Analysis of Text Classification of Dataset using NB-Classification

Shanti Minduru, A Eknath, Venkat Sai Chowdary Velgapudi, Moulana Mohammed

Abstract: Social Media is getting more attention now a days. People are highly influenced by comments, tweets etc. to increase their popularity. Opinions offer organizations to analyze the thinking of the people towards the success of the movie. In this paper we have taken a twitter dataset to explore the sentimental analysis using algorithms Naïve Bayes and K Cross Validation in R Programing [10]. Our implementation results shows that K Cross validation provides the accuracy followed by Naïve Bayes formula.

Index Terms: Naïve Bayes algorithm, K-Fold cross validation, Data Analysis using R.

I. INTRODUCTION

In this era of modernization, we always think of the popularity of movie. But the popularity of the movie depends upon the reviews of the people. Once the movie world introduces the movie to the theatres, we check the popularity of the movie through the reviews of the people [1]. These reviews play an important role for the movie and the industry which has produced it. It is not possible to go through every review because we have very less time to take the next step. And in this point a sentimental analysis plays an important role. Sentimental analysis is an application of mining. It is basically the process of categorizing the sentiments or opinions from a text. In sentiment analysis there are only two decisions positive and negative which make people to like or dislike the movie, or to know us that the movie is best or worse.

People will currently post reviews of product at merchandiser's sites and categorical their views. On nearly something in discussion forums and blogs, and at social network sites. Now if one wants to buy a product, one is no longer limited to asking one's friends and families because there are many user reviews on web [15]. In this paper we get the sentiments from the twitter dataset. To perform sentimental analysis we need to decide that which algorithm makes our work easy and better than other algorithms [2]. And here we provide a clear method to know the analysis of the dataset and here we decide to propose the methods i.e. Naïve Bayes and K Cross Validation. By using this we can observe comments that are seemingly to represent spam considering some indicators: a discontinuous text flow, inadequate and vulgar language or not associated with a selected context. Our approach depends on machine learning algorithms [14].

The user contributions to social media vary form web posts, tweets, reviews and photo etc. An outsized quantity of the info on the net is unstructured text. Opinions expressed in social media in type of reviews or posts represent a crucial and fascinating space value exploration and exploitation. With increase in accessibility of opinion resource like film reviews, tweets, web reviews, the new difficult task is to mine the giant volume of texts and devise appropriate algorithms to grasp the opinion of others[1].The result will be clearly visualized at further discussion of the paper.

In this paper we analyze the following using the three steps given below:

[3]Data Collection and Pre-processing: The user contributions to social media vary form web posts, tweets, reviews and photo etc. An outsized quantity of the info on the net is unstructured text. Opinions expressed in social media in type of reviews or posts represent a crucial and fascinating space value exploration and exploitation. With increase in accessibility of opinion resource like film reviews, tweets, web reviews, the new difficult task is to mine the giant volume of texts and devise appropriate algorithms to grasp the opinion of others[1].As we have given large number of dataset, the accuracy of the mining depends upon the dataset which we had given for training. To increase the efficiency of the model we pre-process the dataset as per our requirements like omitting the unnecessary things like punctuation, white stripes etc.

Mining: In a tweet message, a sentiment is sent in one or other passages, that are rather informal, as well as abbreviations and typos. Finally, a significantly giant fraction of tweets convey no sentiment whatever, like advertisements and links to news articles, which offer some difficulties in knowledge gathering, training and testing[5].Two main algorithms we have used here is K fold cross algorithm and [15] Naïve Bayes which are implemented and evaluated.

Result: The accuracy of the given algorithms is retrieved from the model by performing sentiment analysis [4] on the twitter movie dataset using R Programming. These results are visualized using charts.

Revised Manuscript Received on March 10, 2019.

Shanti Minduru, Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India-522502

A.Eknath, Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India-522502

Venkat Sai Chowdary Velgapudi Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur ,Andhra Pradesh, India-522502

Moulana Mohammed, Associate Professor, Department of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India-522502

II. LITERATURE SURVEY

In this section we take the overlook of some of the papers which has already done some related work on this:

[1] The purpose of this study is to analyze the performance of sentimental analysis in the words of accuracy, precision and recall. In this work they compared the Naïve Bayes algorithm and K Nearest Neighbor Algorithm. Also they used Chi Score test to score the words.

[2] The authors create the structure which divides the positive and negative assessments of each feature in the corpus. The framework comprises of two stages, first stage is to recognition of sentiments [5] and second stage is relative scoring for each and every element.

[3] According to the idea of psychology, human mood derived from the primordial social activities,[6] and a positive or negative mood will influence one another among one community it is being observed that machine learning techniques for sentimental analysis for movie reviews are successful the features which have been picked up had a sensational impact on precisions of classifier.

III. PROPOSED SYSTEM

A. Data Input:

There are 2 ways that to enter input to the film review sentiment analyzer. The number is exponentially rising with the growing quality of twitter website. Sentimental analysis has created it potential to research the moods of someone. It will facilitate North American nation to determine the positive, negative [7].One by providing a list of reviews in JSON file format or by providing TMDB ID of film title. [7]In case of TMDB ID a TMDB JSON API is utilized to fetch and store reviews in MySQL Database. [6]After fetching reviews first ten reviews for a specific title are utilized by the system for sentiment analysis.

A. Part of Speech Tagging:

POS is employed to clear up a sentence so as to extract features from a sentence. In POS tagging every words labelled, it is used to verify word position within the grammatical context.POS tagging helps to seek out nouns,phrases,verbs,adjectives in a very sentence. After POS Tagging there's a bit likelihood selected word may be a discarded word for feature choice and opinion words. Most of previously microblog sentimental analysis focuses on twitter and particularly in English.[13] However, the analysis of Chinese microblogs has some notable variations thereupon of twitter.

B. Features and Opinion Words Extraction:

All the opinion words are chosen from the sentence. The system extracts all nouns, noun phrases, verbs and adjectives from the film review and compares with the present list of words. These words are classified on the basis of their polarity. For Instance "good" word is a positive polarity. On the opposite hand, options are chosen on basis of variety time's incidence of opinion words. If opinion word is an event in review over the edge price then it's more options list. For this method API is trained just for film reviews with the keyword and the phrases wordbook which has "good acting", solid story" and "awesome action".

C. Identify Sentence Polarity:

After extracting all options and opinion word, it's terribly simple to find the polarity of the sentence. Sentence polarity follows the same rules as arithmetic expressions. A negative sentiment contains all negative opinion words and positive sentiment contains all positive opinion words. A negative sentiment might contain a positive opinion word. For Example: "This picture story isn't good" sentence in a picture review. During this sentence, "good" opinion word is of positive polarity however "not" may be a negative word. Therefore, the general polarity of this sentence will be negative.

D. Identify Review Polarity:

Whole review polarity depends on variety of total positive or negative sentences found in a review. Although reviews concerning merchandise and services are extravagantly offered on web, choosing relevant info needs a possible purchaser to pay a major quantity of our time reading reviews and removing comments unrelated to the vital aspects of the reviewed entity [8].

If the amount of positive sentences is larger than the quantity of total negative sentences then review polarity are positive. Similarly, a review polarity are negative if the quantity of total negative sentences is larger than the quantity of positive sentences.

E. Classification of Review:

Once, review polarity is calculated. The percentage of review polarity and polarity (positive or negative) are classified and saved for further analysis. One in all the plain challenges in classifying matter in film reviews is that sentiment words usually relate to the elements of a films instead of reviewer's opinions [9].With additional analysis, workplace collection will be expected and overall performance of show may be predicted.

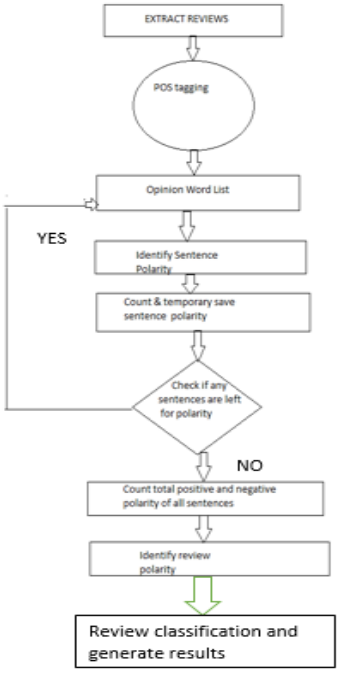

Fig 1.Flowchart of sentimental Analysis

IV. EXPERIMENTAL RESULTS

This model is implemented has given us good outputs along with the accuracy that is proposed by the system. This segment demonstrates the aftereffects of various tests that I executed. Here the dataset we are collected for the experiment is the link given in the reference. To begin with, the consequences of pre-preparing are appeared. These outcomes are trailed by consequences of the component determination and characterization .these consequences of the component determination and the order are joined since they can't be isolated from each other. The order calculation needs the components to group, and the element choice does not accomplish significant outcomes without the arrangement. With the explosive growth of user generated messages, Twitter has become a social website wherever a lot of users will exchange their opinion. Sentimental analysis on twitter knowledge has provided a cost effective and effective thanks to expose opinion timely that is important for deciding in numerous domains [12].
The details of the project experiment is as follows :
Here, we have taken a twitter movie dataset where there are 1000 tweets, these tweets consists of both positive and negative tweets.
We used two algorithms, analyzed which performs better. We have taken the same data set for both the algorithms and the outputs were analyzed depending upon their accuracy and efficiency. The two algorithms are Naïve Bayes algorithm and K Fold Cross validation.
The first thing we have to do is Data Pre-Processing. This helps in improving results through the classified algorithms. It includes the following steps:

1. Remove empty rows.
2. Change the text to lower case. This is required as the programming language interprets the words differently [8].
3. Tokenization: In this every entry in the corpus will be broken into set of words.
4. Remove a stop words, Non-Numeric and perform word stemming.

After the data is pre-processed we then split the model into train and test data set. This is done to find how important a word in document is in comparison to the corpus. Now, the final step includes running the two classification algorithms to classify out data check for accuracy. We perform the following operations on both the classification algorithms:
The arrangement utilizing Naïve Bayesian is done as takes after to start with, every one [1] of the tweets and marks are passed to the classifier. In the subsequent stage, highlight extraction is finished. [14]Presently, both these separated elements and tweets are passed to the Naïve Bayesian classifier. At that point prepare the classifier with this preparation information. At that point the [1] classifier dump record opened in compose back mode and highlight words are put away in it alongside a classifier. After that the document is close.
The accuracy result shown by Naïve Bayes algorithm is:

```
             Actual
Predictions Neg Pos
        Neg 197  59
        Pos  47 196
```

Fig 2.Accuracy Result

The confusion Matrix and statistics using Naïve Bayes Algorithm are:

```
Confusion Matrix and Statistics

          Reference
Prediction Neg Pos
       Neg 197  59
       Pos  47 196

               Accuracy : 0.7876
                 95% CI : (0.749, 0.8227)
    No Information Rate : 0.511
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5754
```

Fig 3.Confusion matrix and statistics using Naïve Bayes

```
Mcnemar's Test P-Value : 0.2853

            Sensitivity : 0.8074
            Specificity : 0.7686
         Pos Pred Value : 0.7695
         Neg Pred Value : 0.8066
             Prevalence : 0.4890
         Detection Rate : 0.3948
   Detection Prevalence : 0.5130
      Balanced Accuracy : 0.7880

       'Positive' Class : Neg
```

Fig4.Accuracy for overall data

```
> conf.mat$byClass
      Sensitivity       Specificity    Pos Pred Value    Neg Pred Value
        0.8073770         0.7686275         0.7695312         0.8065844
        Precision            Recall                F1        Prevalence
        0.7695312         0.8073770         0.7880000         0.4889780
Detection Rate Detection Prevalence Balanced Accuracy
        0.3947896         0.5130261         0.7880023
```

Fig 5.Confusion matrix for class

```
> conf.mat$overall
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull AccuracyPValue
  7.875752e-01   5.753941e-01   7.490314e-01   8.226714e-01   5.110220e-01   2.173991e-37
 McnemarPValue
  2.853336e-01
>
```

Fig 6.Confustion matix for overall result

The graphical visualization of the experiment is described by taking X-axis as class label and Y-axis as number of comments:

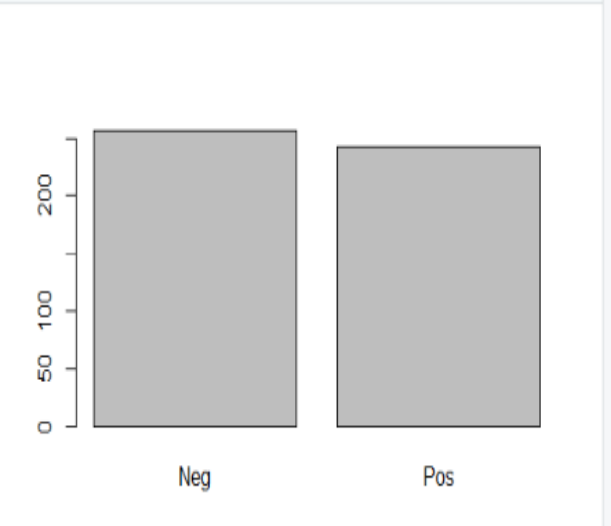

Fig7. Prediction of class as positive and negative

In this we can observe that the negative comments on the movie are more than the positive. From this project we obtained the accuracy score of both naïve Bayes and K fold Cross Validation of the dataset.

V. CONCLUSION

Consequently we reason that the machine learning system is extremely simpler and proficient than typical procedures. These systems are effectively connected to twitter notion investigation. The biggest obstacle to adopting analytics is the lack of knowhow about using it to improve business performance. Business Analytics uses applied mathematics, research and management tools to drive business performance[10].Twitter conclusion investigation is troublesome in light of the fact that it is exceptionally difficult to distinguish enthusiastic words from tweets and furthermore because of the nearness of the rehashed characters, slang words, void areas, incorrect spellings and so on. The abundance of social media information provides opportunities however additionally presents method difficulties for analyzing large scale informal matter information [11].

Grouping exactness of the element vector is tried utilizing classifier like Naïve Bayes. The presumption of Naïve Bayes that the information is free, turned out to be an amazing device in this examination. It was found by the creator that Machine learning calculations were more straightforward to actualize and more effective than different parts of the paper as they delivered a table which considered straightforwardness in the exactness of the Naive Bayes grouping. Generally speaking the half breed way to deal with opinion [13] investigation considered an intensive examination of the information and performs well for a Twitter dataset. In any case, the precision of the Naïve Bayes classifier still leaves opportunity to get better this might be accomplished by better pre-preparing.

FUTURE SCOPE:

The pertinence of slant investigation for future organizations and showcasing in utilizing watch words and examination of the notions around that catchphrase by general society is just going to increment as the notoriety of Twitter becomes throughout the following couple of years. Be that as it may, as far as long haul improvement [2] or research, the capacity of the twitter API to pull information [2] that is more established, ought to be created and in addition another online networking [2] API's so that estimation examination could be performed over some stretch of time, particularly in the domain of sociologies where specialists could enquire into social and political movements of sentiment on the web-based social networking locales [15]. Similarly the absence of progress in feeling after some time on a few issues may be worth seeking after as a point of research for twitter slant examination. The convenience of such an opinion analyzer would take into account an intriguing examination of social and political issues.

REFERENCES

1. Singh, V. K., etal. "Sentiment analysis of movie reviews: “A new feature based heuristic for aspect-level sentiment classification.” Automation, Computing, Communication, Control and Compressed Sensing 1] (iMac4s), international MultiConference on. IEEE, 2013.
2. Kamps, J. Marx, M. Mokken, R. J. De Rijke, M. (2004)"Using word net to measure semantic orientations of adjectives," Found in Neethu, M. Rajasree R.(2013) ‘Sentiment Analysis in Twitter using Machine Learning Techniques' 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT) Tiruchengode India. July 4-6 2013. IEEE.pp1-5 [Accessed April 25th 2014.
3. A. Balahur, J. Hermida and A. Montoyo, 'Building and Exploiting Emotional, a knowledge base for motion detection based on the appraisal theory model', Affective Computing, IEEE Transactions, vol. 3, 188101, 2012.
4. Z. Niu, Z. Yin and X. Kong, 'Sentiment classification for microblog by machine learning,', Computational and Information Sciences(ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, vol. 286289, 2012.
5. A. Celikyilmaz, D. Hakkani-Tur and J. Feng, 'Probabilistic Model-Based Sentiment Analysis of Twitter Messages', Spoken Language Technology Workshop (SLT), 2010 IEEE, vol. 7984, 2010.
6. Y. Wu and F. Ren, 'Learning sentimental influence in twitter', Future Computer Sciences and Application (ICFCSA), 2011International Conference, IEEE, vol. 119122, 2011.
7. Monu Kumar Thapar University, Patiala “Analysing Twitter sentiments through big data”, IEEE, 2016
8. Giuseppe Di Fabbrizio A&T Research Labs, USA “Summarizing Online Reviews Using Aspect Rating Distributions and Language Modelling”, Digital Object Identifier IEEE, 2013
9. Martin Wöllmer Technical University of Munich, Germany “YouTube movie reviews- Sentiment analysis in an audio-visual”, IEEE Computer Society, 2013
10. Shruti Kohli,HimaniSingal, “Data Analysis with R”, 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.
11. Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan, “Mining Social Media Data for Understanding Students’ Learning Experiences”, IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 7, NO. 3, JULY-SEPTEMBER 2014
12. Shu long Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Member, IEEE, Jiajun Bu, Member, IEEE, Chun Chen, Member, IEEE, and Xiaofei He, Member, IEEE,” Interpreting the Public Sentiment Variations on Twitter”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, MAY 2014.
13. Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao, “CMiner: Opinion Extraction and Summarization for Chinese Microblogs”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 7, JULY 2016.
14. Radulescu, C., Dinsoreanu, M., and Potolea, R., “Identification of spam comments using natural language processing techniques”, Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on, September 2014.

15. Liu, B., "Sentiment Analysis: A Multi-Faceted Problem", IEEE Intelligent Systems, 2010.
16. http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/movie-pang02.zip

AUTHORS PROFILE:

Shanti Minduru, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram, Andhra Pradesh. She is doing her research work in knowledge engineering.

A Eknath, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram ,Andhra Pradesh. She is doing her research work in knowledge engineering.

Dr Moulana Mohammed received his Ph.D.in Computer Science from Bharathiar University in 2018 and M.Tech in CSE from JNTUK in 2009.He is an Associate Professor in the Computer Science & Engineering Department at KLUniversity. His research areas include data mining, bioinformatics, IoT and big data analytics.

Venkat Sai Chowdary Velgapudi, is a student at the department of Computer Science and Engineering at K L Educational foundation, Deemed to be University, Vaddeswaram ,Andhra Pradesh. He is doing his research work in knowledge engineering.