GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

Analysis Of Text Classification Of Dataset Using NB- Classifier

A Report for the Evaluation 3 of Project 2

Submitted by

ASIF ANSARI

(1613101202/16SCSE101841)

in partial fulfillment for the award of the degree of

Bachelor of Technology

IN

Computer Science and Engineering

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr . SREENARAYANAN NM

Assistant Professor

April/2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "**Analysis Of Text Classification Of Dataset Using NB- Classifier** " is the bonafide work of **ASIF ANSARI** (1613101205)" who carried out the project work under my supervision.

SIGNATURE OF HEAD
Dr. MUNISH SHABARWAL,
PhD (Management), PhD (CS)
Professor & Dean
School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR
SREEMARAYANAN NM
Assistant Professor
School of Computer science & Engineering

TABLE OF CONTENTS

Abstract

This is era of Modernization, when we want to buy something we always think of the genuine and popularity of the product. But the popularity of the product depends upon the review which is given by the people. For example once movie released, we check the popularity of the movie through the reviews, comments and rating which is given by different-different people. These reviews play an important role for the movie and the industry and also for public.

In this busy life people don't have lot of time to go through each and every reviews. And at this point sentimental analysis plays an important role. This is basically the process of categorizing the sentiments or opinions of the people. In sentiment analysis there are only two decisions positive and negative which makes people to like or dislike the movie or anything, People will currently post their reviews of product at commercial sites about the products. Now if anyone wants to buy a product, one is simply just go through the reviews, ratings and one make the decision very easily .In this paper Naives Bayes classifier use to perform the sentimental analysis. To perform the process Machine learning algorithm use to reach to the approaches. For that Features selection does improve the performance of sentiment classification, so we have to select more and more feature for accurate classification. So Naives Bayes classification performs better and fast than other classification methods.

Social Media is getting more attention now a days. People are highly influenced by comments, tweets etc. to increase their popularity.. In this paper we have taken a dataset to explore the sentimental analysis using algorithms Naïve Bayes. Our implementation results shows that K Cross validation provides the accuracy followed by Naïve Bayes formula. And in this point a sentimental analysis plays an important role. Sentimental analysis is an application of mining. It is basically the process of categorising 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.

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.
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. Opinions offer organisations to analyse the thinking of the people towards the success of the movie. 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 .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.

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 other. 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. 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 on the twitter movie dataset using Python programming .These results are visualized using charts.

And in this point a sentimental analysis plays an important role. Sentimental analysis is an application of mining. 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. 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. The result will be clearly visualized at further discussion of the paper.

The World Wide Web and therefore the web give a forum through that an individual's process of higher cognitive could also be influenced by the opinions of others. For instance, the customer feedback system utilized by ebay.com permits customers to use free-form of text to rate the products and services received whereas create the ratings on the market to alternative customers to review before they create a sale call in impact permitting a client to form an additional informed call. Client feedback and merchandise evaluations may be found at several online sites together with epinions.com and amazon.com. Online sites like rotten tomatoes.com, permit picture buffs to go away opinions and comments. Other online sites, like cnn command globe and mail.com, permit readers to go away comments

These varieties of online media have resulted in massive Analysis of Text Classification of Dataset Using NB-Classifier quantities of matter knowledge containing opinion and facts. Over the years, there has been intensive analysis geared toward analyzing and classifying text and knowledge, wherever the target is to assign predefined class labels to documents primarily based upon
learned models. However, more modern analysis has tried to analyze matter knowledge to work out however a personal "feels" a couple of specific topic(i.e., the individual's sentiment towards that topic).

This has diode to the event of sentiment analysis and classification systems. Sentiment Analysis and classification are technically challenging As a result of opinions will be expressed in delicate and complicated ways that, involving the use of slang, ambiguity, sarcasm, irony and idiom. Sentimental analysis and classification is performed for may reasons, for instance to track the ups and downs of combination attitudes to a complete or product to match the attitudes of online customers between one complete or product and another, and to tug out of specific analysis and classification systems. Sentiment analysis and classification are technically challenging as a result of opinions will be expressed in delicate and complicated ways that, involving the use of slang, ambiguity, sarcasm, irony and idiom. Sentiment analysis and classification is performed for many reasons, for instance to track the ups and downs of combination attitudes to a complete or product to match the attitudes of on-line customers between one complete or product and another ,and to tug out examples of specific varieties of positive or negative statements on some topic. It's going to even be performed to enhance client relationship management and to assist alternative potential customer create wise choices. The remainder of the chapter is organized as follows, we give a statement of the matter. We tend to describe the steps concerned in sentiment analysis and classification.

Literature Survey

[1] Balahur et al (2009) presented a comparison on the techniques as well as resources which may be utilized for mining opinions from quotations in news articles. The challenges in the task were shown, motivated by the possibility of various targets as well as a huge set of affect phenomena which quotes comprise. The proposed methods evaluated were evaluated utilizing annotated quotations taken from news given by the EMM news collecting engine. A general OM system needs usage of both huge lexicons as well as specialized training as well as testing data.

[2] Schneider et al., (2009) proposed a novel matrix learning strategy for extending relevance learning vector quantization (RLVQ), an effective prototype-based classification protocol, toward a general adaptive measure. Through introduction of a full matrix of relevance factors in the distance metric, correlations between various attributes as well as their significance for classification occurs at the time of training. When contrasted with weighted Euclidean measure utilized in RLVQ as well as its variants, a complete matrix is more powerful for representing the internal structure of data adequately. Huge margin generalization bounds may be transferred to the case resulting in bounds that are not dependent on input dimensionality.
This is true for local measures attached to all prototypes that correspond to piece-wise quadratic decision bounds. The protocol was evaluated in contrast to alternate LVQ strategies through usage of artificial dataset, a benchmark multiclass issue from UCI repository, as well as a problem from bio-informatics, the recognition of splice sites for C.

[3] Martin Wollmer et al. proposed method performer sentiment classification for audio plus video reviews of user. Review for a movie is given in 2 minute YouTube video. For sentiment classification of such reviews method use automatic speech recognition system and
video recognition system. For better classification of reviews vocal and face expression play vital role. Richard Socher et al.shows that semantic word space are very useful but they can't used with long sentences. That why, Sentiment Treebank was introduced. This Treebank consist of various parse trees to classify the sentence into the one of classes of sentiments.

[4] Martin Wollmer et al., [12] proposed method performer sentiment classification for audio plus video reviews of user. Review for a movie is given in 2 minute YouTube video for sentiment classification of such reviews method use automatic speech recognition system and video recognition system. For better classification of reviews vocal and face expression play vital role. Richard Socher et al., shows that semantic word space is very useful but they can't used with long sentences. That why, Sentiment Treebank was introduced. This Treebank consist of various parse trees to classify the sentence into the one of classes of sentiments. Recursive Neural Tensor network is the example of such method. One example is taken, to understand how this method works. Example review is "This film doesn't care about cleverness, wit or any other kind of intelligent humours. it divide the sentence into token and make tree structure which divide the comment into one of the class label. Sentence is taken and then using Treebank concept it is accurately classify into one of five classes. Five class labels are very negative (- -), negative (-), neutral (0), positive (+), and very positive (+ +).

[5] Li et. al studied online forums hotspot and forecast using sentiment analysis and text mining approaches. First of all, to inspect the sentiment polarity for each piece of text, an algorithm was created. Afterwards to develop unsupervised text mining approach the algorithm was joined with k-means clustering and support vector machine (SVM). Described text mining approach had been used to group forums into various clusters, whose centre represent a hotspot forum within the current time span. The datasets had been taken from SINA sports forum. Experimental results showed that SVM forecasting gets high consistent results with k-means clustering. The top 10 hotspot forums given by SVM forecasting resembles 80% of k-means clustering results. Both SVM and k-means achieved the same results for the top 4 hotspot forums of the year. In this paper they had created an algorithm that automatically analyze the sentiment polarity of a text with the help of which text values were obtained. Influential power of text was represented by absolute value and sentiment polarity by the sign of text. Previously created algorithm was then combined with k-means clustering and SVM classification to integrated approach for online sports forums cluster analysis. Unsupervised algorithm had been applied to group the forums into various clusters, whose center represent hotspot forum with the current time span. In addition to clustering the forums based on data from the current time window, forecasting for the next window was also done by them. Proof for existence of correlations between post

text sentiment and hotspot distribution was given by empirical studies. Results showed that both SVM and k-means produce consistent natural groupings.

Several companies could be benefited from these hotspot predicting approaches in different Analysis of Text Classification of Dataset Using NB-Classifier ways. These companies could also combine results for market basket analysis to yield comprehensive decision support information. A firm in financial sector or the financial of a giant company might get profit from such a sentimental and text mining process. In financial market, right before a security market opens and trading begins, analysts people on sales and trading desks usually try to get an overall fix on market sentiment and for particular investments. First, algorithm design could be improved to yield a more accurate calculation of sentiment. Even for supervised learning, algorithms other than SVM, or variations of SVM, could be joined as well. Secondly, they had incorporated topic extraction. Third, a practical system, in the form of a website portal, was desired as their major future work.

Proposed model

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. One by providing a list of reviews in JSON file format or by providing TMDB ID of film title. In case of TMDB ID a TMDB JSON API is utilized to fetch and store reviews in MySQL Database. After fetching reviews first ten reviews for a specific title are utilized by the system for sentiment analysis.

B.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 labeled, 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 micro blog sentimental analysis focuses on twitter and particularly in English. However, the analysis of Chinese micro blogs has some notable variations there upon of twitter.

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.

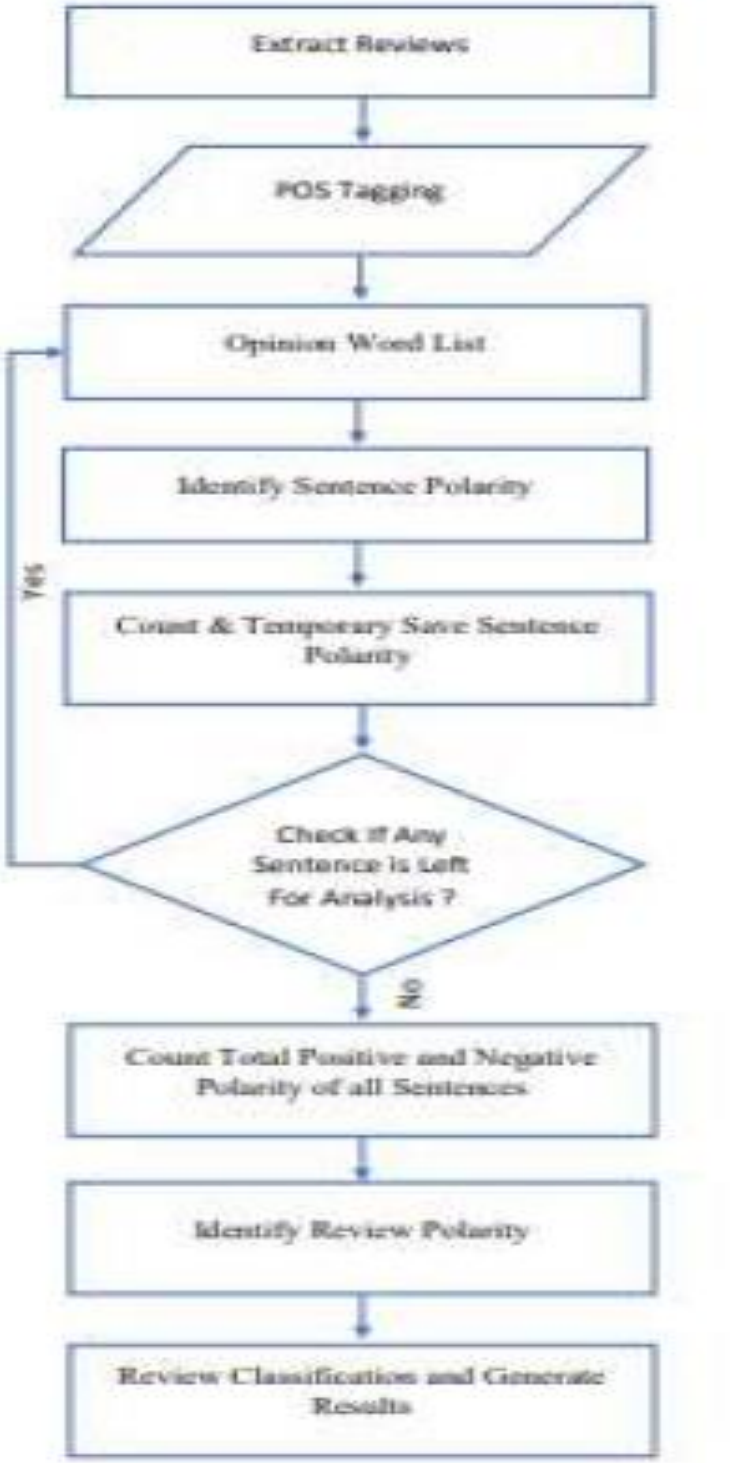
D.Features and Opinion Words Extraction:
All the opinion words are chosen from the sentence. The system extracts all nouns, noun phrases, verbs and adjectives from the film review and compares with the present list of words. These words are classified on the basis of their polarity. For Instance "good" 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".

E.Identify Review Polarity:

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.

F.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 .With additional analysis, workplace collection will be expected and overall performance of show may be predicted.



Sentimental Analysis in Proposed System

Existing System

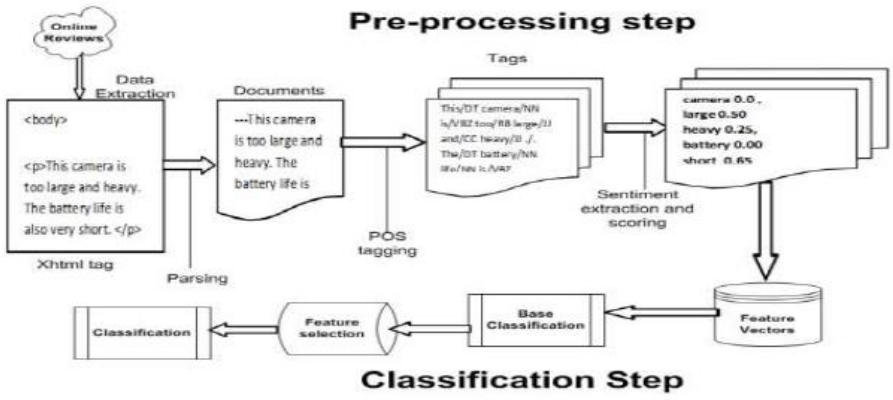
This model is implemented has given us good outputs along with the accuracy that is proposed by the system. This segment demonstrates the aftereffects of various tests that I executed. Here the dataset we are collected for the experiment is the link given in the reference. To begin with, the consequences 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. 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 .
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 of the tweets and marks are passed to the classifier. In the subsequent stage, highlight extraction is finished. 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 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. Each dataset is assumed to be restricted to one type of object. For example, all the movie reviews will be about movies.

 Unigrams (individual terms in a review e.g., clear, noisy) are used as features (i.e. we do not use bi-grams and higher forms of n-grams).
 Stemming (the process of reducing inflected words to their base or root form. e.g. “clearly” is stemmed to “clear”) could distort the part-of-speech of a term, so it is not used.
 For tagging, our aim is to find the sentiment of the whole document, thus whole documents are tagged rather than sentences.
 For terms with multiple polarity scores, the average of all these scores is used.
 The percentage of the number of unique terms found in the documents which will not be found in the SentiWordNet dictionary is considered to be minimal and so will be ignored.
 Terms not found in the SentiWordNet dictionary will be regarded as non sentiment bearing words.



Implementation

We use NLTK's Naive Bayes classifier for our task here. In the feature extractor function, we basically extract all the unique words. However, the NLTK classifier needs the data to be arranged in the form of a dictionary. Hence, we arranged it in such a way that the NLTK classifier object can ingest it.

Once we divide the data into training and testing datasets, we train the classifier to categorize the sentences into positive and negative. If you look at the top informative words, you can see that we have words such as "outstanding" to indicate positive reviews and words such as "insulting" to indicate negative reviews. This is interesting information because it tells us what words are being used to indicate strong reactions.

How to Perform Sentiment Analysis:-

Step 1: Create a new Python file, and import the following packages.

Step 2: Define a function to extract features.

Step 3: We need training data for this, so we will use movie reviews in NLTK.

Step 4: Let's separate these into positive and negative reviews.

Step 5: Divide the data into training and testing datasets.

Step 6: Extract the features.

Step 7: We will use a Naive Bayes classifier. Define the object and train it.

Step 8: The classifier object contains the most informative words that it obtained during analysis. These words basically have a strong say in what's classified as a positive or a negative review. Let's print them out.

Step 9: Create a couple of random input sentences.

Step 10: Run the classifier on those input sentences and obtain the predictions.

Step 11: Print the output.

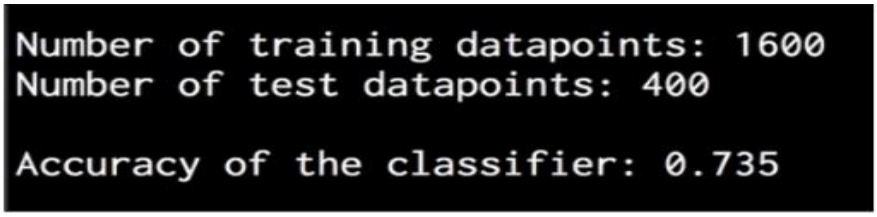
#CODE

```
import nltk.classify.util
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import movie_reviews
def extract_features(word_list):
 return dict([(word, True) for word in word_list])
if __name__=='__main__':
 positive_fileids = movie_reviews.fileids('pos')
 negative_fileids = movie_reviews.fileids('neg')
features_positive = [(extract_features(movie_reviews.words(fileids=[f])),'Positive')
for f in positive_fileids]
features_negative =
[(extract_features(movie_reviews.words(fileids=[f])),'Negative') for f in
negative_fileids]
# Split the data into train and test (80/20)
threshold_factor = 0.8
threshold_positive = int(threshold_factor * len(features_positive))
```

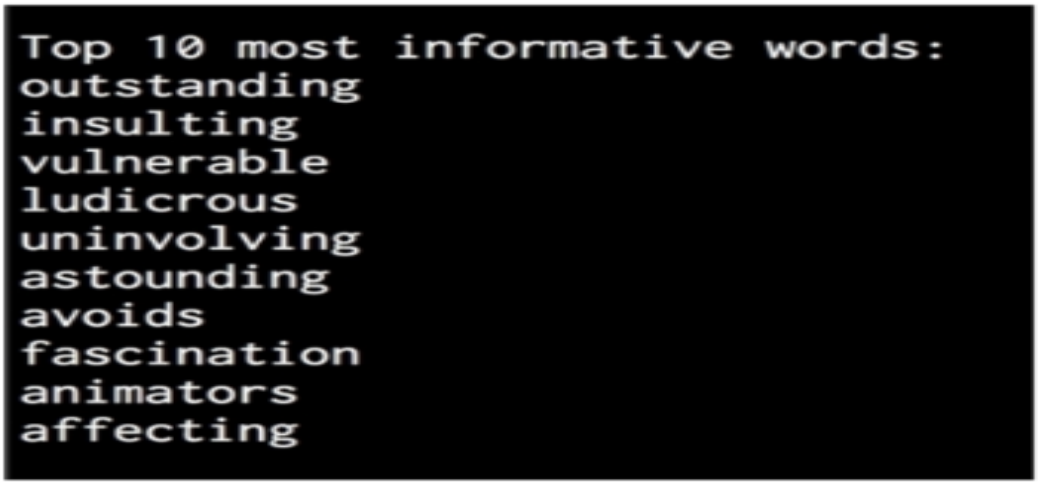
```
threshold_negative = int(threshold_factor * len(features_negative))
features_train = features_positive[:threshold_positive] +
features_negative[:threshold_negative]
features_test = features_positive[threshold_positive:] +
features_negative[threshold_negative:]
print("Number of training datapoints:\n")
len(features_train)
print("Number of test datapoints:\n")
len(features_test)
# Train a Naive Bayes classifier
classifier = NaiveBayesClassifier.train(features_train)
print("\nAccuracy of the
classifier:"),nltk.classify.util.accuracy(classifierfeatures_test)


# Sample input reviews
input_reviews = [
"It is an amazing movie",
"This is a dull movie. I would never recommend it to anyone.",
"The cinematography is pretty great in this movie",
"The direction was terrible and the story was all over the place"]
print("\nPredictions:")
for review in input_reviews:
 print(("\nReview:"),review)
probdist=classifier.prob_classify(extract_features(review.split()))
pred_sentiment = probdist.max()
print(("Predicted sentiment:"),pred_sentiment)
print("Probability:",round(probdist.prob(pred_sentiment), 2))
```

Result

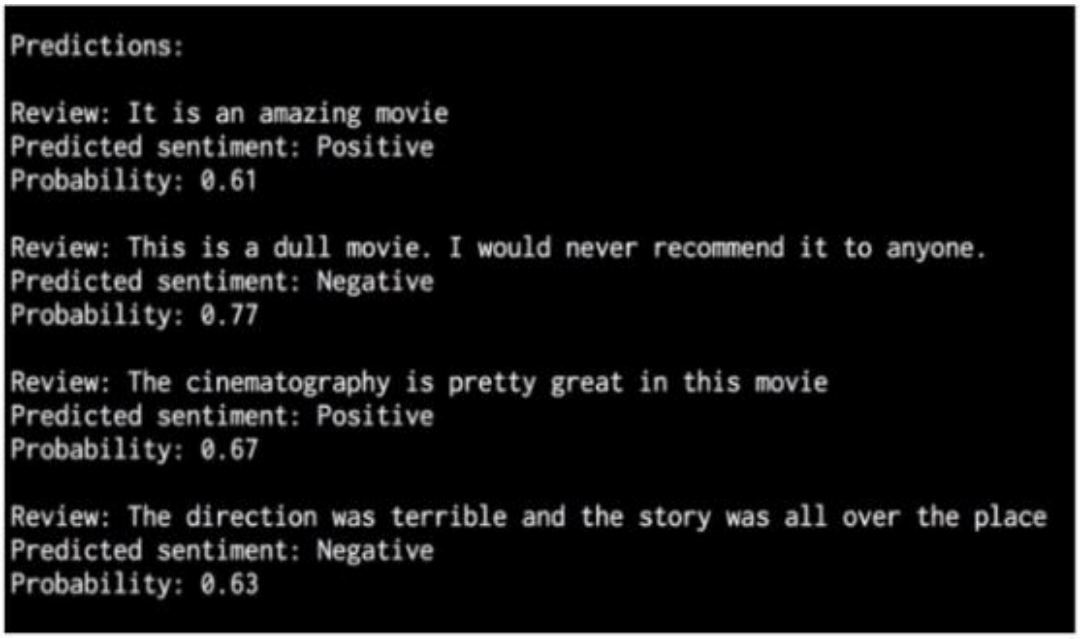
The first is the accuracy, as shown in the following image:



The next is a list of most informative words:



The last is the list of predictions, which are based on the input sentences:



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.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. 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 straight forward to actualize and more effective than different parts of the paper as they delivered a table which considered straightforwardness in the exactness of the Naïve Bayes grouping. Generally speaking the half breed way to deal with opinion 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 Work

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 or research, the capacity of the twitter API to pull information that is more established, ought to be created and in addition another online networking 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 . 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]. Shruti Kohli, Himani Singal, “Data Analysis with R”, 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.

[2]. Singh, V. K., etal. "Sentiment analysis of movie reviews: “A new feature based heuristic for aspect-level sentiment classification.” Automation Computing, Communication, controland compressed Sensing (iMac4s), international Multi Conference on. IEEE-2013.

[3]. Martin Wöllmer Technical University of Munich, Germany “YouTube movie reviews- Sentiment analysis in an audio-visual”, IEEE Computer Society, 2013

[4]. Liu, B., “Sentiment Analysis: A Multi-Faceted Problem”, IEEE Intelligent Systems, 2010.

[5]. 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]. Monu Kumar Thapar University, Patiala “Analysing Twitter sentiments through big data”, IEEE, 2016

http://boston.lti.cs.cmu.edu/classes/95-865-K/HW/HW3/movie-pang02.zip

https://www.sciencedirect.com/science/article/pii/S2090447914000550#s0015

http://ww25.gbsheli.com/2009/03/twitgraph-en.html

http://help.sentiment140.com/for-students.