Twitter Sentiment Polarity Prediction Using Naïve Bayes Algorithm

[1]Y. Shivasree, [2]Manideep Chindham, [3]Modepalli Venkata Hrudai, [4]Mohammed Mubashir Ahmed

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

shivasreeyerrabati@tkrec.ac.in

[2]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
manideeppatel0007@gmail.com

[3]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
mvhrudai512@gmail.com

[4]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
mohammedmubashir204@gmail.com

Abstract: *Sentiment Analysis (SA) is the computational treatment of opinions, sentiments and subjectivity of text. Aspect based Sentiment Analysis (ABSA) is a specific SA that aims to extract most important aspects of an entity and predict the polarity of each aspect from the text. Aspect based sentiment analysis consists of aspect and sentiment extraction, and determination of the sentiment's orientation. In this project, we propose a system to extract the aspect sentiment pair and compute the rating for each grouped aspect. Our approach starts with selecting the subjective sentences in the reviews. Then, it extracts aspects and opinions from the sentences, and determine the orientation of the sentiment. Twitter is the popular micro blogging site where thousands of people exchange their thoughts daily in the form of tweets. The characteristics of tweet is to be short and simple way of expressions. though this thesis will focus on sentiment analysis of twitter data. The research area of sentiment analysis are text data mining and NLP. By using different supervised machine learning techniques, we will perform the sentiment analysis on twitter data. However, we will focus on techniques and types of sentiment analysis where we will perform how to extract tweets from twitter. Further we will compare different machine learning techniques on the same dataset and also find some standard measures.*

Keywords: *Sentiment analysis, label data, sentiment polarity, sentiment classification.*

I. INTRODUCTION

With the increasing popularity of social networking, blogging and micro-blogging websites, every day a huge amount of informal subjective text statements are made available online. The information captured from these texts, could be employed for scientific surveys from a social or political perspective. Companies and product owners who aim to ameliorate their products/services may strongly benefit from the rich feedback. On the other hand, customers could also learn about positivity or negativity of different features of products/services according to users' opinions, to make an educated purchase. Furthermore, applications like rating movies based on online movie reviews could not emerge without making use of these data. "Sentiment Analysis On Twitter Data" is increasing popularity of social networking and Sentiment Analysis (SA) is one of the most widely studied applications of Natural Language Processing (NLP) and Machine Learning (ML). This field has grown tremendously with the advent of the Web 2.0. The Internet has provided a platform for people to express their views, emotions and sentiments towards products, people and life in general. Thus, the Internet is now a vast resource of opinion rich textual data [1].

Nowadays twitter, Facebook, WhatsApp are getting so much attention from people and also, they are getting very much popular among people. Sentiment analysis provides many opportunities to develop a new application. in the industrial field, sentiment analysis has big effect, like government organization and big companies, their desire is to know about what people think about their product, their market value. the aim of sentiment analysis is to find out the mood, behaviour and opinion of person from texts. for the sentiment analysis purpose, social networking used the various sentiment analysis techniques to take the public data. Sentiment analysis widely used in various domain such as finance, economics, defence, politics. The data available on the social networking sites can be unstructured and structured. almost 80% data on the internet is unstructured. Sentiment analysis techniques are used to find out the people opinion on social media. Twitter is also a huge platform in that different idea, thought, opinion is presented and exchanged. It does not matter where people came from, what religious opinions they hold, rich or poor, educated or uneducated, they comment, compliment, discuss, argue, insist [2].

Sentiment Analysis

Sentiment analysis is a process of computationally identifying and categorizing opinions from piece of text, and determine whether the writer's attitude towards a particular topic or the product is positive, negative or neutral.

Instance suppose you want to buy a product. so, before purchasing a product. You look for the feedback like what the other customer has to say about that particular product whether it is good or bad and you analyse it manually by looking at their feedback. now consider at the company level how did the company analyse what their customer is thinking about their product. Generally, they do not have one or more customer. they do have millions of customers. So, what they will do. So here company needs to do sentiment analysis. To know whether their product is actually doing good in the market or not[3].

Sentiment Analysis Classification Based on the different perspective. Sentiment analysis has different variety of class. In which only one is used in sentiment classification techniques. This is classified into two other approaches i.e., machine learning approach and lexicon-based approach. We can add one more technique i.e., hybrid approach. There are three main classification level i.e., sentence level, document level, and last one is aspect level. Based on sentiment analysis, polarities can be classified into three classes such as positive neutral or negative.

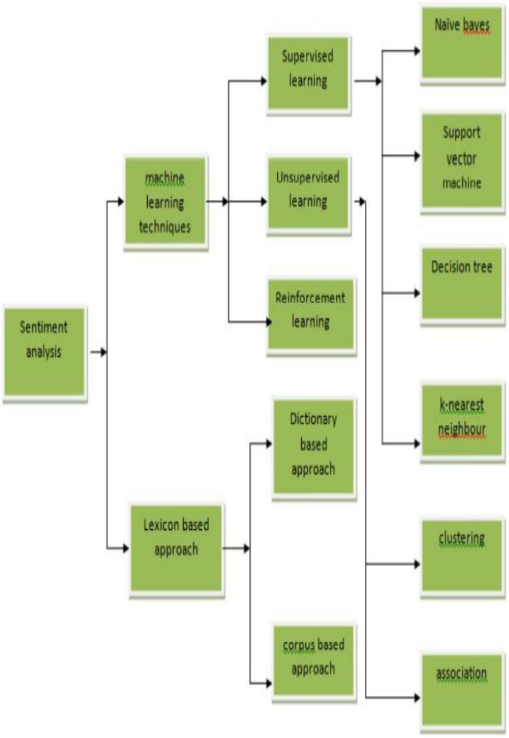


Fig.1 Sentiment Classification Techniques

Sentiment analysis is area where we can classify the various techniques. It is most popular area of research. It is notably classified into two types such as machine learning based approach and lexicon-based approach. Lexicon based approach basically focused on negative and positive term and it is further classified into two types i.e., dictionary based and corpus based. Moreover, machine learning approach is focused on two techniques namely supervised and unsupervised approach.

The goal of Sentiment Analysis is to harness this data in order to obtain important information regarding public opinion, that would help make smarter business decisions, political campaigns and better product consumption. Sentiment Analysis focuses on identifying whether a given piece of text is subjective or objective and if it is subjective, then whether it is negative or positive.

II. LITERATURE SURVEY

In this chapter, we will discuss different literature survey of different author. There are many researchers who has done great contribute in this area. In sentiment analysis, researchers have done various research on sentiment analysis by using different techniques. Here we are going to discuss some researches that will help us to know about the sentiment analysis in depth. Dhiraj gurkhe, Niraj pal and Rishit Bhatia discussed how twitter data is processed firstly they collected data from various sources and eliminate those features which does not contribute to find any polarity and then this data send into the sentiment classification engine i.e., naïve bayes classification algorithm which will calculate the probabilities i.e., how much data is corrected and predict the sentiment for the given query [4].

M.bouazizi, T. ohtsuki have discussed the tweets which contain more than one sentiment called as multi class sentiment analysis. Where they have identified the exact sentiment conveyed by the user rather than the whole sentiment of the tweet. To identify this thing, they have also used SENTA tool. They proposed an approach, with the help of this approach they have calculated the sentiment score whoever sentiment is having highest score that will be considered this process is called as “Quantification” [5].

Geetika gautam, Divakar Yadav have discussed about customer review classification for which they have used twitter dataset which is already labeled. In this task they have used machine learning based algorithm i.e., naïve bayes, SVM, maximum entropy. They have worked on Python and NLTK for training the SVM, naïve bayes, maximum entropy. Naïve bayes is better techniques in term of accuracy and gives the better result compare to Maximum entropy. We can get the better result with compare to SVM by using the SVM with unigram model. And then further accuracy can be improved by semantic analytic followed by wordNet.[6].

Akshay Amolik and m.venkatesan, a highly suitable model have discussed int his paper which will take the twitter data of upcoming Hollywood and Bollywood movies. They are able to this task with the help of classifier and features like SVM and naïve bayes. Both of them are used for high accuracy but in terms of precision naïve bayes is better than SVM and if we talk about recall then SVM is better than naïve bayes. By increasing the dataset, we can increase the classification accuracy[7].

Subhabrata Mukherjee, Akshat Malu, Balamurali A.R, Pushpak Bhattacharya have discussed a hybrid system named as TwiSent which will resolve problem like spam tweet, pragmatics, noisy text. Twisent consist of spell checker and pragmatics handler. spell checker finds the noisy text whereas pragmatics handler handles the pragmatics in tweets. Twisent gives better result compare to C-feel-IT system. The accuracy of finding the negative sentiment of TwiSent system is high the C-Feel-IT[8].

Dmitry Davidov, Oren Tsur, Ari Rappoport in this paper they have proposed a supervised sentiment classification framework which is based on twitter data. They have used K nearest neighbor and feature vector. the basic purpose of this framework is to identify and distinguish between sentiment types defined by smiley and tags[9].

Neethu M S, Rajasree R, the author has used the machine learning techniques in this survey paper to explore the twitter data related to electronic product. They have used feature vector for the tweet's classification. they have used three types of classifiers i.e., SVM, naïve bayes, maximum entropy, and these classifiers were tested using Matlab simulator. SVM and naïve bayes classifier are implemented using built in function. Whereas MaxEnt classifier is used by MaxEnt software. So basically, the whole classifier have nearly the same performance[10].

Pulkit et al. built and proposed a model which extract tweet from twitter based on the post terror activities. they made their study on terrorist attack which was occurred in uri on 18 september 2016. They considered 59,988 tweet which had taken after the attack. They consider only those tweets which has #UriAttack, #uriattack. #Uriattacks. They have used the naïve bayes and SVM to extract the last re-tweet time and number of re-tweet 20 Sudarshan Sirsat et al. proposed a technique in sentiment analysis on twitter data where they have collected reviews of the product. They have used naïve Bayes algorithm which perform better in term of accuracy and efficiency. They have extracted 200 tweets where the average length of tweet was 70.105. the aim of this research is to identify the characteristic of tweet like how many times the tweet was liked and how many times they have re-tweet the tweet[11].

Hetu et al. proposed a model in sentiment analysis on twitter data based on anaconda python. They extract the dataset from kaggle in which they classify the people emotions based on positive and negative reviews. This model gives high accuracy on large dataset. Ali hasan et al. proposed a model using the hybrid approach that comprise sentiment analyzer machine learning. They took only those tweet that is followed by the hashtag (#) and contain the current political trends. Basically, this model converts the urdu tweet into English tweet. They took 1690 tweet for training data and 400 for testing the data. They have used the naïve bayes and SVM classifier for training the dataset in weka and building a model. They have used three different libraries to calculate the subjectivity and polarity. Feddah AlhumaidiAl Otaibi et al. proposed a model by using the supervised and unsupervised algorithm [12].

III. PROPOSED WORK

Here we have used a Naïve Bayes classifier for sentiment classification. The user review is analyzed and each feature wise score rating of the product is determined. The reviews are preprocessed to eliminate noises using various tools and methods such as stop word removal, stemming, etc. The extracted words are classified into positive and negative through unigrams using naive Bayesian classifier. We are planning to improve the User Interface and further improve the efficiency and address humor-based analogy in reviews for better understanding the sentiment.

PROPOSED MACHINE LEAERNING TECHNIQUES

To classify the text classification problem in sentiment analysis, machine learning is used. In this to train a model, training data records is used which later used to identify the predict model without level. Each and every record is labelled into different classes. When we give new unlabelled record to model, then the model will label that dataset into different classes. There are three types of different classes such as positive, negative and neutral. Generally neutral class is mixed opinion. Rarely we consider the neutral class. Eventually machine learning techniques is of three types i.e., supervised learning techniques, unsupervised learning techniques and reinforcement techniques.

Naïve Bayes:

Naïve bayes theorem is a classification method with the independent assumption between the predictors. In other words, the approach of particular predictor of one class is not connected to closeness of some other class. Naïve bayes is a “probabilistic classifier”. Let’s take an instance, an apple may be considered a fruit if it is red in colour, and if it is round in shape and if its diameter is considered to be three inches approximately. Despite of these feature are dependent on one another or in the presence of another feature. All these independent properties contribute to find the probability of naïve classifier that this is an apple. Naïve bayes is beneficial for big data sets and can be built easily. Let us consider a class variable ‘y’ and a dependent vector from x1 to xn. So according to naïve bayes

$$P(y \mid x_1,\ldots,x_n) = \frac{P(y)P(x_1,\ldots x_n \mid y)}{P(x_1,\ldots,x_n)}$$

So according to mutually independent assumption

$$P(x_i|y,x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = P(x_i|y),$$

For each value of i this function behaves

$$P(y \mid x_1,\ldots,x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1,\ldots,x_n)}$$

NATURAL LANGUAGE TOOLKIT (NLTK):

The Python programming language provides a wide range of tools and libraries for attacking specific NLP tasks. Many of these are found in the Natural Language Toolkit, or NLTK, an open-source collection of libraries, programs, and education resources for building NLP programs. The NLTK includes libraries for many of the NLP tasks listed above, plus libraries for subtasks, such as sentence parsing, word segmentation, stemming and lemmatization (methods of trimming words down to their roots), and tokenization (for breaking phrases, sentences, paragraphs and passages into tokens that help the computer better understand the text). It also includes libraries for implementing capabilities such as semantic reasoning, the ability to reach logical conclusions based on facts extracted from text.

K-Nearest Neighbour (KNN)

K-NN is very simple algorithm that stores all the available cases and classifies the new data or case based on similarity measure. it uses the entire dataset in its training phase. Figure 1.3 recommender system using KNN algorithm 8 For instance if apple look most similar to banana, orange, melon rather than a monkey, dog or cat most likely apple belong to the group of fruits. In general k-nn is used in search application where you are looking for the similar item. In KNN, k denotes the number of nearest neighbours which are holding class of the new data or the testing data. KNN is used at the industries level. Figure 3 shows concept search using KNN algorithm the biggest use case of KNN search is recommended system. recommended system is an automated form of shop counter guy. When you will ask for the product it not only show to you the relevant product but also suggest you or recommend you the product related to your relevant product that you want to buy. KNN algorithm applying to recommending product like in amazon and for recommending media in Netflix.

amazon
Recommender System
Industrial Application of KNN Algorithm
9:41
Apple iPhone 8 (Space Grey, 64GB)

Fig.2 K-Nearest Neighbour

Natural Language Processing:

Natural language processing deals with techniques that can analyse, compute, and represent the data at various level analysis of languages for the purpose to make the machine process like human language for different disciplines and applications. NLP algorithm highly depends on machine learning techniques with the major being numerical. there are various types of natural language processing namely

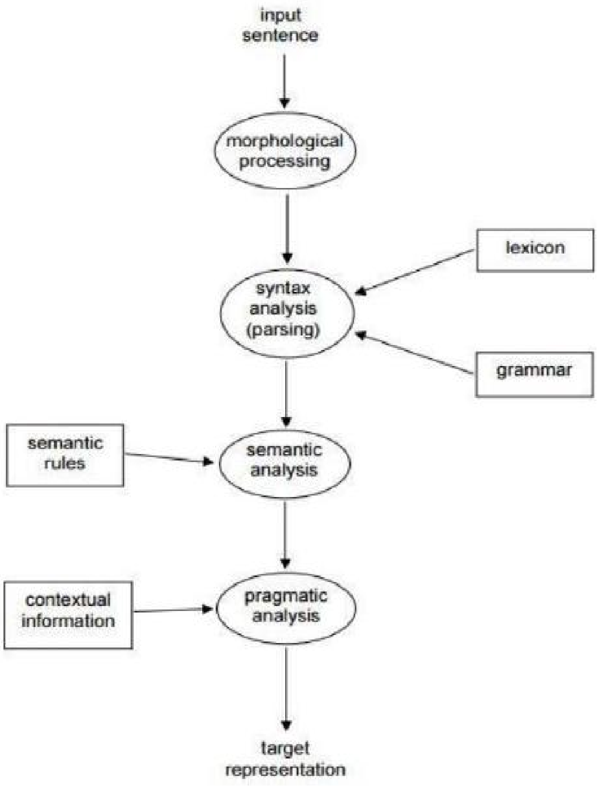


Fig.3 Steps of Natural Language Processing.

IV. RESULTS

ISSN: 2366-1313

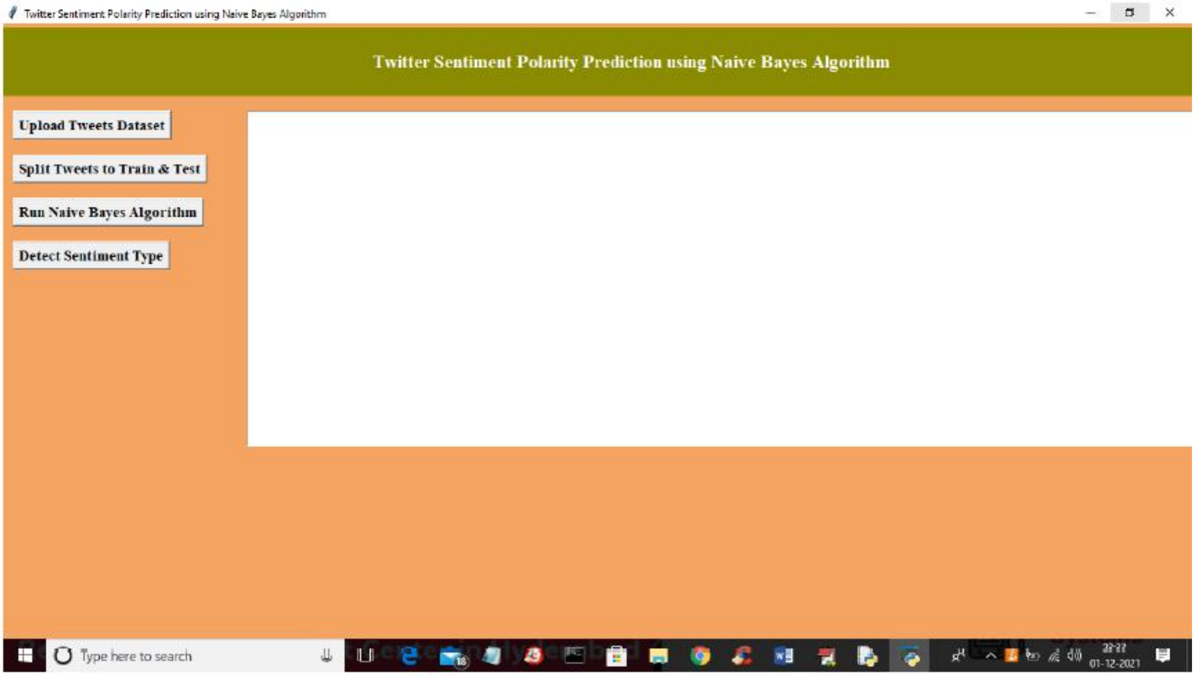
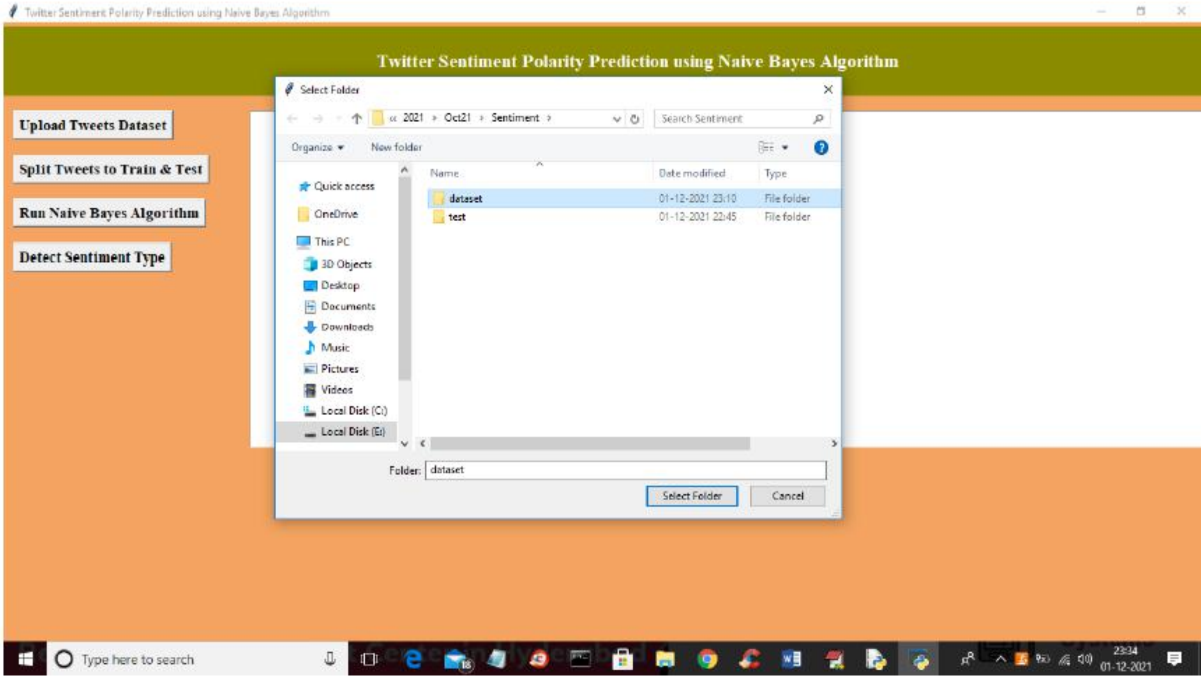
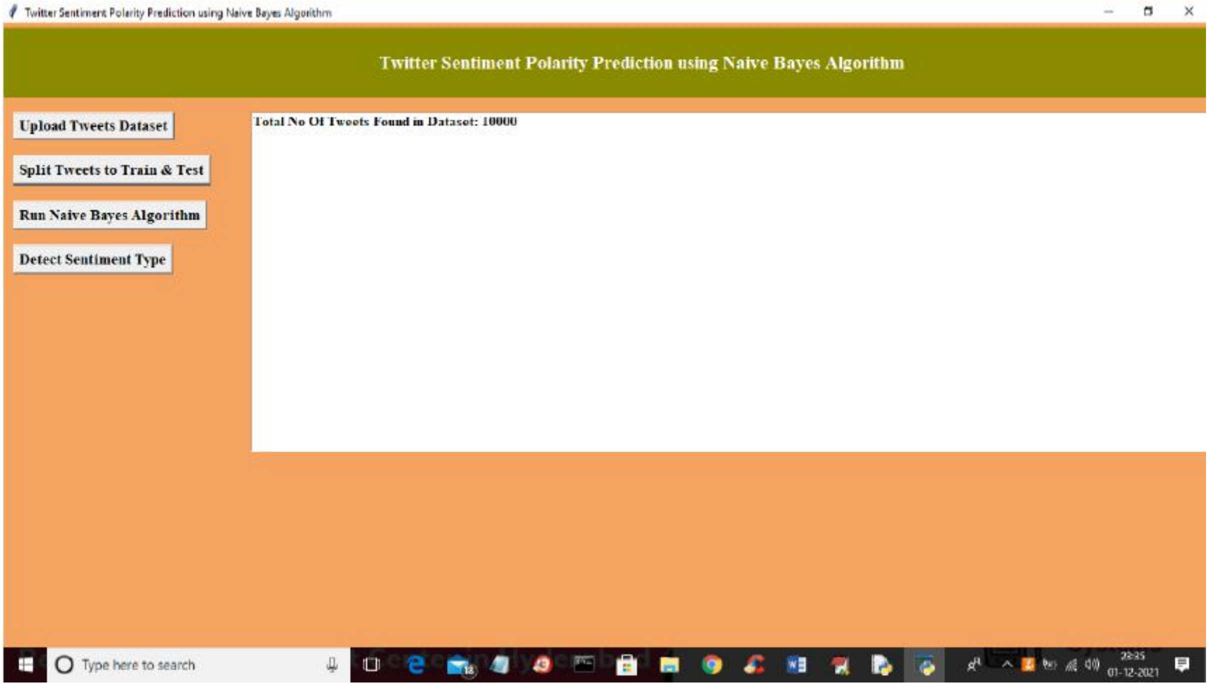

Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Upload Tweets Dataset
Split Tweets to Train & Test
Run Naive Bayes Algorithm
Detect Sentiment Type
Type here to search
22:22
01-12-2021

Fig.4 upload tweet dataset



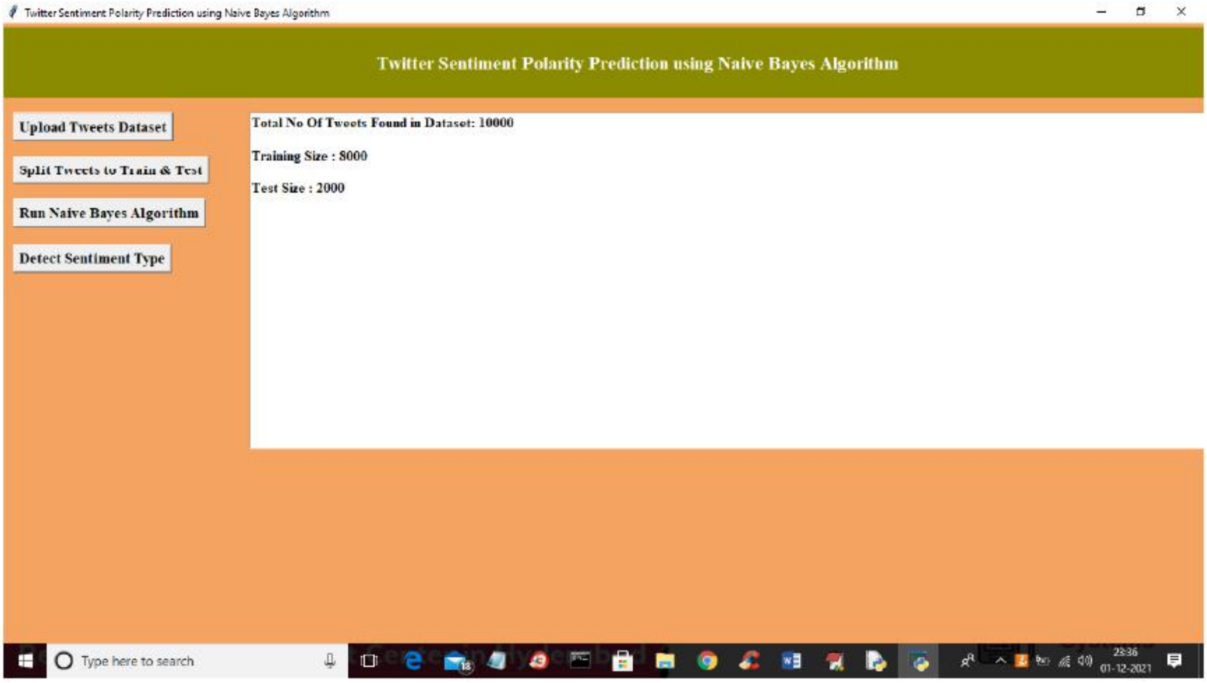
Folder: dataset

Fig.5 In above screen selecting and uploading ‘dataset’ folder which contains positive and negative tweets dataset and after loading will get below screen


Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Upload Tweets Dataset
Split Tweets to Train & Test
Run Naive Bayes Algorithm
Detect Sentiment Type
Type here to search

Total No Of Tweets Found in Dataset: 10000

Fig.6 In above screen tweets loaded and dataset contains 10000 tweets and now click on 'Split Tweets to Train & Test' button to split dataset into train and test.

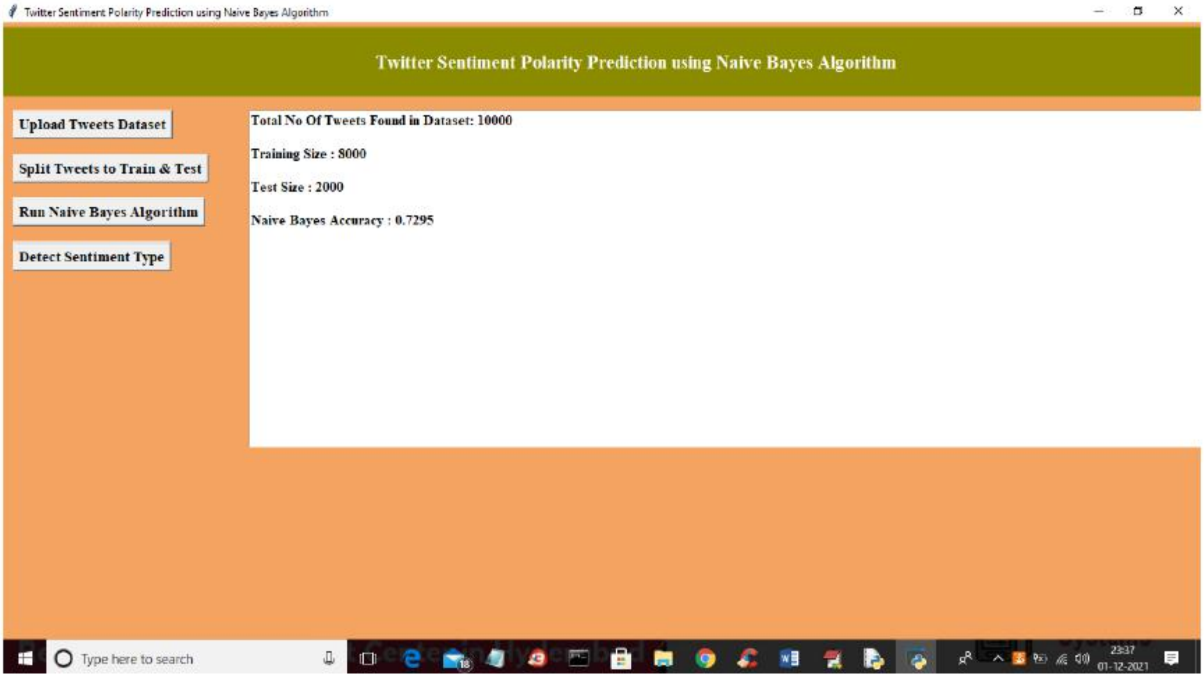

Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Upload Tweets Dataset
Split Tweets to Train & Test
Run Naive Bayes Algorithm
Detect Sentiment Type
Type here to search

Total No Of Tweets Found in Dataset: 10000

Training Size : 8000

Test Size : 2000

Fig.7 In above screen to train naïve bayes application using 8000 tweets and then 2000 tweets are using as test to calculate prediction accuracy. Now train and test data is ready and now click on 'Run Naïve Bayes Algorithm' button to train Naïve Bayes with above dataset and to get below accuracy.


Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Upload Tweets Dataset
Split Tweets to Train & Test
Run Naive Bayes Algorithm
Detect Sentiment Type
Type here to search
23:37
01-12-2021

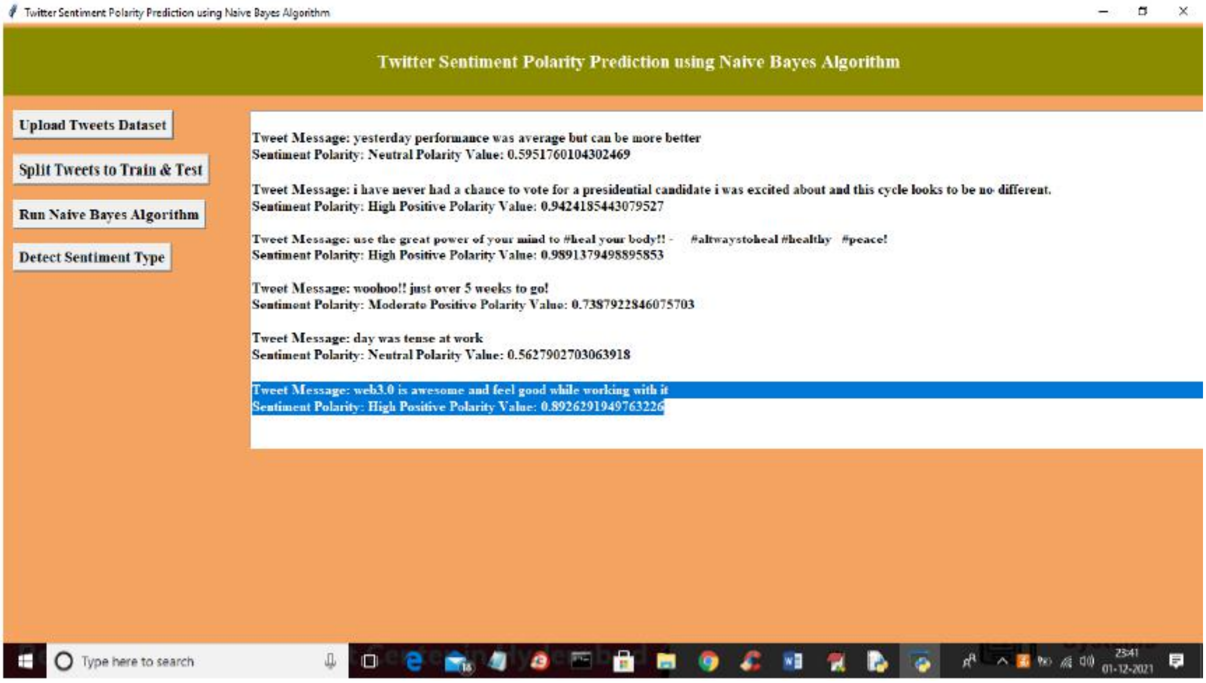
```
Total No Of Tweets Found in Dataset: 10000

Training Size : 8000

Test Size : 2000

Naive Bayes Accuracy : 0.7295
```

Fig.8 In above screen I am selecting and uploading ‘test.txt’ tweets test file and then click on ‘Open’ button to load test tweets and to get below prediction result and if u want you can add new tweets in that file


Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Twitter Sentiment Polarity Prediction using Naive Bayes Algorithm
Upload Tweets Dataset
Split Tweets to Train & Test
Run Naive Bayes Algorithm
Detect Sentiment Type
Type here to search
23:41
01-12-2021

```
Tweet Message: yesterday performance was average but can be more better
Sentiment Polarity: Neutral Polarity Value: 0.5951760104302469

Tweet Message: i have never had a chance to vote for a presidential candidate i was excited about and this cycle looks to be no different.
Sentiment Polarity: High Positive Polarity Value: 0.9424185443079527

Tweet Message: use the great power of your mind to #heal your body!! -    #altwaystoheal #healthy   #peace!
Sentiment Polarity: High Positive Polarity Value: 0.9891379498895853

Tweet Message: woohoo!! just over 5 weeks to go!
Sentiment Polarity: Moderate Positive Polarity Value: 0.7387922846075703

Tweet Message: day was tense at work
Sentiment Polarity: Neutral Polarity Value: 0.5627902703063918

Tweet Message: web3.0 is awesome and feel good while working with it
Sentiment Polarity: High Positive Polarity Value: 0.8926291949763226
```

Fig.9 Final tweet message

V. CONCLUSION

Sentiment Analysis is one of the most widely used applications of NLP. It leverages the data existing in vast amounts on public platforms. And provides useful insights to businesses helping them improve their services, and in turn increase customer satisfaction. Aspect based sentiment analysis is a step ahead of the conventional sentiment analysis. Models like BERT have proven to be very effective for sentiment analysis. However, practical application has taught me that simpler models can give a decent performance with much lesser compute requirements and training time. Moreover, when working with languages other than English, it is quite hard to find good pre-trained models! In today's world, spacious amount of data is generated by various communication such as social media, organizations etc. these data may or may not be in structured form. Therefore to understand the polarity of data first we need to do the sentiment analysis of data. Opinion mining can be performed in various field such as marketing and customer feedback. large number of organizations are taking the valuable feedback of person and performing opinion mining on those data so that they could provide the better services to the customer and this data helps the organizations to enhance their future services. Furthermore, there are various scopes where we can perform the opinion mining such as sentence, paragraph, documents, sub sentences levels.

REFERENCES

[1] Gurkhe D., Pal N. and Rishit B. "Effective Sentiment Analysis of Social Media Datasets using Naïve Bayesian Classification." (2014).

[2] Bouazizi, M., Ohtsuki, T.: Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer. IEEE Access. 6, 64486-64502 (2018).

[3] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). (2014).

[4] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016): 1- 7.

[5] Mukherjee S., Malu A., Balamurali A.R, Bhattacharyya P."TwiSent: A Multistage System for Analyzing Sentiment in Twitter".

[6] Davidov D., Tsur O., Rappoport A." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys".

[7] Neethu, M., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). (2013).

[8] Pulkit Garg, Himanshu Garg, VirenderRanga "Sentiment Analysis of the Uri Terror Attack UsingTwitter" International Conference on Computing, Communication and Automation (ICCCA2017). [9] Prof. SudarshanSirsat, Dr.Sujata Rao, Dr.BhartiWukkadada"Sentiment Analysis on Twitter Data forproduct evaluation" IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719PP 22-25.(2019)

[10] Hetu Bhavsar, Richa Manglani" Sentiment Analysis of Twitter Data using Python"International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN: 2395-0056 p-ISSN: 2395-0072 49.

[11] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management, Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/ 10.1109/MDM.2013. [12] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175