Sentimental Analysis of Movie Reviews using Twitter

Monisha V. N
Dept. of Computer Science and Engineering,
Cambridge Institute of Technology,
Bengaluru, India.

Madhumitha N
Dept. of Computer Science and Engineering
Cambridge Institute of Technology
Bengaluru, India.

Nimishamba B
Dept. of Computer Science and Engineering,
Cambridge Institute of Technology,
Bengaluru, India.

Poornima N
Dept. of Computer Science and Engineering,
Cambridge Institute of Technology,
Bengaluru, India.

Dr. Shashikumar D. R
HOD, Dept. of Computer Science and Engineering,
Cambridge Institute of Technology,
Bengaluru, India.

Abstract*—One of the most well known Microblogging site is twitter. Numerous individuals utilize this stage to communicate their feelings by tweeting. Sentimental analysis is process where one can foresee the feelings of individual utilizing text investigation strategies. It is one of the main study areas in Natural Language Processing (NLP). This will be useful for many applications from product review to political critics. In this paper we will anticipate the feelings of tweets which are extricated from twitter account utilizing twitter programming interface. We utilize two calculations to manufacture the model one is Naive Bayes and another is SVM. These two algorithms give best outcomes for text arrangement where it learns how to classify the given input from trained dataset and hence it is supervised learning. The preparation dataset is film audits which are utilized to prepare the model. The forecast is made on tweets extricated as positive, negative and neutral. The point is to assemble an effective model to foresee the emotions of tweet with the goal that watcher can know whether a film merits is worth viewing or not and furthermore to accomplish better precision.

Keywords*— *Sentimental analysis; Microblogging; Twitter API; Twitter; NLP; Navie Bayes; SVM.

I. INTRODUCTION

These days, web based life is turning out to be increasingly more mainstream since cell phones can get to interpersonal organizations effectively from any place. In this way, web based life is turning into a significant theme for research in numerous fields. As the quantity of individuals utilizing the interpersonal organization is developing step by step, to speak with their friends so they can share their own inclination consistently, sees are made for a huge scope. Observing or following online networking is the most significant point in the current situation. The test dataset is tweets obtained from twitter account using twitter programming interface. Today numerous organizations have been utilizing internet based life showcasing to publicize their items or brands. For developing an online life checking, different devices have been required which includes two segments: one to assess what number of clients of their image are pulled in because of their advancement and second to discover people's opinion of the specific brand. Different stages like Facebook, Twitter, LinkedIn, Instagram permits various individuals to remark, see, judge on wide themes going from instruction, governmental issues, and diversion. These stages contain mass information in type of tweets, websites, and posts extra. Sentimental analysis plans to decide the contrariety of feelings like joy, distress, misery, scorn, outrage and warmth and suppositions from the content, audits, posts which are accessible online on these stages. Nostalgic examination that is Sentimental analysis is confounded procedure because the content contains slang words, emojis and rehashed letter use and so on. Subsequently it is one of the wide exploration zones and has its application in pretty much every area from business to social.

In this paper we utilize two calculations Naive Bayes and SVM to assemble the model and compute their precision independently. The training dataset is film reviews which contains 8500 reviews downloaded from kaggle, it is labeled dataset. Kaggle is a platform where various Data Scientists around the world can challenge, compete and learn about every aspect of Data science field practically. Right off the bat the information is preprocessed, then it is vectorized and fed into the model to foresee the feeling. At last the framework is built for a watcher to know whether a film merits watching or not and furthermore to give better efficiency.

II. WHAT IS SENTIMENTAL ANALYSIS?

Sentimental analysis alludes to the utilization of common language preparing to distinguish and separate uneven data in source materials or just alludes to the way toward recognizing the extremity of the content. It is additionally alluded to as assessment mining, as it infers the sentiment or the mentality of a client. A typical way to deal with utilizing this is portrayed how individuals consider a specific point.

Estimation investigation helps in deciding the musings of a speaker or an author as for some topic or the general relevant extremity of an archive. The disposition might be their choice or gauge, the enthusiastic condition of the client while composing. Feeling Investigation can be utilized to decide the assessment on an assortment of levels. It will rank as either positive or negative, and it will likewise rank the response of words or expressions in the information.

Slant Investigation can follow a specific subject, numerous organizations use it to follow or watch their items, administrations or status by and large. For instance, in the event that somebody is assaulting your image via web-based networking media, conclusion investigation will score the post as massively negative, and you can make alarms for posts with hyper-negative assessment scores. The below figure Fig.1 give the complete idea about the sentiment analysis in step by step manner for easy understanding.

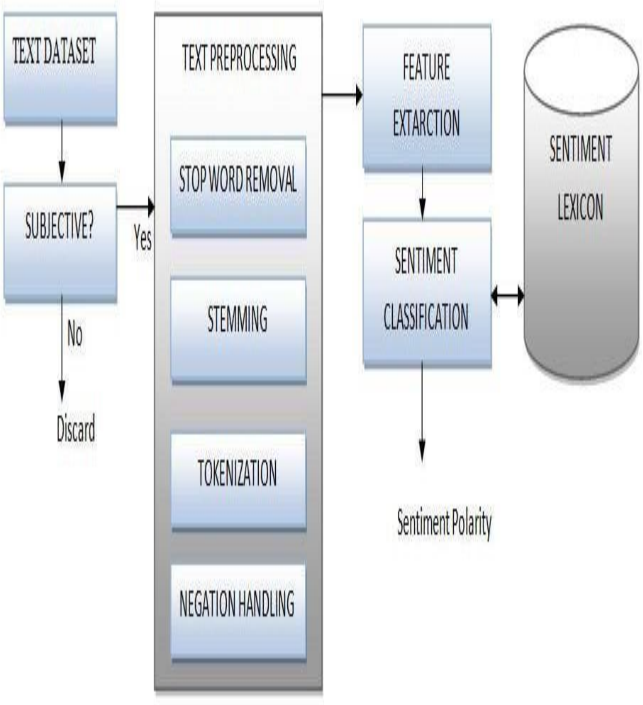


Fig.1. Sentimental Analysis

III. ALGORITHMS USED

Two algorithms used are Naive Bayes and SVM (support vector machine) the tweets will be classified as positive, negative and neutral. The tweets will be converted into Vector format using CountVectorizer and then fed into the two different algorithm models developed and output will be obtained individually.

A. CountVectorizer

The information retrieval from any dataset or document is very important step. Raw data cannot be used as input to machine learning algorithms. Hence converting the text into vector or number format is a crucial step. CountVectorizer is one of basic word embedding model used to give a word vector form. It counts each and every occurrence of word present in the document and increases its count every time if it finds the same word again

Example:

Data= [‘The’, ‘cat’, ‘saw’, ‘the’, ‘mouse’, ‘the’, ‘mouse’, ‘ran’, ‘away’]

	the	cat	saw	mouse	ran	away
Data	3	1	1	2	1	1

Here, in the example the word “the” has occurred three times, each time the word “the” occurs the count increases. Hence, count for the word “the” is three. In the same way, word “mouse” has occurred twice and hence the count for the word “mouse” is two. Remaining word “cat”, “saw”, “ran” and “away” has occurred only once in the example data. Hence the count for these words will remain one.

This was an example to show how CountVectorizer algorithm is used to convert text into number or vector format.

B. SVM (Support Vector Machine)

Support Vector Machines employ the technique of Finding the decision boundary that maximizes the distance between two classes is the basic principle of SVM. This model enables us to classify the test data

(1). It gives the importance of each feature retrieved

(2). Gives support vector based on subject information to classify it to closer boundary using hyper planes

(3) It gives specific decisions not just binary classification. Radial basis function kernel is a type of kernel method which we have used here in project. Radial basis function kernel is a function whose value depends on the distance from the origin or from some point.

$$K(X_1, X_2) = exponent(-\gamma\|X_1 - X_2\|^2)$$

Where,

a. ||X1 — X2 || = Euclidean distance between X1 & X2

b. Gamma is used only for radial basis function kernel. As this value increases model gets over fit and if decreased model gets under fit.

C. Multinomial Naive Bayes

Naive Bayes is a supervised machine learning algorithm, where it is used for classification purpose. This works on Bayes rule where it calculates probability to classify the given input. The basic principle of Naive Bayes is that the features are independent to all other features in a given dataset, example a fruit may be considered to be an apple if its color is red, has round shape and diameter of three but all this features of the fruit is independent to one another. Bayes rules goes like this,

P (c/x) = p(x1/c)*p(x2/c)...p(xn/c).

Where,

a. P (c/x) is the posterior probability of class (c, target) given predictor (x, attributes).
b. P(c) is the prior probability of class.
c. P (x|c) is the likelihood which is the probability of predictor given class.
d. P (x) is the prior probability of predictor.

IV. PROPOSED METHODOLOGY

The dataset used for training the model is downloaded from kaggle where it has sentimental which is unique, phrases that is movie reviews as text and values as 1,2,3 and 4 where 1 specifies the phrase is negative, 2 as neutral and 3,4 as positive. It has 8500 movie reviews each being unique. The test data is tweets extracted from twitter account using twitter developer tools.

Firstly the reviews are preprocessed:

i. **Lower Case**: converting the phrase into lowercase.

ii. **Remove URLs:** If the phrase contains URL it is deleted completely from the text.

iii. **Stop words removal:** Stop words like at, the, is, at extra is removed from the phrase as it does not contribute to the sentiment.

iv. **Stemming:** When phrase has words repeated example “Its awesomeeeeee” the extra e’s will be removed and made into standard format as "awesome".

We use scikit learn library to perform all the above operations. Second step is converting the preprocessed phrase into vector format. We use CountVectorizer where it will count the presence of each word and increases the count gradually, if the word is present in the vocabulary. It is one of most common used word embedding model. It is very important step because the input to the model should be in number format. Vector size is set to 5000; finally the vector is converted into array. The below figure Fig.2 gives the complete flow diagram of proposed system.

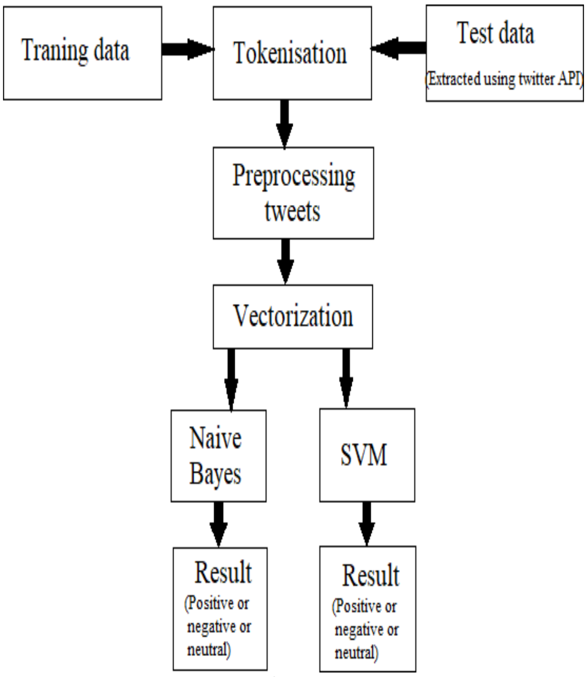

Traning data
Tokenisation
Test data
(Extracted using twitter API)
Preprocessing tweets
Vectorization
Naive Bayes
SVM
Result
(Positive or negative or neutral)
Result
(Positive or negative or neutral)

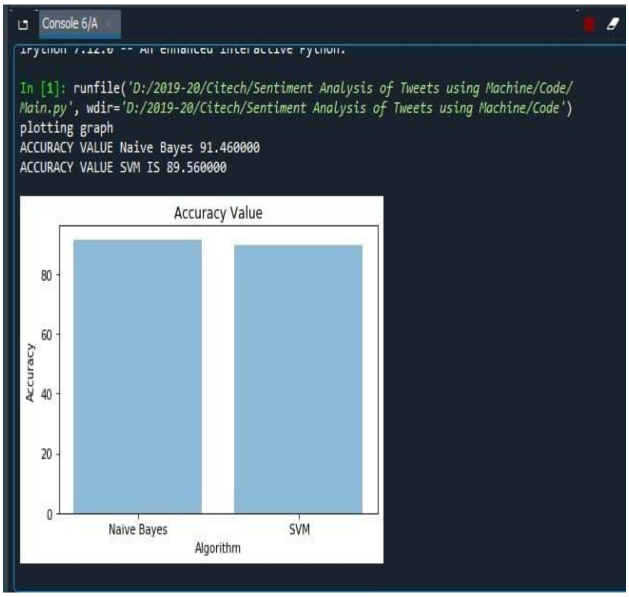
Fig.2. Flow diagram of Proposed System

V. ACCURACY

Accuracy is one of the metrics used to evaluate the developed model. It is the total no of correctly classified input to the total size of input given. In this paper we make use of Naive Bayes which gives accuracy of 91% and SVM (Support Vector machine) which gives accuracy of 89%.

VI. EXPERIMENTAL RESULTS AND EVALUATION

In the proposed work, we created two classification models one is Naive Bayes and another one is Super Vector Machine. A series of operations is run to evaluate the tweets that are test data extracted using twitter API to classify the sentiments. The below figure Fig.3 is a screenshot of the result.



```
IPython 7.12.0 -- An enhanced Interactive Python.

In [1]: runfile('D:/2019-20/Citech/Sentiment Analysis of Tweets using Machine/Code/
Main.py', wdir='D:/2019-20/Citech/Sentiment Analysis of Tweets using Machine/Code')
plotting graph
ACCURACY VALUE Naive Bayes 91.460000
ACCURACY VALUE SVM IS 89.560000
```

Fig.3. Experimental result

VII. CONCLUSION

This paper gives a clear view of two best classification algorithm Naive Bayes and SVM (Support vector machine) and also about their working principle. The word embedding model used was CountVectorizer. In First step we preprocessed the reviews and made into pure text form. In second step the text was converted into vector format using CountVectorizer with size of 5000 and the model was trained. Lastly the test data extracted from twitter account was fed into both models i.e. Naive Bayes and SVM.

Where the accuracy shown by Naive Bayes was 91% and that of SVM was 89%. It was seen that Navie Bayes model classifies the tweets into proper sentiments giving more accuracy than SVM, but the accuracy rate is dependent on the data set. If the data set varies or changes then accuracy rate of both Navie Bayes and SVM will also change according to data set.

ACKNOWLEDGEMENT

This work was able to be completed with the guidance of our guide and Hod of Computer Science and Engineering department Dr. Shashikumar D.R. We would like to thank him for his support and guidance.

REFERECS

[1] V.Lakshmi, K.Harika, H.Bavishya, Ch.SriHarsha, “SENTIMENT ANALYSIS OF TWITTER DATA," vol.04, February2017.Link "https://www.irjet.net/archives/V4/i3/IRJET-V4I3581.pdf” https://en.wikipedia.org/wiki/Twitter

[2] Using TF-IDF to Determine Word Relevance in Document Queries by Juan Ramos(Department of Computer Science, Rutgers University)

[3] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.

[4] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[5] [6]J. R. Quinlan, C4.5: Programs for Machine Learning,
Morgan Kaufmann Publishers, 1993

[6] Predicting Students Final GPA Using Decision Trees: A Case Study by Mashael A. Al-Barrak and Muna Al-Razgan

[7] https://www.lexalytics.com/lexablog/2014/sentiment- analysis-added-to- oxford-dictionaries

[8] A Short Introduction to Boosting by Yoav Freund and Robert E. Schapire (AT&T Labs, Research, Shannon Laboratory)

[9] A Hybrid Approach for SupervisedTwitter Sentiment Classification by K. Revathyand Dr. B. Sathiyabhama

[10] Using Objective Words in SentiWordNet to Improve Word-of-Mouth sentiment classification by Chihli hung and Hao-kai ling

[11] R. Parikh and M. Movassate, “Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques”