistics, Dibrugarh University, Dibrugarh, Assam, India

**Corresponding Author:**
**Pal Deka**
Assistant Professor, Department of Agricultural Statistics, Biswanath College of Agriculture, Biswanath Chariali, Assam, India

# Twitter sentiment analysis of national stock exchange, India

**Pal Deka and Manash Pratim Barman**

**Abstract**
Sentiment analysis is the term used to understand the sentiments of the people or individuals. The sentiments of the investors towards certain stocks and stock exchanges could be classified into positive, negative and neutral. Twitter is one such social networking platforms used to express such sentiments. Here in this study, perceptions or sentiments of the investors towards one of the India's largest stock exchange is examined. The tweets regarding #National Stock Exchange, India were extracted from the twitter platform from 2016-2021. The results shows that the neutral sentiment tweets were highest in number than positive and negative tweets during this period of time. Further the study also performed sentiment group classification. Multinomial logistic regression, Naïve Bayes and Support vector machine classification algorithms was considered. Among these learning algorithms, accuracy score of support vector machine algorithm was found better in classification of the tweets. The F-1 score of support vector machine algorithm was satisfactory then rest of the algorithms.

**Keywords:** National stock exchange, twitter, sentiment analysis, machine learning algorithms, tweets

## 1. Introduction
The future prediction of stock market is always an area of research for the researchers and investors [2]. In recent times, machine learning and deep learning has fruitful applications in stock market. Machine learning algorithms were applied in both classification and prediction of the stock market. The dynamic nature of stock market has been also explained by some deep learning methods. These algorithms in some studies have shown better results than classical statistical methods. The perception or sentiments of the investors towards a company stock or stock market is an important aspect in stock market analysis. In the past, investors' sentiments or views about certain stocks took sufficient amount of time to reach among other investors [13]. Nowadays individual expressed their inner sentiments about anything in internet(web) such as wiki, blogs, forum and social media platforms like facebook, twitter etc. [3]. Among all the other social networking platforms, twitter could be taken into consideration. Twitter has been considered as the 10$^{th}$ most widespread social network platform around the world with 300 million active users per month [8]. The information shared by users in twitter could be used by others for investment in a particular stock [7]. Sentiment analysis is also known as opinion mining which address opinion oriented natural language processing [11]. Sentiment analysis is used in research studies to know about the review of peoples towards a particular commodity in the market, new movie, preference of political party etc. This study of sentiments of people could be broadly considered under sentiment analysis. Sentiment analysis is the term used to understand the sentiments of the people or individuals. It is considered under Natural language processing (NLP). The sentiment analysis has overcome the subjective analysis in observing the views as positive, negative or neutral but not as subjective and objective. There are several research studies regarding application of sentiment analysis in stock market. Sentiment analysis plays important role in prediction of movement of NIFTY and SENSEX [1]. In this 21$^{st}$ century the technology has provided everyone a platform to express their sentiments in a less than a fingertip. This technological advancement generated huge amount of data which also contributed for the change in economic and business strategies [12]. The present research study consists of understanding the sentiments of twitter users regarding National Stock exchange. The sentiment or view point regarding this stock exchange

is an important aspect of this study. The behavior of the investors towards the companies listed under National stock exchange could be also known with the help of this study. The study has used three machine learning algorithms to classify and predict the filtered tweets. The accuracy of these machine learning algorithms depends on the confusion matrix and classification report generated by each algorithm. The next section of the paper consists of data and methodology adopted for preprocessing of the tweets and later section consists of results and conclusion.

## 2. Material and Methods

### 2.1 Data

The study consists of unstructured data i.e. twitter feeds data from anonymous users with tweets regarding #National Stock Exchange. Primarily, twitter developer account was created on developer.twitter.com [17] and its application programming interface (API) helped to extract the tweets. It was collected from the year 2016 to 2021. A total of 13377 tweets were gathered in this particular time frame. The feelings and opinions expressed through tweets during the pandemic year 2020 were also included for the study. Such sentiments helps to understand the post covid-19 affect on companies [6]. A single tweet consists of maximum 280 characters and it generates 70 billion characters per day on twitter [14]. The tweets were preprocessed through removal of stop words, special characters, URLs etc. The re-tweets and duplicate tweets were also dropped from the study.

### 2.2 Methodology

The analysis of the study is performed using python programming language. Sentiment classification is also one of the important aspects of sentiment analysis for classifying text into polarity group (Medeiros and Borges, 2019). The algorithms employed for sentiment group classification are Support vector machine, Multinomial logistic regression and Naïve Bayes. The preprocessed tweets were divided into seventy percent training and thirty percent testing sets. The algorithms considered ten words at a time to classify and later predict the tweets into respective sentiment groups. The conversion of tweets into bag of words is done through the method of n-gram of natural language processing. The following algorithms used in the study are discussed below:

#### 2.2.1 Multinomial logistic regression

Logistic regression algorithm could be considered as one of the important supervised machine learning algorithms. Apart from its application in regression, it is widely used in text mining, binary as well as multinomial classification problems. In multinomial logistic regression, the dependent variable consists of three classes/ categories. The independent variables of the model should not have multi-co linearity among them. It is used to predict the response variable on the basis of independent variables (continuous or categorical) [4]. In logistic regression, the curve shaped is S shape in contrast to straight linear regression line.

#### 2.2.2 Support Vector Machine

Support Vector Machine (SVM) is an important and widely used machine learning algorithms. It has been used for classification and regression related problems. Some recent research studies shows that the Support vector machine gives better classification accuracy results than other classification algorithms [15]. The algorithm creates a hyperplane which will divide the n-dimensional space into classes. The hyperplane is develop through support vectors or points. The algorithm scope is broad due to its applications in most of the research areas.

#### 2.2.3 Naïve Bayes Classifier

Naïve Baye's algorithm is based on Bayes's theorem. It follows the independence assumption of features related to one class is independent to the other. This algorithm is also used for both classification and regression machine learning related problems. It is used in sentiment analysis, text classification, spam detection etc. It can be used for both linearly and non-linearly separable cases [9]. The algorithm is considered to be simple and provide better accuracy then most of the other complex algorithms. It deals with real world applications and independent of noise.

### 2.3 Evaluation of the fitted models

In order to evaluate the performance of the three selected machine learning models, the algorithms were trained with 70 percent of the tweets and rest 30 percent is used for testing of these models. Following are the measures used for evaluation of the fitted models on the basis of test samples.

**Accuracy Score**

The score refers to the percentage of correctly classified test samples. The score is calculated with the help of formula developed by (Harikrishnan, 2019).

$$\text{Accuracy Score} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

TP= True Positive, TN= True Negative
FP= False Positive, FN= False Negative

**F-1 Score**

F-1 score is considered as an alternative evaluation metrics of machine learning algorithms. In contrast to accuracy score, it concentrates on predictive skill of group/class wise performance of the algorithms. The score uses two important evaluation metrics precision and recall. It is the harmonic mean of the precision and recall.

$$\text{F-1 score} = \frac{2}{\frac{1}{\text{precison}} + \frac{1}{\text{recall}}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

**Confusion matrix**

Confusion matrix is a performance measure for machine learning classifiers. The matrix could be constructed for two or more than two classes. It consists of true and predicted values in rows and columns.

## 3. Results and Discussions

The filtered tweets are classified based on polarity and subjectivity score into three sentiments i.e. Positive, Negative and Neutral. The individual tweets which score was less than 0 were considered as negative and for the tweets scored greater than 0 were positive. The rest of the tweets were in neutral group. For training the machine learning algorithms, 70 percent randomly selected tweets is used and rest 30 percent is for testing purpose. The following table 1 shows the number of tweets based on positive, negative and neutral.

**Table 1:** Classification of Tweets based on different sentiments

| Sentiments | Number of tweets |
|---|---|
| Positive | 3750 |
| Negative | 1802 |
| Neutral | 7825 |

From table 1, the neutral tweets are more in number in comparison to positive and negative tweets. The negative tweets are least in number among all three tweet groups.

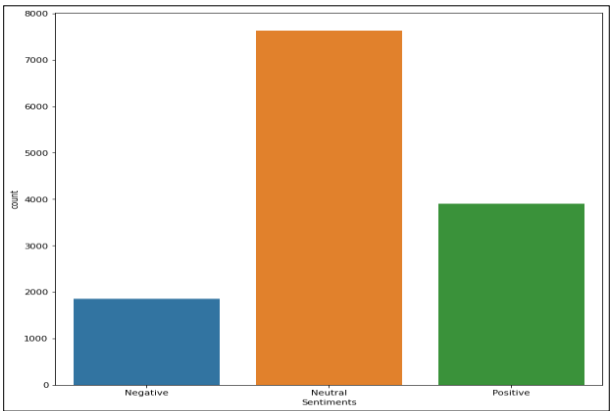


**Fig 1:** Graphical representation of sentiments of tweets

From the bars of sentiment group it is observed that neutral tweets are highest in number then positive and negative tweets.

### 3.1 Evaluation of Supervised machine learning algorithms

The evaluation of the supervised algorithms is done through the calculation of accuracy score. The table 2 below shows the percentage of accuracy of the algorithms.

**Table 2:** The accuracy score for the supervised machine learning algorithms

| Supervised learning algorithms | Accuracy Score | Percentage |
|---|---|---|
| Support Vector Machine | 0.8244 | 82.44% |
| Multinomial Logistic Regression | 0.8106 | 81.06% |
| Naïve Bayes Classifier | 0.7968 | 79.68% |

The accuracy score of Support vector machine algorithm is better in comparison to Multinomial Logistic regression and Naïve Bayes Classifier. The accuracy percentages of three algorithms have little difference among them.

### 3.2 Confusion Matrices

The confusion matrices of the supervised learning algorithms are given as follows:

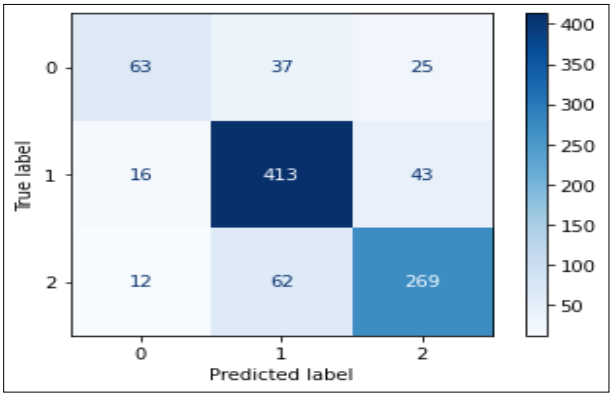


**Fig 2:** Confusion matrix of Support Vector Machine Algorithm

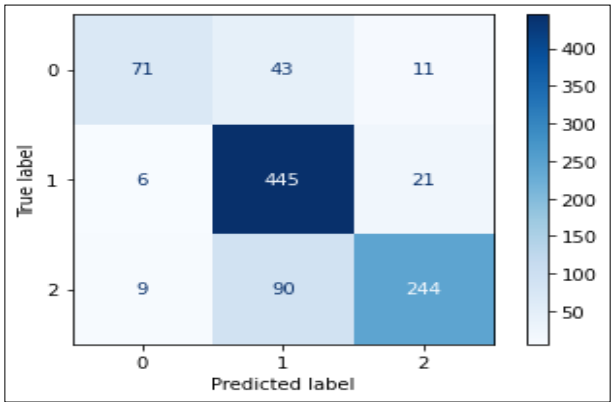


**Fig 3:** Confusion matrix of Multinomial Logistic Regression

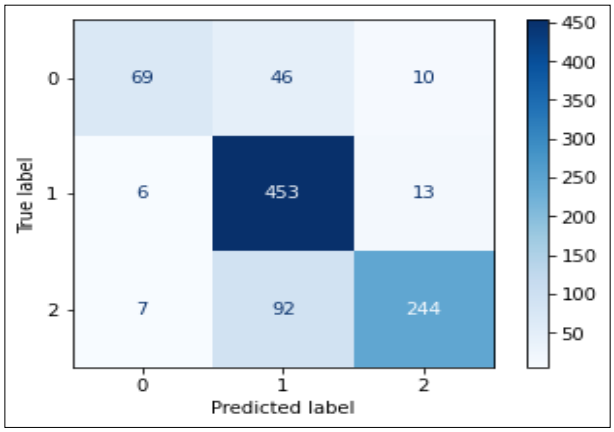


**Fig 4:** Confusion matrix of Naïve Bayes Classifier

The confusion matrix of the selected machine learning algorithms consists of three classes (0, 1 and 2) or Negative, neutral and positive sentiment tweets. For prediction of negative tweets, multinomial logistic regression shows better than rest of the algorithms. The Naïve Bayes classifier shows good prediction for neutral tweets in comparison to other two models. Further from the matrices, it is observed that support vector machine has shown good prediction skill for positive tweets.

### 3.3 Classification Reports

Classification report of supervised learning algorithms is given below

#### 3.1.1 Support Vector Machine

| Labels | Precision | Recall | F1-score |
|---|---|---|---|
| Positive | 0.94 | 0.80 | 0.86 |
| Negative | 0.92 | 0.76 | 0.83 |
| Neutral | 0.88 | 0.98 | 0.93 |

#### 3.1.2 Multinomial Naïve Bayes

| Labels | Precision | Recall | F1-score |
|---|---|---|---|
| Positive | 0.78 | 0.78 | 0.78 |
| Negative | 0.89 | 0.68 | 0.77 |
| Neutral | 0.87 | 0.92 | 0.89 |

#### 3.1.3 Multinomial Logistic regression

| Labels | Precision | Recall | F1-score |
|---|---|---|---|
| Positive | 0.93 | 0.78 | 0.85 |
| Negative | 0.93 | 0.73 | 0.82 |
| Neutral | 0.87 | 0.98 | 0.92 |

The classification report of the machine learning algorithms is more or less similar. The F1-score for each sentiment group of support vector machine is better in comparison to other respective algorithms. The precision and recall value of the algorithms is satisfactory to some extent.

## 4. Conclusion

The study was undertaken to observe the sentiments of investors towards National Stock exchange, India. Among several other social networking platforms, twitter was selected for collection of tweets regarding one of the India's important stock exchanges for last five years. It has revealed that neutral sentiment tweets were highest than positive and negative tweets during this period of time. For sentiment group classification, Multinomial logistic regression, Naïve Bayes and Support vector machine was considered. The accuracy score of support vector machine algorithm is better followed by other two selected learning algorithms. The F-1 score for the three sentiment groups of support vector machine algorithm was satisfactory then rest of the algorithms.

## 5. Declarations

### Availability of data materials

The datasets generated and/or analyzed during the current study are available in twitter, https.developer.twitter.com

### Competing interest

The authors declare that they have no competing interests

### Funding

No funding agencies were involved in this research study

### Author's contribution

The idea behind this research was of MPB. PD collected, analyzed and interpreted the data regarding twitter sentiment analysis. The programming was also done by PD and was a major contributor in writing the manuscript. The final manuscript was read and approved by the authors.

## 6. Acknowledgements

I on behalf of all the authors associated with this research study acknowledge Twitter social network platform for providing the data.

## 7. Reference

1. Bharadwaj A, Narayan Y, Vanraj P, Pawan, Dutta M. Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. In: 4th International Conference on Eco-friendly Computing and Communication Systems; 2015. Elsevier. Procedia Computer Science. 2015;70:85-91.
2. Bharathi S, Geetha A. Sentiment Analysis for Effective Stock Market Prediction. International Journal of Intelligent Engineering and Systems. 2017;10(3):146-154. DOI: 10.22266/ijies2017.0630.16.
3. Chavan V, Dr. Sannaki SS, Sambrekar KP. Share Market Prediction using Twitter Sentiment Analysis. International Journal of Innovative Research in Technology. 2021;8(5):540-548. ISSN: 2349-6002.
4. El-Habil AM. An Application on Multinomial Logistic Regression Model. Pakistan Journal of Statistics and Operation Research. 2012;8(2):271-291.
5. Harikrishnan NB. Confusion Matrix, Accuracy, Precision, Recall, F1 score. Medium. 2019. Available from: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-scoreade299cf63cd.
6. Jain Y, Tirth V. Sentiment Analysis of Tweets and Texts Using Python on Stocks and COVID-19. International Journal of Computational Intelligence Research. 2020;16(2):87-104. DOI: https://dx.doi.org/10.37622/IJCIR/16.2.2020.87-104.
7. Kedar SV, Kadam S. Stock Market Increase and Decrease using Twitter Sentiment Analysis and ARIMA Model. Turkish Journal of Computer and Mathematics Education. 2021;12(1S):146-161.
8. Kumar SK, Aolo A, *et al*. Stock Price Prediction Using Optimal Network Based Twitter Sentiment Analysis. Intelligent Automation and Soft Computing. 2022;33(2):1217-1227. Tech Science Press. DOI: 10.32604/iasc.2022.024311.
9. Mathapati PM, Shahapukar AS, Hanabaratti KD. Sentiment Analysis using Naïve-Bayes Algorithm. International Journal of Computer Science and Engineering. 2017;5(7):75-77. E-ISSN: 2347-2693.
10. Medeiros CM, Borges PRV. Twitter Sentiment Analysis Regarding the Brazilian Stock Market. In: VIII Brazilian Workshop on Social Network Analysis and Mining; 2019. DOI: 10.5753/brasnam.2019.6550.
11. Ozturk SS. A Sentiment Analysis of Twitter Content as a Predictor of Exchange Rate Movements. Review of Economic Analysis. 2014;6:132-140.
12. Ranco G, Aleksovski D, Caldarelli G, Grcar M, Mozetic I. The Effects of Twitter Sentiment on Stock Price Returns. PLOS ONE. 2015;10(9):e0138441. DOI: 10.1371/journal.pone.0138441.
13. Rao T, Srivastava S. Analyzing Stock Market Movements Using Twitter Sentiment Analysis. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining; 2012. IEEE Computer Society. DOI: 10.1109/ASONAM2012.30.
14. Shah N. Stock Market Movements Using Twitter Sentiment Analysis. International Journal of Scientific Development and Research. 2019;4(4):105-109. ISSN: 2455-2631.
15. Srivastava DK, Bhambhu L. Data Classification using Support Vector Machine. Journal of Theoretical and Applied Information Technology. 2010;1-7.
16. Tabari N, Praneeth B, Biswas P, *et al*. Causality Analysis of Twitter Sentiments and Stock Market Returns. In: Proceedings of the First Workshop on Economics and Natural Language Processing; 2018. Melbourne, Australia: Association for Computational Linguistics; 11-19.
17. Twitter Developer. Available from: https://developer.twitter.com.