Analysis of user emotions and opinion using Multinomial Naive Bayes Classifier

Surya Prabha PM
Department of Computer Science and Engineering
Thiagarajar College of Engineering
Madurai, India
suryaprabhab.e3796@gmail.com

Seetha Lakshmi V
Department of ComputerScience and Engineering
Thiagarajar College of Engineering
Madurai, India
seethavenkatesan1997@gmail.com

Subbulakshmi B
Department of ComputerScience and Engineering
Thiagarajar College of Engineering
Madurai, India
bscse@tce.edu

Abstract*— The user emotion analysis has been a great deal, Here primarily, the twitter dataset is used for analysis. [6] In the current trend, twitter application usage is getting drastically increased where people share their thoughts with each other throughout the world. Generally people use social media as a platform to express their opinion to others, In the meantime Facebook and Twitter serve as the major source for them. Hence they make use of this when needed. The volume of the tweets remain quite high since it is a collection of so many tweets throughout the world, and also it consists of both positive and negative tweets, For the purpose of analysis it is necessary to find the positive and negative tweets and categorize them accordingly. In this project the input dataset consists of a compilation of a number of tweets which expresses the thoughts of various people regarding the election results and their own thoughts about different election candidates who compete in the election, correspondingly each person may have varying opinion on each candidate which will be positive or negative. The analysis of these tweets will help in recording a small survey over those different candidates. The dataset consists of more than 2000 records. The analysis is being done with the help of python programming by importing various packages and implementing a number of functions required for analysis. The results are obtained in two different categories namely positive and negative. Multinomial Naive Bayes classifier has been used for the purpose of analysis. The work is further extended and tested for the imdb movie review dataset, mobile review dataset and the results are categorized as positive and negative.

Keywords—Multinomial Naive Bayes, user tweets, confusion matrix, sentimental analysis, positive, negative.

I. INTRODUCTION

Sentimental Analysis is a broader field in text mining which has a great role in text classification.[1],[3],[4], Twitter Analysis is one of the subareas in Sentimental analysis where the tweets are being classified into different categories, Twitter serves as a source for the society in gathering the people's thoughts, Often twitter contributes very high in marketing.[7], People share their thoughts regarding various products in the market, which may include the quality of the product, the most current trending product in the market, which marks out a varying graph of the thoughts of different people. [10], Not only in marketing, twitter also has a major effect in many fields, in many cases the tweets of people change the situation. Analysis over the tweets is necessary since it gives an overall opinion in many cases and gives a clear cut idea of what various people think in a situation. For the analysis purpose a proper algorithm is required in order to provide accurate results.

A. *Machine Learning Algorithms*

[10],[11],[12] Machine learning algorithms serve as a great source in text classification, For good efficiency and better understanding probabilistic algorithms are adopted for analysis. In such way Naïve Bayes, SVM, C4.5 algorithms serve better in classification. There are also many other classification techniques available.

B. *Platforms available for classification*

There are two different platforms for classification, namely i) Tool based classification ii) Code level classification.

C. *Tool based Classification*

There are various tools available for classification, such as R, Weka, Rapid Miner, XL Miner, Knime, Orange etc. Some of the tools require just drag and drop of functions, whereas tools like R require importing and loading of different packages required for the particular algorithm to work.

D. *Code Level Classification*

There are different programming languages available for analysis and classification of the text namely Java, Python, etc.

E. *Data sources for Classification*

There are many different kinds of dataset available for the purpose of sentiment classification. The source for the datasets is not limited, there are many websites available which provide the dataset for sentimental analysis.

There also many open source dataset repositories available, which provides the necessary dataset required. The authors could also make a request to other authors who acquire the corresponding dataset and avail it.

II. EXPLANATION OF METHODOLOGY

The methodology used here for implementation is Multinomial Naïve Bayes Classifier (MNB), which is a modified form of Naïve Bayes Classifier, MNB is also a probabilistic approach which is similar to Naïve Bayes. MNB is specially designed for text documents in order to calculate the occurrence of each words.

A. Difference between Multinomial Naïve Bayes and Naïve Bayes classifiers

Generally [5], [2], Naïve Bayes (NB), works based on the conditional probability (It considers the conditional independence of the features), while the Multinomial Naïve Bayes works based on the multinomial distribution. The multinomial Naïve Bayes classifier considers the multiple occurrences of each term.

B. Why Multinomial Naïve Bayes?

Naïve Bayes classifier is one of the most widely used method for text mining, Multinomial Naïve Bayes could be said as a upgraded version of the existing Naïve Bayes classifier, and it effectively manipulates the word count by calculating the frequency of each word, whereas in Naïve Bayes classifier the frequency of the words does not have much effect on the working of the algorithm. It is known that the frequency of each text has a higher impact in categorizing the text into different categories. Hence Multinomial Naïve Bayes is considered to be best for the purpose of classification of the text.

III. PROCESS FLOW

The classification process includes a number of sequential steps. They are,

1) Step 1: Dataset Gathering

The dataset required for the project is being gathered.

a) Pre-Processing

The collected dataset is preprocessed, and the unwanted symbols, values are removed so that the dataset is completely fit for the classification process.

2) Step 2: Implementation

The implementation is done by loading different packages required for the algorithm. The dataset is loaded as a csv file and then the steps such as feature extraction, finding the term frequency for each corresponding terms is being done and then the algorithm is applied on the processed testing dataset leading to the expected result.

3) Step 3: Gathering of Results

The results are being obtained as two categories as positive and negative which denotes if the corresponding tweet is positive or negative.

4) Step 4: Performance Comparison

The existing work of Naïve Bayes is being compared with the current proposed method with the help of confusion matrix.

The Work flow of the proposed method is being visualized in the "Fig.1."

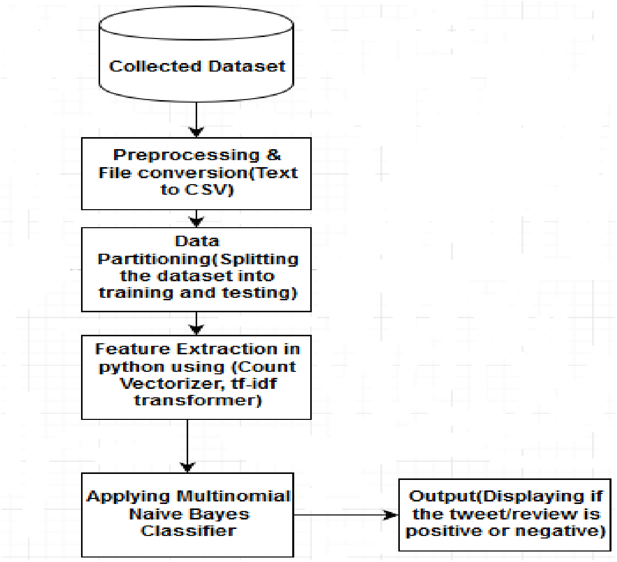

Collected Dataset
Preprocessing & File conversion(Text to CSV)
Data Partitioning(Splitting the dataset into training and testing)
Feature Extraction in python using (Count Vectorizer, tf-idf transformer)
Applying Multinomial Naive Bayes Classifier
Output(Displaying if the tweet/review is positive or negative)

Fig.1. Work Flow of the proposed Method

IV. PSEUDOCODE FOR MULTINOMIAL NAÏVE BAYES

```
1.  Read the input Dataset T.
2.  Partition the dataset as training t1 and
    testing t2 where (t1,t2 )∈T.
3.  Function Multinomial Naïve Bayes (MNB)
4.  Consider t1 as labelled set.
5.  Consider t2 as data fragments (DF).
6.  Calculate the term frequency (TF)
    for each DF.
    TF (T) = (Nt (T) in D) / (T(Nt(t))
    Where,
    T→Corresponding term in a document
    Nt →Number of occurrences of a term in the
          document
    D→The Corresponding document
     t→The Corresponding term
7.  Calculate the inverse document frequency
    (IDF) for each DF.
    IDF(T)= log e(tn (D)/ N(D) with term t in it.
    tn  (D)→Total number of documents.
    N(D)→Number of documents
8.  Calculate TF-IDF
    TF-IDF= TF * IDF
9.  Find the emotion of the text using
    MNB.
10. Display the result ($P_C$ , $N_C$)
    Where,
    $P_C$→Positive category
    $N_C$→Negative category
11. End function
```

V. DETAILS OF THE DATASET

For the preliminary analysis, the data set used in this project is collected from dataworld.io. Two types of dataset is required for the process (training and testing).

A. File conversion

The dataset will be downloaded in text file format, hence a file conversion is required where the file is being converted as CSV format (Comma Separated Value). The file after conversion could be viewed in Microsoft Excel. The file conversion here is much required for the processing for better understanding of the dataset, and for making the process easier. Since the file in text format will appear as clumsy form and will not be that easy for processing.

B. Training Dataset

The training Dataset is a csv file which consists of a collection of number of words, and their corresponding label which is being considered while processing of the words in the document. The file consists of positive, negative columns which corresponds to 0 or 1, for each word depending upon the word.

For example, i) Bad- For this word the negative category will be 1 and the positive category will correspond to 0. ii) Good- For this word the positive category will be 1, and the negative category will correspond to 0.

Similarly the dataset consists of thousands of words and for each word the corresponding label (0 or 1) of the respective category (positive or negative) is being mentioned. This labelling helps in the processing of the testing dataset using the proposed algorithm.

C. Testing Dataset

The testing dataset consists of the tweets of various users, The dataset is a compilation of about more than 2000 records. The algorithm is applied in the testing dataset. The implementation of the proposed algorithm will be carried out in the testing dataset only. The process is being done with the help of the collection of the labelled words in the training dataset.

VI. SOFTWARE DESCRIPTION

The IDE used here for implementation of python coding is Spyder3.3. This IDE requires a minimum of 550MB RAM availability, 1 GB of free disk space (For better performance), the system may be of 32 bit version or 64 bit version. The IDE is available for both windows and linux operating systems.

A. Installation

This IDE is installed through Anaconda Distribution, which is a open source python distribution. The anaconda navigator comes along with the anaconda package and could be installed through anaconda prompt. The Anaconda navigator provides a list of various IDE available for python, Spyder IDE is one of the widely used IDE's among other packages. Spyder 3.3.3 is the recently released version and it is used for the implementation of this project. This software could be downloaded from the official website of spyder (https://www.spyder-ide.org).

VII. WORKING OF THE CODE

The algorithm is designed with the help of python code

A. Packages Used

1) Scikit learn:

The Multinomial Naïve Bayes is being imported using scikit learn package in the python code.

2) Pandas:

Pandas is one of the packages in python which is used for analysis of unlabeled data. This is imported as pd.

3) Numpy:

Numpy package is used for processing multidimensional arrays, and the data could be loaded as data frames in the code using this package.

B. Functions used

The count vectorizer and tf-idf transformer are used for processing of the document.

1) Count Vectorizer

The code consists of Count Vectorizer which is used for tokenization.

2) Tf-idf transformer

a) Term frequency (Tf):

The term frequency refers to the number of the occurrences of different words, which is generally calculated based on how often the term occurs in a document. It doesn't ignore the most commonly occurring words. Hence the issue arises while weighting.

b) Inverse Document Frequency (Idf):

The term Inverse Document frequency is quite different from the term frequency. i.e., it does not considers the most commonly occurring words such as I, he, she, is, the, as, etc. It only takes into account of the rarely occurring words in the document based on the type of document and relevance of the words. In short it only considers the more meaningful words occurring in the document which is required for processing.

C. Considering the dataset as fragments(DF)

Each dataset is considered as a fragment for processing. The data is depicted as columns and there are two variables considered here namely x, y. Two categories are present namely cls_neg, cls_pos.

D. Applying MNB

The MNB algorithm is being applied for classification. MNB is applied to the DF, and the category is found.

VIII. RESULTS AND ANALYSIS

A. Analysis using multiple datasets

For the analysis of the Multinomial Naïve Bayes Classifier multiple datasets are used here, namely the twitter dataset and imdb movie review dataset, mobile review dataset in which the results display the positive and negative categories.

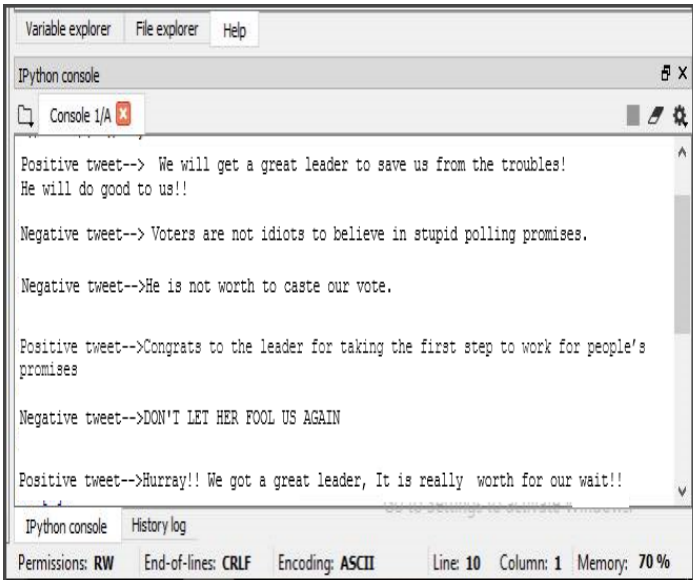
1) Twitter Dataset

The primary analysis of the proposed method is being done with the twitter dataset, the results consists of the tweets representing the positive and negative category of each tweet. The classification is done based on the analysis of every single word, each word and their label will correspond to the

analysis, the sum of the labels of each word in a sentence will help in the prediction of the category of the tweet.
The results will be displayed in the console.

The output for the proposed work is being depicted below in the "Fig.2."

"Fig.3." displays the categorisation for the movie review dataset.



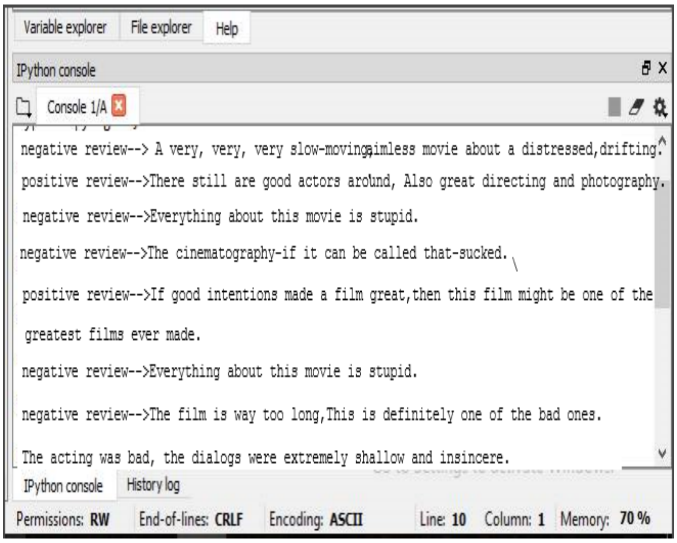
```
Positive tweet-->  We will get a great leader to save us from the troubles!
He will do good to us!!

Negative tweet--> Voters are not idiots to believe in stupid polling promises.


Negative tweet-->He is not worth to caste our vote.


Positive tweet-->Congrats to the leader for taking the first step to work for people's
promises

Negative tweet-->DON'T LET HER FOOL US AGAIN


Positive tweet-->Hurray!! We got a great leader, It is really  worth for our wait!!
```

Permissions: RW End-of-lines: CRLF Encoding: ASCII Line: 10 Column: 1 Memory: 70 %

Fig.2. Screenshot displaying the category of tweets

2) Movie Review Dataset

The analysis is further extended using the imdb movie review dataset, for which the result effectively visualizes the positive and negative reviews for each corresponding review of the user.



```
negative review--> A very, very, very slow-moving,aimless movie about a distressed,drifting.
positive review-->There still are good actors around, Also great directing and photography.
negative review-->Everything about this movie is stupid.
negative review-->The cinematography-if it can be called that-sucked.
positive review-->If good intentions made a film great,then this film might be one of the
 greatest films ever made.
negative review-->Everything about this movie is stupid.
negative review-->The film is way too long,This is definitely one of the bad ones.
The acting was bad, the dialogs were extremely shallow and insincere.
```

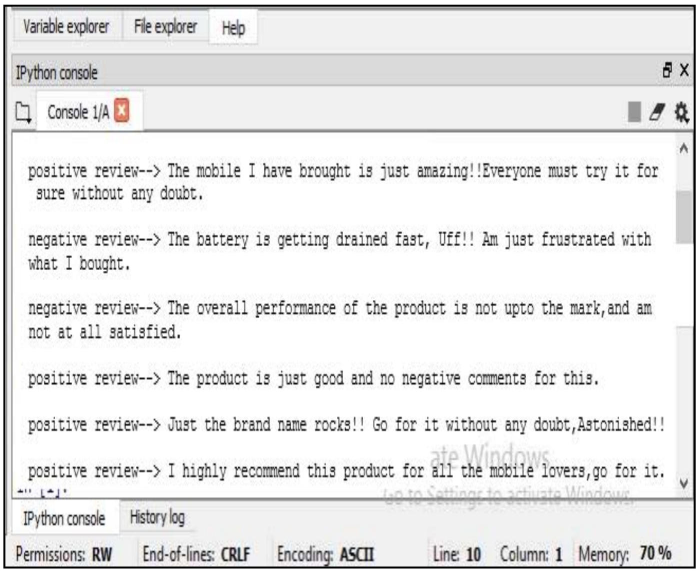
Permissions: RW End-of-lines: CRLF Encoding: ASCII Line: 10 Column: 1 Memory: 70 %

Fig.3. Screenshot displaying the category of movie reviews

3) Mobile Reviews Dataset

The mobile reviews dataset which gathers the reviews of the various users is also used for the analysis purpose and the results are obtained accordingly.

"Fig.4." displays the categories of the mobile reviews.



```
positive review--> The mobile I have brought is just amazing!!Everyone must try it for
 sure without any doubt.

negative review--> The battery is getting drained fast, Uff!! Am just frustrated with
what I bought.

negative review--> The overall performance of the product is not upto the mark,and am
not at all satisfied.

positive review--> The product is just good and no negative comments for this.

positive review--> Just the brand name rocks!! Go for it without any doubt,Astonished!!

positive review--> I highly recommend this product for all the mobile lovers,go for it.
```

Permissions: RW End-of-lines: CRLF Encoding: ASCII Line: 10 Column: 1 Memory: 70 %

Fig.4. Screenshot displaying the category of mobile reviews

IX. PERFORMANCE COMPARISON

For the purpose of justification of performance of the current proposed approach the work extension has been done by comparing it with the performance of the naïve Bayes classifier.

A. Naïve Bayes classifier

The implementation for Naïve Bayes classifier has been done with the help of amazon product reviews dataset which consisted of about 1000 observations and the result visualized the two categories of the dataset namely positive and negative category in the form of a confusion matrix which consisted of the true positive, false positive, true negative and false negative values, where the true negative and false negative values indicate the misclassified text in the dataset.

Correspondingly, the accuracy of this method was calculated which was about 78%.

B. Multinomial Naïve Bayes

The currently proposed method is a modified form of Naïve Bayes classifier, which aims to provide higher accuracy than the previous method, it is proved by comparing the misclassification rate of both the methods, where the Multinomial Naïve Bayes has lesser misclassified values than the previously proposed method, and also the accuracy of the proposed method was found out to be 89%. Hence it is literally understood that the currently proposed method provides better results than the Phase-I method.

C. Work Comparison with same input set

For the better comparison results both the methods has been checked with the same input set (twitter dataset) which consisted of 1000 records, and the correctly classified values for Naïve Bayes corresponds to 400 positive values, 380 negative values, and the remaining values are the misclassified values.

Similarly for Multinomial Naïve Bayes the correctly identified values was about 420 positive values, 450 negative values, while the remaining tweets were found out to be misclassified. While comparing the performance of both the algorithms, it is clearly found that Multinomial Naïve Bayes performed well than the Naïve Bayes classifier.

"Table.1","Table.2" effectively visualizes the correctly predicted and misclassified values obtained after processing.

Table.1.Classification Results of NB

Category	Positive	Negative
Positive Value	400	165
Negative Value	50	380

Table.2.Classification Results of MNB

Category	Positive	Negative
Positive Value	420	75
Negative Value	50	450

E. Accuracy Prediction

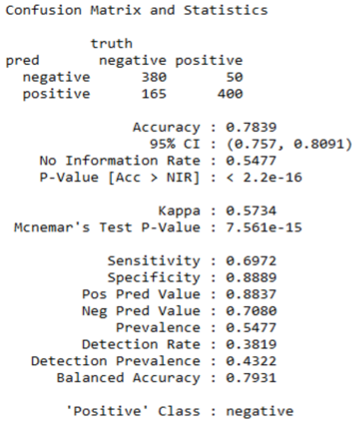
1) Confusion Matrix Explanation

In "Fig.5.","Fig.6.", For the Purpose of accuracy prediction confusion matrix is being obtained, Furtherly the Specificity(True Negative(FP) rate)(which refers to the correctly classified negative values), Sensitivity(True Positive((TP) or Recall rate),(which refers to the correctly classified positive values) values are being calculated and being displayed. The Kappa value defines the relation between the two different values (TP & FP). The process is being done with the help of R tool. The accuracy obtained after processing both the classifiers using same dataset is correspondingly 78% for Naïve Bayes Classifier and 87% for Multinomial Naïve Bayes Classifier.

2) Accuracy Depiction

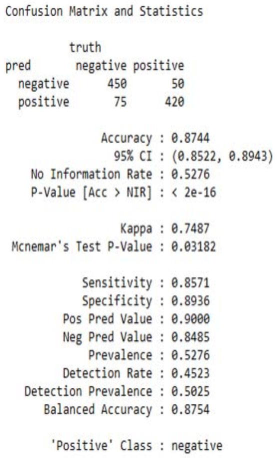
In "Fig.7.", The graphical representation of the accuracy of both the classifiers is being depicted.

In "Fig.8.", The performance of the Multinomial Naïve Bayes for various datasets is being visualized, which shows that the accuracy of MNB for all the datasets is almost greater than 85%.



```
Confusion Matrix and Statistics

          truth
pred       negative positive
  negative      380       50
  positive      165      400

               Accuracy : 0.7839
                 95% CI : (0.757, 0.8091)
    No Information Rate : 0.5477
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5734
 Mcnemar's Test P-Value : 7.561e-15

            Sensitivity : 0.6972
            Specificity : 0.8889
         Pos Pred Value : 0.8837
         Neg Pred Value : 0.7080
             Prevalence : 0.5477
         Detection Rate : 0.3819
   Detection Prevalence : 0.4322
      Balanced Accuracy : 0.7931

       'Positive' Class : negative
```

Fig.5. Confusion Matrix for Naïve Bayes Classifier



```
Confusion Matrix and Statistics

          truth
pred       negative positive
  negative      450       50
  positive       75      420

               Accuracy : 0.8744
                 95% CI : (0.8522, 0.8943)
    No Information Rate : 0.5276
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7487
 Mcnemar's Test P-Value : 0.03182

            Sensitivity : 0.8571
            Specificity : 0.8936
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.8485
             Prevalence : 0.5276
         Detection Rate : 0.4523
   Detection Prevalence : 0.5025
      Balanced Accuracy : 0.8754

       'Positive' Class : negative
```

Fig.6. Confusion Matrix for Multinomial Naïve Bayes Classifier

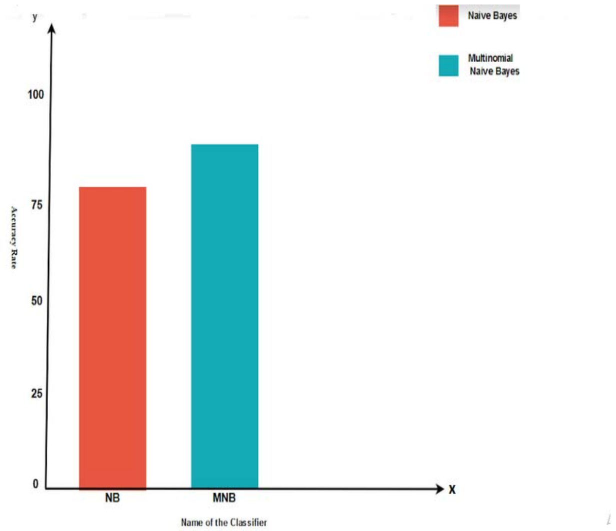


Fig.7. Accuracy depiction Graph for two classifiers

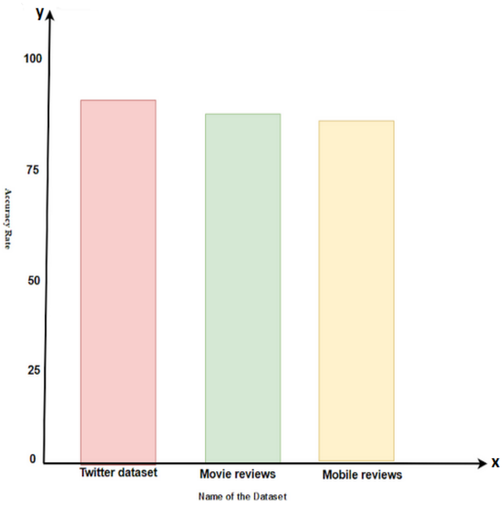


Fig.8. Accuracy Graph for multiple datasets using MNB

X. CONCLUSION

The dependency of the people [9], on online reviews is getting increased rapidly day by day, hence the gathering of the thoughts of various people is necessary for providing a platform to know what others think. The sentimental analysis is not limited to a particular dataset, It could be implemented to any kind of review dataset or social media tweets.

The proposed project majorly deals with twitter dataset in which negation handling issue is carefully handled in order to provide a better accurate algorithm, For negation handling each word is considered separately for analysis and the label for each word is being considered in order to avoid misclassification. Though the proposed algorithm is not able to provide 100% of accuracy, it has the ability to provide a better accuracy of about 87%, which could be corrected for still more accuracy in future.

XI. FUTURE WORK

The Future work is aimed to provide a better own proposed algorithm which may be a combination of two existing algorithms or a completely new algorithm, this algorithm will be applied to a different type of dataset, Negation handling issue will be dealt still more seriously.

XII. REFERENCES

[1] SoYeopYooJeInSong, OkRanJeong,“Social media contents based sentiment analysis and prediction system”, Expert Systems with Applications, Volume 105, 1 September 2018.

[2] AbinashTripathy,AnkitAgrawal,SantanuKumarRath“Classification of Sentimental Reviews Using Machine Learning Techniques”, Procedia Computer Science, Volume 57, 2015.

[3] G. Subramaniam, R.Aswini, M.Ranjitha ,Praveen Kumar Rajendran , “Survey on user emotion analysis using Twitter data”,2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).

[4] Avinash Chandra Pandey,Dharmveer Singh Rajpoot Mukesh Saraswat, “Twitter sentiment analysis using hybrid cuckoo search method”, Information Processing & Management, Volume 53, Issue 4, July 2017.

[5] Hanhoon KangSeong, JoonYooDongil Han, “Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews” Expert Systems with Applications Volume 39, Issue 5, April 2012.

[6] Noor FarizahIbrahim, XiaojunWang, “A text analytics approach for online retailing service imprsovement: Evidence from Twitter”, Decision Support Systems, Volume 121, June 2019.

[7] Eleonora D, Andrea, Pietro Ducange, AlessioBechini, Alessandro Renda, FrancescoMarcelloni, “Monitoring the public opinion about the vaccination topic from tweets analysis”, Expert Systems with Applications, Volume 116, February 2019.

[8] Mahmoud AlAyyoub, Abed Allah, Khamaiseh,YaserJararweh, Mohammed N,Al-Kabi,“A comprehensive survey of arabic sentiment analysis”, Information Processing & Management, Volume 56, Issue 2, March 2019.

[9] Xiaojiang Lei, Xueming Qian, Guoshuai Zhao “Rating Prediction based on social Sentiment from textual reviews”, IEEE Transactions On Multimedia, Volume:40, Issue:4, June 2017.

[10] RatabGull, UmarShoaib, SabaRasheed, Washma Abid, BeenishZahoor,“Pre Processing of Twitter's Data for Opinion Mining in Political Context”, Procedia Computer Science Volume 96, 2016.

[11] Kim Schouten , Onne van der Weijde,FlaviusFrasincar ,Rommert Dekker “Supervised and Unsupervised Aspect category detection for sentiment analysis with co-occurrence data”, IEEE Transactions on CyberneticsVolume: 48, Issue: 4, April 2018.

[12] Xiaojiang Lei, Xueming Qian, Guoshuai Zhao “Rating Prediction based on social Sentiment from textual reviews”, IEEE Transactions On Multimedia, Volume:40, Issue:4, June 2017.