SENTIMENT AND EMOTIONAL ANALYSIS OF USER OPINIONS ON TWITTER DATA USING MACHINE LEARNING TECHNIQUES

[1]S.Joshua Johnson,[2]B.V.D.S.Sailendra,[3]K.Katyayini,[4]B.Pooja,[5]G.V.Sai Prasad

[1]Assistant Professor, [2,3,4,5] Student

Department Of Computer Science & Engineering,

Anil Neerukonda Institute Of Technology and Sciences, Visakhapatnam, India.

ABSTRACT

Technology today has become a momentous driving vehicle for communication world-wide.Social media platforms like twitter,facebook, instagram are the most important arenas for expressing views on transformations happening in and around the world everyday.Twitter is a rich origin of info for mining of user opinions.This paper reflects the idea of taking user opinions into consideration performing sentiment, emotion analysis and establishing conclusions on interested topics using Machine Learning algorithms.Naive Bayes and Support Vectors Machines in Machine Learning are tuned-up using supervised learning to obtain outputs for sentiment emotion analysis respectively.Sentiment analysis desires to obtain sentiment polarity (positive or negative) and Emotion analysis intent to obtain emotion (eg., empty , sadness , anger..etc) from user data. Such analysis essentially serves a gateway for consumer needs and generates growth opportunities in businesses.

Keywords : Sentiment Analysis , Emotion analysis, Naive Bayes , Support Vector Machines, Preprocessing , Tokenization, Stemming ,Lemmatization , Twitter.

1.INTRODUCTION

Social media sentiment analysis has turn out to be a distinguished area of study and experimentation in current years..Twitter a microblogging site, has lion's share in social media info. Most research has been confined to classify tweets into positive,negative categories .Human emotions are extremely diverse and cannot be restricted to certain metrics alone.Polarity analysis gives limited information on the actual intent of message delivered by author and just positive or negative classes are not sufficient to understand nuances of underlying tone of a sentence.This brings the need to take one step above sentiment analysis leading to emotion analysis. In this paper we throw light on methods we have used to derive sentiment analysis and how we have accomplished emotion analysis of user opinions.A supervised learning technique provides labels to classifier to make it understand the insights among various features.Once the classifier gets familiarized with train data it can perform classification on unseen test data.We have chosen Naive Bayes and Support Vector Machine classification algorithms to carry out sentiment and emotional analysis respectively.

Performing SA(sentiment analysis) and EA (emotion analysis) will help organizations or companies to improve services , track products and obtain customer feedback in a normalized form.Gaining insights from large volumes of data is a mountain of a task for humans hence using an automated process will easily drill down into different customer feedback segments mentioned on social media or elsewhere .Effective business strategies can be built from results of sentiment and emotion analysis.Identifying clear emotions will establish a transparent meaning of text which potentially develops customer relationships ,motivation and extends consumer expectations towards a brand or service.

Emotion detection involves a wide platter of emotions classified into states like joy,fear,anger,surprise and many more.We here examine sentiments and emotions of short texts coined as tweets from the famous social media, twitter.

2.PREVIOUS WORK

Traditional methods for identifying sentiments have been carried out ignoring major elements like emoticons and symbols.Ample amount of research has been done to accomplish sentiment analysis and emotion analysis using machine learning and Natural language processing techniques.Various methods for sentiment analysis have evolved over time and few of which are mentioned in [1] by Shiv Dhar, Suyog Pednekar, Kishan Borad, Ashwini Save.Author conveys various approaches and algorithms such as lexicon based approach,SVM,convolution neural network and IML algorithm help in sentiment classification of text.

[1](Methods for Sentiment Analysis) conveys that two good machine learning algorithms to implement sentiment analysis are SVM (Support Vector Machines) and Naive Bayes.1.Sentiment Analysis on Twitter using streaming API: It uses NLP where it helps in tokenization, stemming, classification, tagging, parsing and sentiment reasoning Its basic feature is to convert unstructured data into structured data. It makes use of Naive Bayes for classification which feels necessity for number of linear parameters.SVM is a machine that intakes the input and contain them as a vector then using SentiWordNet it scores then it decides the sentiment. It also segregates the opinion into one among the three classes positive, negative or Neutral.

Furthermore author in [2](the impact of feature extraction on sentiment analysis) specifies feature extraction techniques such as bag-of-words,TF-IDF,NLP and he analysed the importance of extracting features and how they play a crucial role in the performance of the classifier and its outcome.

Here he analysed the impact of two features TF-IDF word level and N-Gram on a Twitter data set and found that performance of classifier is 3-4% in high when TF-IDF feature is chosen than N-Gram and analysis is done using many classification algorithms like Naïve Bayes, Support Vector Machines etc..Eventually he concludes that the feature selection process affects the sentiment evaluation result.

In [3] (Analyzing Sentiment of Twitter Data using MachineLearning Algorithm)a procedure has been proposed for sentiment analysis saying tweets can be extracted from an API called tweepy.Then perform preprocessing by removing stop words,URLs and punctuations.Stages of procedure include data collection,pre-processing setup,sentiment classifier and evaluation result.And each step is further classified like in the pre-processing step

some sub-steps like stemming, stop word extractor are also included. The efficiency of a classifier mostly rely upon the pre-processing step.

For emotion analysis [4](Using a Heterogeneous Dataset for Emotion Analysis in Text) a supervised machine learning algorithm (SVM) had been used Different feature sets such as bag-of-words,N-Grams were used.Five datasets are taken into account to correlate among diverse approaches.The Support Vector Machines classifier (SVM) outperformed other classifiers, and it generalized well on unseen examples.

The author concludes some research studies in sentiment analysis claimed that N-grams features improve performance beyond the BOW approach.

[5](Multiclass Emotional Analysis on Social Media Posts) conclusions have been drawn that SVM has outperformed Naive Bayes and Softmax regression .This paper briefs SVM algorithm for supervised learning.

[6](Emotion Detection from Text) Here a framework was proposed which is divided into two components namely: Emotion ontology and emotion detector.In emotion ontology weightage is assigned to each emotion class based on emotion word hierarchy.The emotion detector algorithm is responsible for estimating weight for a particular emotion by adding weights at each level of emotion hierarchy and does the same for its counter emotion, then compares the both scores and greater one is taken as the detected emotion.The foremost step is evaluation of parameters. Different parameters include Parent-child relationship if a text document belongs to a child it also indirectly refers to the parent of it.Another parameter depth in ontology gives an idea about how specific is the term in relation to its corresponding ontology structure.The proposed algorithm calculate the score for each emotion word with the help of parameters from previous steps. Score will be calculated for every primary level emotion class. Finally Emotion class having the highest score will win the race and be declared as Emotion state of the corresponding text document.

3.PROPOSED SYSTEM

3.1 Extraction Of Data

Tweets based on a keyword of user's choice of interest have been collected using a famous twitter API known as Tweepy and stored into a csv file.This data set collected for sentiment analysis have tweets based on a keyword e.g.,cybertruck. Tweets mimicking various emotions as a dataset downloaded from kaggle is used for emotional analysis.Since both the machines are trained using supervised learning and work on different parameters different data sets have been considered.

3.2 Preprocessing Of Data

Following are the Preprocessing steps that have been carried out:

3.2.1 Removing Html tags and urls:

Html tags and urls often have minimum sentiments thus they are removed from tweets.Using regular expressions.

3.2.2 Conversion to lowercase:

To maintain uniformity all the tweets are converted to lowercase .This will benefit to avert inconsistency in data.Python provides a function called lower() to convert sentences to lower case.

3.2.3 Tokenization:

In this step a sentence is deconstructed as a collection of words each word,letter or a symbol

is recognized as a token having its own significance.As an illustration it be viewed as

Text :" I love Icecream"

Tokens: "I","love","Icecream"

3.2.4 Removing punctuations and special symbols:

Apart from the considered set of emoticons punctuations and symbols like &,\,; are removed.

3.2.5 Stop words removal:

Sentences often contain fillers and prepositions such as and,this,that which do not reflect any meaning .Suchlike words are called stop words.Since they do not add value to deep meanings thence can be avoided.NLTK library provides arbitrary methods to counterattack the stop words.

Before stop words filtration : ["The",`` weather", `` is", `` nice", `` today"]

After stop words filtration :["weather","today"]

3.2.6 Stemming and lemmatization:

Sentences are always narrated in tenses,singular and plural forms making most words accompany with -ing,-ed,es and ies. Therefore,extracting the root word will suffice to identify sentiment behind the text.

Base forms are the skeleton for grammar stemming and lemmatization reduces inflectional forms and derivational forms to common base forms .

Example: Cats is reduced to cat ,ponies is reduced to poni.

This task has been implemented using Snowball Stemmer and WordNet lemmatizer in NLTK

3.2.7 Feature Extraction:

Text data demands a special measure before you train the model.Words after tokenization are encoded as integers or floating point values for feeding input to machine learning algorithm. This practice is described as vectorization or feature extraction. Scikit-learn library offers TF-IDF vectorizer to convert text to word frequency vectors.

3.3 Fitting Data to Classifier and predicting test data:

Train data is fitted to a suitable classifier upon feature extraction ,then once the classifier is trained enough then we predict the results of the test data using the classifier,then compare the original value to the value returned by the classifier.

3.4 Result Analysis:

Here the accuracy of different classifiers are shown among which the best classifier with highest accuracy percent is the chosen. Some factors such as f-score,mean,variance etc., also accounts for consideration of the classifiers.

3.5 Visual Representation:

Our final results are plotted as pie charts which contains different fields such as positive,negative,neutral in case of sentiment analysis.where as happy,sad,joy etc., in case of emotional analysis. Pictorial representation is the best way to convey information without much efforts.Thus it is chosen.

4.ARCHITECTURE DIAGRAM

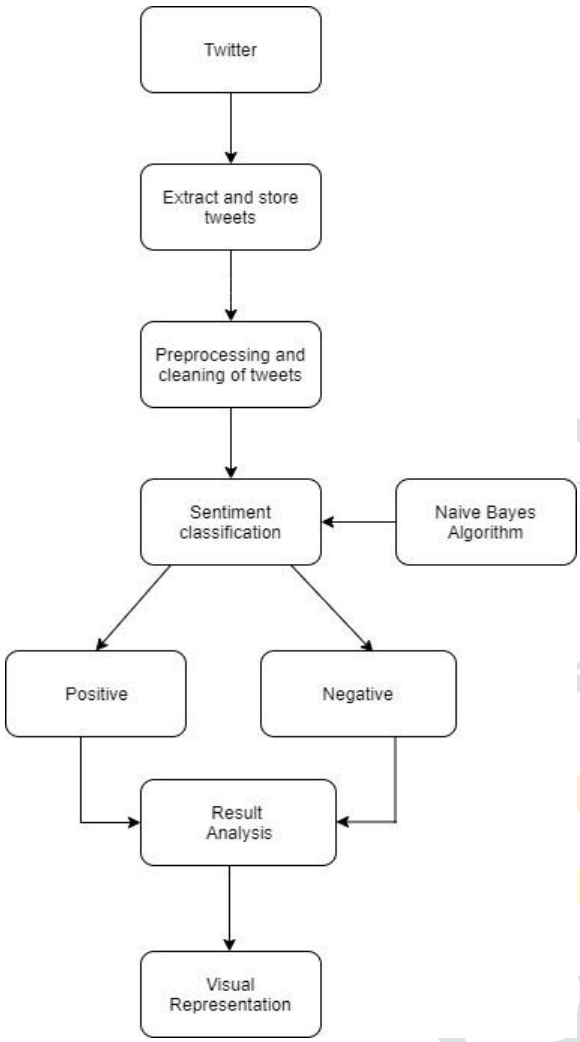

Fig 4.1 Architecture diagram for sentiment analysis using Naive Bayes

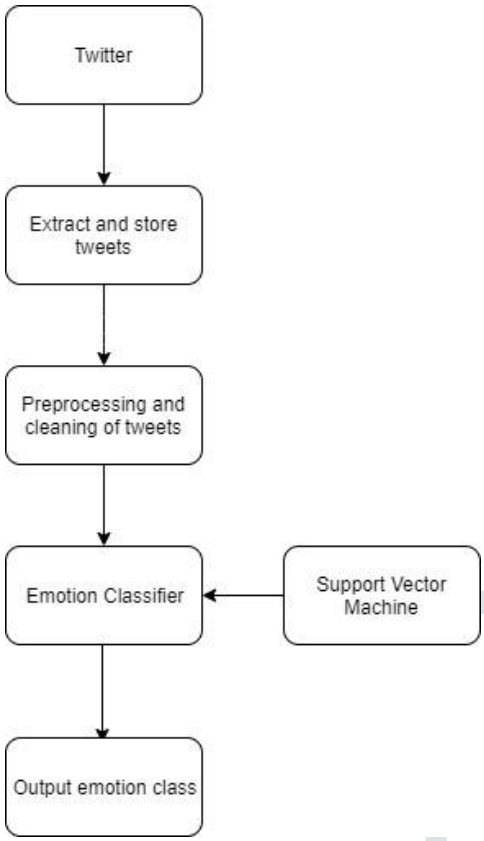

Fig 4.2 Architecture diagram for Emotion analysis using Support Vector Machines

Naive Bayes Algorithm :

Naive Bayes algorithm which is based on well known Bayes theorem which is mathematically represented as

$$P(A/B)=\frac{P(B/A)P(A)}{P(B)}$$

Where ,

A and B are events

P(A/B) is the likeliness of happening of event A given that event B is true and has happened,Which is known to be as posterior probability .

P(A) is the likeliness of happening of an event A being true,Which is known to be as prior probability.

P(B/A) is the likeliness of happening of an event B given A was true ,Which is known to be as Likelihood.

P(B) is the likeliness of happening of an event B, Which is known to be as Evidence .

Bayes theorem can now be applied on data sets in following way

$$P(y/X)=\frac{P(X/y)\,P(y)}{P(X)}$$

(Reference from Tom Mitchell)

Where y is a class variable and X is feature vector

Support Vector Machine:

Support Vector Machines is a supervised machine learning algorithm , adopted conventionally for classification as well as regression problems.SVMs for classification, work by figuring out the right hyperplane among the classes. After being trained by a labeled data set , SVM outputs an optimal hyperplane that categorizes new examples. Classification by SVMs for different data sets is governed by tuning parameters namely kernel , regularization, gamma and margin.When data is 2 dimensional Support vector classifier is a line, if it is 3 D SVC forms a plane instead of a line. When data is more than 4D then classifier is a hyperplane.For highly distributed data Maximal margin and support vector classifier fail and hence SVMs are used.For linearly separable patterns optimal hyperplane is formed and for non-linearly separable patterns transformation of original data into a new space is performed determined by kernel function.The trouble of discovering an optimal hyperplane is an optimization problem and can be worked out using optimization techniques (eg. Lagrange).To classify tweets into different emotion classes a linear kernel has been utilized.Linear kernel is preferable for text classification problems because text has lot of features,linear kernel is faster and less parameters are optimized . When SVM is trained with a linear kernel only C regularization parameter need to be optimized whereas for other kernels you need to optimize gamma parameter also.

5.INPUT

ItemId	Sentiment	SentimentText
1	0	@elonmusk The generation of graphics before #cybertruck
2	1	I received an awesome surprise #Cybertruck gift package today for joining the pre-order/reservation club!! Thanksâ€¦ https://t.co/dVo8YbFoQs
3	0	@Tesla did you forget about the most comfortable driving position that truck owners will want? Yup. #cybertruck https://t.co/creyXJ9EoQ
4	0	@elonmusk I'm very curious as to how the #cybertruck or any electric vehicles will do over a long term with daily mâ€¦ https://t.co/7NEB2W8V57
5	0	The New Tesla Cybertruck is Dang Easy to Model in Blender https://t.co/jIoZSZRQj7 @Tesla #Cybertruck #Tesla #futuretechnology

Fig 5.1 Train dataset for sentiment analysis

tweet_id	sentiment	author	content
1.957E+09	empty	xoshayzers	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[
1.957E+09	sadness	wannamam	Layin n bed with a headache ughhhh...waitin on your call...
1.957E+09	sadness	coolfunky	Funeral ceremony...gloomy friday...
1.957E+09	enthusiasm	czareaquinc	wants to hang out with friends SOON!
1.957E+09	neutral	xkilljoyx	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.
1.957E+09	worry	xxxPEACHES	Re-pinging @ghostridah14: why didn't you go to prom? BC my bf didn't like my friends
1.957E+09	sadness	ShansBee	I should be sleep, but im not! thinking about an old friend who I want. but he's married now. damn, & he wants me 2! scandalous!

Fig 5.2 Train dataset for emotional analysis

6.OUTPUT

Below are the results for sentiment and emotional analysis represented as a pie-chart for users using matplotlib.

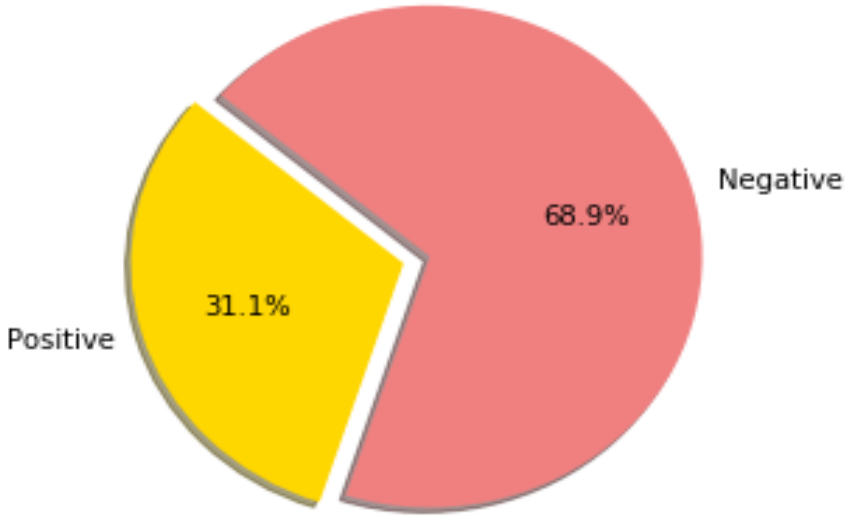

Fig 6.1 Result for Sentiment Analysis of Twitter Data

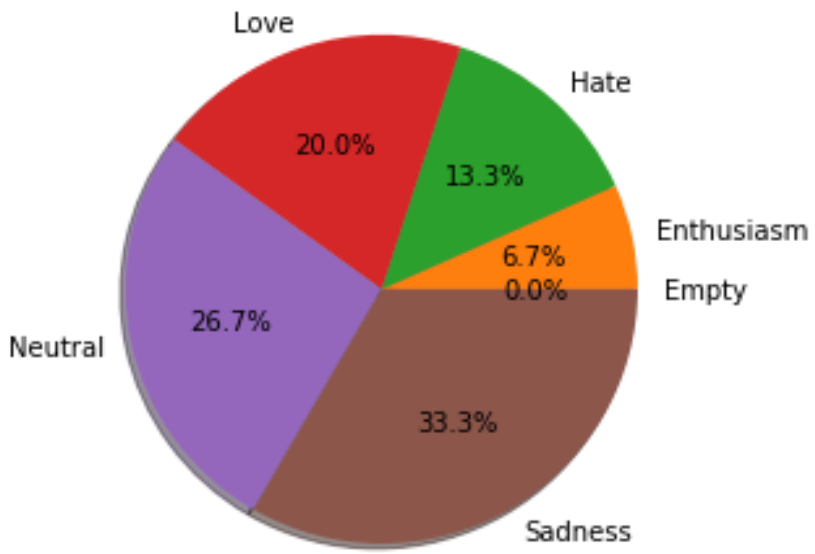

Fig 6.2 Result for Emotional Analysis of Twitter Data

7.RESULT ANALYSIS

```
=== logitic regression ===
scores =  [0.80959922 0.81016909 0.80595209 0.81198447 0.80780817 0.81363307
 0.81083696 0.81422158 0.81576402 0.81074569]
mean =  0.8110714362289213
variance =  7.933866749593147e-06
score on the learning data (accuracy) =  0.8718567836324151

=== bernoulliNB ===
scores =  [0.79016702 0.78374591 0.78744018 0.78276063 0.7815876  0.78967136
 0.79251899 0.79093912 0.7961597  0.79197936]
mean =  0.7886969879731057
variance =  2.0108782575180654e-05
score on the learning data (accuracy) =  0.9029031889358784

=== multinomialNB ===
scores =  [0.80779944 0.80702918 0.80614631 0.80742148 0.80748963 0.80853918
 0.80896121 0.81008703 0.8120753  0.81028724]
mean =  0.8085836007084021
variance =  2.901807591884587e-06
score on the learning data (accuracy) =  0.899074179906275
```

Fig 7.1 Scores generated using different algorithms at test-split=0.8

```
Out[71]: Text(0.5, 0, 'test-size percent')
```

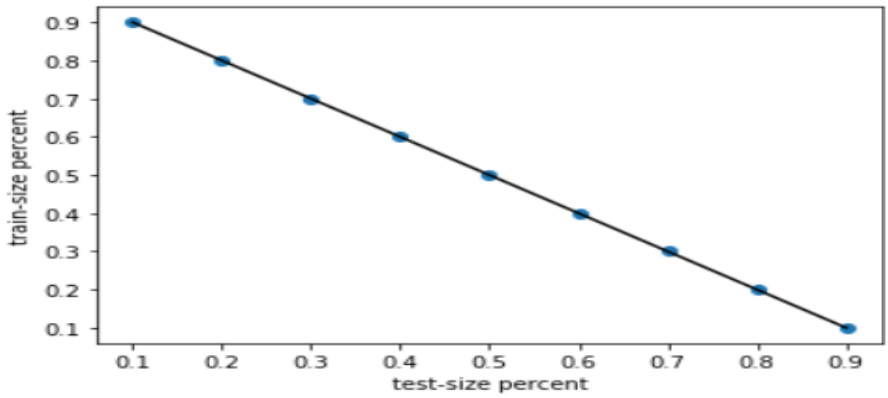

Fig 7.1 Train data and Test data(Their values affect final scores Fig 7.2)

```
Out[70]: Text(0.5, 0, 'test_split')
```

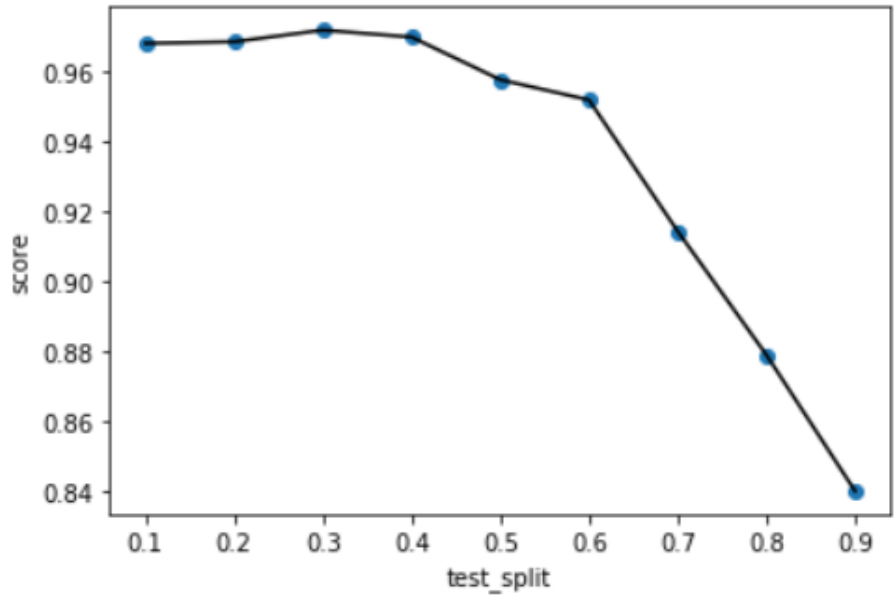

Fig 7.2 Scores based on test_split

Thus it can be inferred from these figures that test-split (which means selecting some percent as test data and the rest as train data) affects the final score of our classifier.So choosing an optimal value is a necessity in our case test-split=0.3 ended up with the maximum score.

8.CONCLUSION

We furnished results for Sentiment and Emotional Analysis on twitter data . On applying Logistic regression, Bernouille Naive Bayes and Multinomial Naive Bayes for sentiment analysis Multinomial Naive Bayes stands out with 96.4% accuracy at test_split=0.3.Users topic of interest for sentiment analysis has been considered ,So that they may get to know the statistics of sentiment behind the topic of their own interest. We firmly conclude that implementing sentiment analysis and emotional analysis using these algorithms will help in deeper understanding of textual data which can essentially serve a potential platform for businesses .

In future work , we aim to handle emoticons ,sarcasm, dive deep into emotional analysis to further detect idiomatic statements .We will also explore richer linguistic analysis such as parsing and semantic analysis.

9.REFERENCES

[1]. Shiv Dhar, Suyog Pednekar, Kishan Borad- Methods for Sentiment Analysis Computer Engineering, VIVA Institute of Technology, University of Mumbai, India.

[2].Ravinder Ahujaa, Aakarsha Chuga, Shruti Kohlia,Shaurya Guptaa,Pratyush Ahujaa- The Impact of Features Extraction on the Sentiment Analysis Noida, India.

[3].Richa Mathur, Devesh Bandil, Vibhakar Pathak-Analyzing Sentiment of Twitter Data using Machine Learning Algorithm

[4].Soumaya Chaffar and Diana Using a Heterogeneous Dataset forEmotion Analysis in Inkpen, School of Information Technology and Engineering, University of Ottawa ,Ottawa, ON, Canada.

[5]. Alec Glassford and Berk Coker Multiclass Emotion Analysis of Social Media Posts Stanford University.

[6].Shiv Naresh Shivhare and Prof.Saritha Khetawat Emotion Detection from Text Department of CSE and IT,Maulana Azad National Institute of Technology ,Bhopal, Madhya Pradesh ,India.

[7].Kavya Suppala, Narasinga Rao "Sentiment Analysis Using Naïve Bayes Classifier"International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019.

[8]. Bhagyashri Wagh,J.V. Shinde ,N.R.Wankhade"Sentiment Analysis on Twitter Data using Naive Bayes" International Journal of Advanced Research in Computer and Communication Engineering ISO 3297: 2007 Certified

[9]. Akshi kumar, Prakhar Dogra and Vikrant Dabas "Emotion Analysis of Twitter using Opinion Mining" Dept. of Computer Engineering , Delhi Technology University,New Delhi India

[10].Introduction to Machine Learning Alex Smola and S.V.N. Vishwanathan Yahoo! Labs Santa Clara –and– Departments of Statistics and Computer Science Purdue University –and– College of Engineering and Computer Science Australian National University

[11].Machine Learning Tom M. Mitchell

[12].An Idiot's guide to Support vector machines (SVMs) R. Berwick, Village Idiot

[13].https://www.svm-tutorial.com/2014/10/svm-linear-kernel-good-text-classification/

[14].https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[15].https://www.ijitee.org/wp-content/uploads/papers/v8i8/H6330068819.pdf