Essay Tone Detector
(Sentiment Analysis)

A

PROJECT REPORT

Submitted in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE

Under the supervision of

Dr. Vivek Sehgal

(Associate Professor)

by

Ridhi Soni (161362)

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, SOLAN – 173234
HIMACHAL PRADESH, INDIA

December – 2019

Candidate's Declaration

I hereby declare that the work presented in this report entitled "**Essay Tone Detector(Sentiment Analysis)**" in partial fulfilment of the requirements for the award of thedegree of **Bachelor of Technology in Computer Science and Engineering Information Technology** submitted in the department of Computer Science &Engineering andInformation Technology, Jaypee University of Information Technology Waknaghat is anauthentic record of my own work carried out over a period from August 2019 toDecember 2019 under the supervision of **Dr. Vivek Sehgal**, Associate Professor(Senior Grade), Department of Computer Science & Engineering and InformationTechnology.

The matter embodied in the report has not been submitted for the award of any other degree or diploma**.**

Ridhi Soni
(161362)

This is to certify that the above statement made by the candidate is true to the best of myknowledge.

Dr. Vivek Sehgal
Associate Professor (Senior Grade)
Department of Computer Science & Engineering and Information Technology
Dated: -

ACKNOWLEDGEMENT

It is our privilege to express our deep gratitude and regards to our project supervisor **Dr. Vivek Sehgal** for his mentoring, valuable inputs, guidance and constructive criticism throughout the duration of this project. We also express our sincere thanks for encouraging and allowing us to present the project on the topic “Essay Tone Detector(Sentimental Analysis)” for the partial fulfilment of the requirements leading to the award of B.Tech. degree. We would also like to thank Dr Satya Prakash Ghrera, Head of Department (CSE) for providing us a great opportunity to work on such an interesting project. Last but not least we would like to express our sincere gratitude to our family members who stood by us and stopped us in every phase of this project and gave us the much required moral support in carrying out this project successfully.

TABLE OF CONTENTS

TABLES

List of Figures

ABSTRACT

Social media has emerged as a platform to raise user's opinions, views and influence the way any business is commercialized. The Internet has led researchers to have access to a massive amount of data in various fields such as- such as machine learning, natural language processing (NLP) and data mining, management and marketing and even psychology-to detect the sentiment of the publicly available dataset.

The report focuses on various challenges and applications of Sentiment Analysis and discusses the different strategies to implement a computational approach of one's convictions. We will examine techniques for future vision the discourse formation. We will consider some distinct topics in Sentiment Analysis and modern accomplishments in those areas by using different machine learning algorithms to check the accuracy of the analysis.

Keywords: social media, twitter, sentiment analysis, sequence of words.

Chapter 1

INTRODUCTION

The determining of opinions on brands and services or understand customer's attitudes is the key to sentiment analysis, so we have preferred to serve with twitter as well as essay contents since we feel its better approach of common sensibility as compared to popular articles and blogs. In this project, we have used the collection of words limited to 140 characters in the case of twitter, but for the future, we don't define the words as it is for essay content and the expression that is expressed. Human annotators did an illustration of topics and feelings. The primary reason for taking Twitter as an example is that the amount of important data is much more substantial for twitter as compared to any web sites. One more reason is that the response on twitter is swift and is widespread since the number of users who tweet is more than those who write blogs daily.

When microblogging platforms are preferred, the task of sentiment analysis becomes even a lot of exciting. It introduces a new way of expressing the emotions, where only short texts are required, which would help us to detect the emotion of the person, so containing new kind, abbreviations, and grammatical mistakes that were generated by choice. For instance, Twitter is an internet microblogging and social networking platform that allows users to jot down short standing updates of most lengths of 140 words**.**

1.1 Machine Learning

This term applies to the “machine-driven detection of significant models in data.”Due to large amount of data size, this technique has become a conventional procedure for data extraction. As spam mail filtering and customized reportage to look engine optimizations and period of time object exposure systems, employed in numerous fields. Whereas the range of contemporary algorithms depends on the training task, purposeful literature contrasts in step with the character of the communication amongst the computer and also the atmosphere. As such, the division is formed between supervised and unsupervised algorithms.

1.1.1 Supervised Learning:

In this type of algorithmic program, the data to be trained "includes samples of the data vectors beside their corresponding target vectors". Example- A system can be trained to identify images of different animals. A set of marked pictures will be processed by the algorithm during the training level. At this time, the computer 'knows' which images include which type of animal. New unlabelled images are shown based on recognized algorithm before any kind of animal image. Therefore the goal of this learning algorithm outlines the input to the actual value.

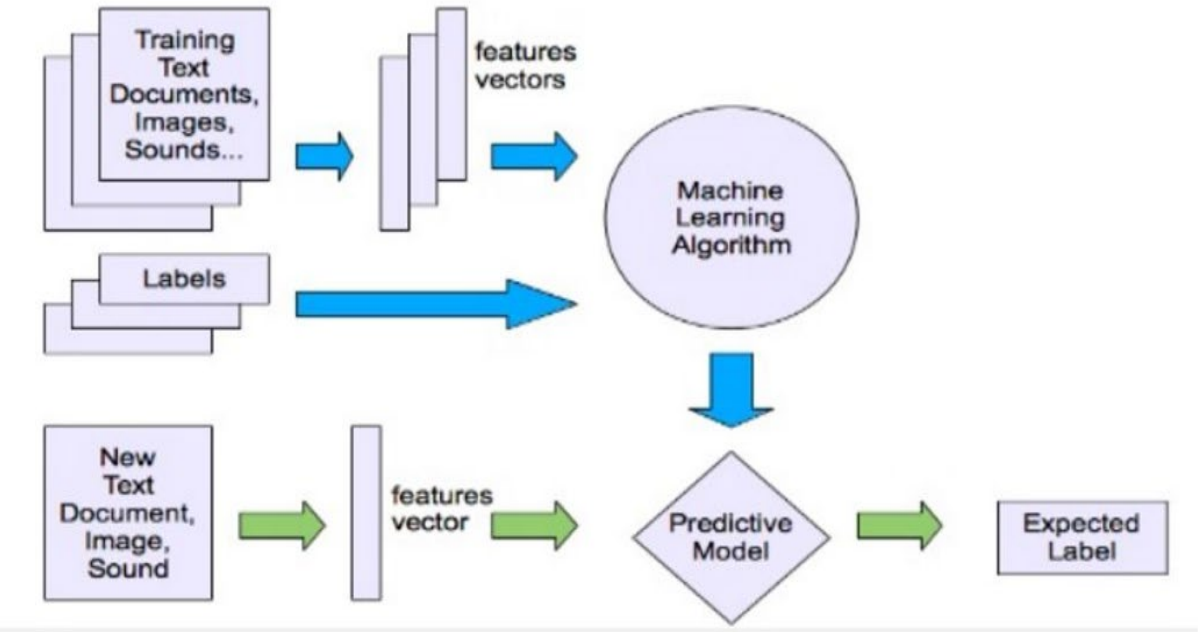


Fig 1: Supervised Learning Pipeline

1.1.2 Unsupervised Learning:

This learning algorithms has the same extent as supervised learning, which is to predict the input to actual value. But in this type of algo input is not determines earlier as we are not provided by any independent variable. This algo consists of dataset without any labelled responses.

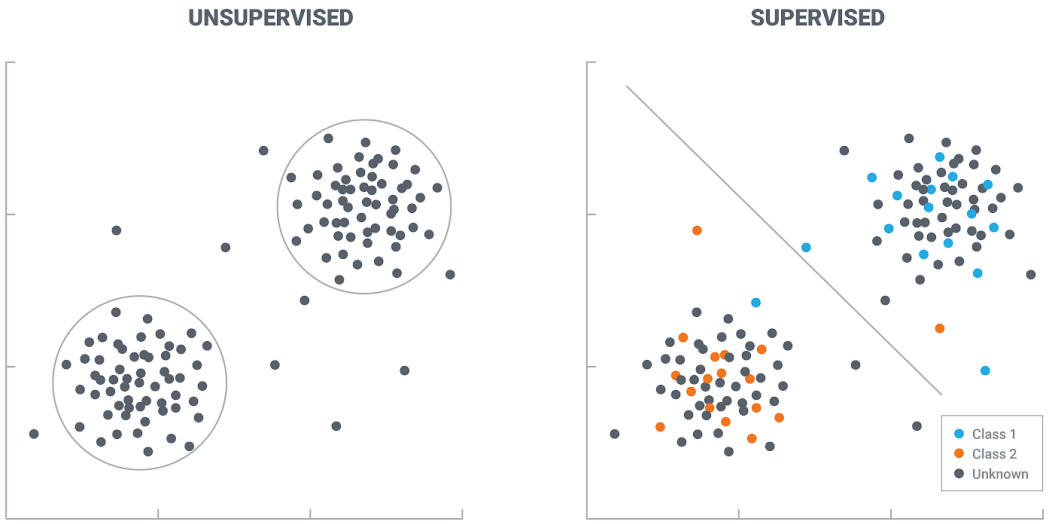


Fig 2: Supervised VS Unsupervised

1.2 OBJECTIVE

Essay Tone Detector classifies the expressions in a text that exhibits to have some sensibility. A decent amount of study has been done on this topic with the use of reviews, blogs, articles, documents, and various phrases. This differs from twitter as it is a socialnetworking platform that limits users to write the updates of 140 characters only. The motive of this work is to work on an algo that can accurately classify any textas joy, sadness, anger, disgust, and fear, and our main objective is that we can achieve the most accurate value on incorporating sentiment on the provided dataset using machine learning algorithms.

1.3 PROBLEM STATEMENT

This sentiment analysis as if we consider in current times plays a vital role in depicting the emotions of the particular sequence of sentence therefore research in this area is growing.

There is still much room left for further research in this area. For a given a message, we need to classify whether the message is of joy, anger, sadness, fear, disgust, and surprise. For messages conveying all the sentiments in the text in the form of a graph. The researchers test new features and different classification techniques and compare the result to the baseline performance. Comparisons between the results came from various features and classification algorithms to achieve the most suitable features and most structured classification technique.

Chapter 2

LITERATURE REVIEW

Paper – 1

Suresh, H., 2016, October. An unsupervised fuzzy clustering method for twitter sentiment analysis. In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 80-85). IEEE.

Main Contribution:-

In this paper, the author introduces a unique clustering model to analyze the tweets regarding the sentiments of all the different kinds using the real dataset handled over the last year. Corresponding analysis is created with the help of the currently existing cluster techniques that helps to achieve high accuracy, precision, recall and execution time.These are calculated using K Means algorithm and the Expectation-Maximization algorithms which mainly comprisesof statistics.

Algorithm Used:-

The performance was estimated according to the variable counts and semantic relationship with the use of K-Means algo. The scientists used twitter statistics for identifying the vulnerabilityreaction on a dataset which was collected on destructive Hurricane study and Boston marathon bombardments. The tests were executed using the Naïve-Bayes algorithm. No data was found which included the data and the instruments which were used for the evaluation of the experiment.

The more the data, the better is the output.

Result Analysis:-

As per the experiment, the recommended procedure is verified to be beneficial in carrying out quality reslts in this domain i.e of twitter sentiment analysis. The K-means clustering

method is more effective than Expectation-Maximization clustering method because of the 5% more efficiency and time required to form a better model requires 0.1 seconds as to 1.06 seconds against EM. The given model gives a pretty good efficiency of about 76.4% and less time is required to create a model as compared to the other two methods. Therefore this model is determined to be better for real time applications considering efficiency and execution time as our primary parameters.

Future Scope:-

The further work under this includes addition of n number of samples of tweets, the performance would be examined upon distinct clustering techniques and a better and efficient method would be built to get better results in analyzing the twitter feeds.

Paper – 2

Sarlan, A., Nadam, C. and Basri, S., 2014, November. Twitter sentiment analysis. In *Proceedings of the 6th International Conference on Information Technology and Multimedia* (pp. 212-216). IEEE.

Main Contribution:-

This paper includes the design of sentiment analysis, extorting a large amount of twitter feeds. For development, Prototyping was used.

Algorithm Used:-

- NLP-Natural Language Processing
- CBR-Case-Based Reasoning
- ANN-Artificial Neural Network
- SVM-Support Vector Machine

Result Analysis:-

Results classify clients' prospects via twitter feeds into positive, negative and Null, which was depicted in a pie chart and a HTML page.

Future Scope:-

The researches include that the furthur work be done in building up the web application with the help of django as it only runs on the LinuxServer & LAMP, therefore it was not completed. Hence the author includes this element to be focused on in future.

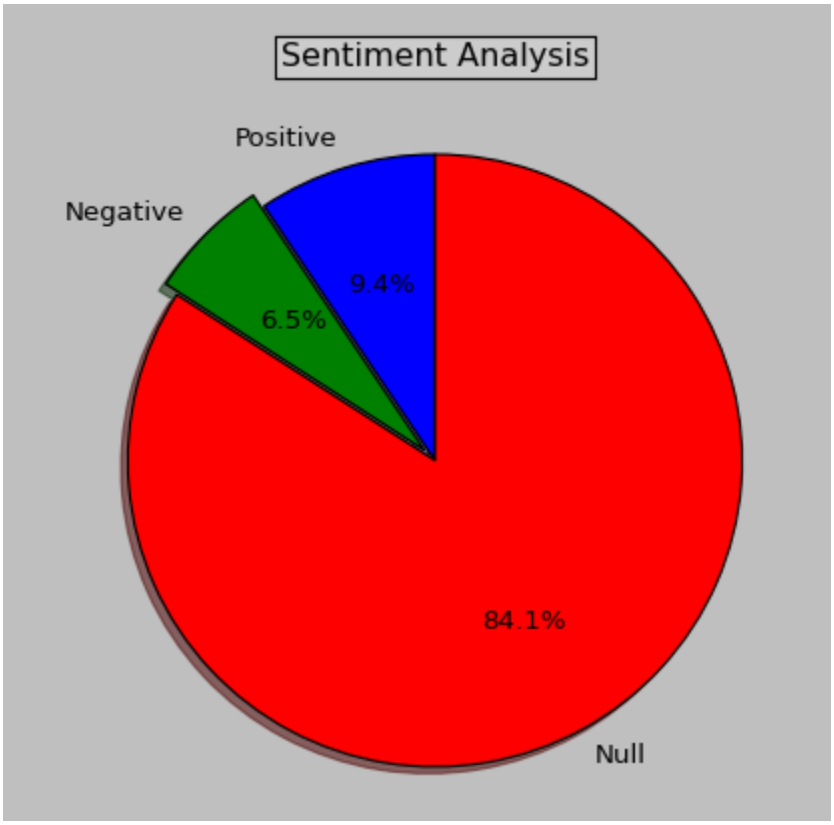


Fig 3: Pie chart For Result

Paper – 3

IEEE. Pandarachalil, R., Sendhilkumar, S. and Mahalakshmi, G.S., 2015. Twitter sentiment analysis for large-scale data: an unsupervised approach. *Cognitive computation*, *7*(2), pp.254-262.

Main Contribution:-

In this paper, the researcher uses unsupervised machine learning algorithms for analyzing twetter sentiments. The polarity of twitter feeds are assessed by using 3 sentiment lexicon which are SenticNet, SentiWordNet, and SentislangNet. Sentislang_Net is a sentiment lexicon built from SenticNet and SentiWordNet for dialects and abbreviations.

Algorithm Used:-

Twitter Sentiment Analysis (TSA) has two stages:

(1) pre-processing stage

(2) sentiment analysis stage.

The twitter feeds are normalised and converted into uni-grams (fourgrams, trigrams, bigrams and unigrams) in the first phase. The next phase explains the sentiment contradiction in every tweet from the created unigrams. Sentiment polarity of every lingo is assessed by operating the sentiment analysis algorithm on keys using SenticNet and SentiWordNet and then they are eliminated or corrected manually.

Result Analysis:-

The results show the performance of the sentiment analyzer on two data sets. The outcomes determines that our unsupervised approach achieved a rationally reliable recall & Fmeasure and surpasses all basic approaches.They discovered that applying other sentiment lexicons (SenticNet, SentislangNet) along with SentiWordNet has enhanced the efficiency of sentiment analysis.

Future Scope:-

Utilizing other sentics (affective learning) when compared with every other theory might bean assuring future scope for twetter sentiment analysis.

Paper – 4

Ramadhan, W.P., Novianty, S.A. and Setianingsih, S.C., 2017, September. Sentiment analysis using multinomial logistic regression. In *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)* (pp. 46-49). IEEE.

Main Contribution:-

This paper includes the sentiment analysis on the Jakarta Governor Election, where the initial step was collection of tweets using a twitter API which included names of every contender on Jakarta Governor Election. The tweets that were collected were used as an input for preprocessing. Then the following step was to obtain useful data from the tweets. The features were reconstructed into a vector in 0/1 form and modified again using the Tf/idf method. The dataset contains two sorts of data i.e trainingdata and testingdata.

Algorithm Used:-

Multinomial Logistic regression: It is a regression design which causes the logistic regression to classification problems. It is a type of regression where the output can take more than two possible values. K-Fold CrossValidation was used to test the efficiency of the algo.

Result Analysis:-

After all tests, the most significant factor in receiving a decent result is the creation of training as well as testing data. To attain higher accuracy, more training data needs to be associated to the number of testing data.

Future Scope:-

The future scope of this paper would be working on more powerful twitter API along with powerful machine learning algorithm.

Paper – 5

Da Silva, N.F., Hruschka, E.R. and Hruschka Jr, E.R., 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, pp.170-179.

Main Contribution:-

The author presents a way by using classifier ensembles and lexicons which automatically analyzes the sentiment of twitter feeds. Twitter feeds categorized as either positivee or negative. This method is beneficial for users that examine outputs for companies whichfulfil the sentiment of their brands along with various other applications. Indeed, sentiment analysis in microblogging (e.g., FB, IG, and Twitter) through classifier ensembles and lexicons were not thoroughly examined in the research paper.

Algorithm Used:-

The algorithms used were:-

- Random Forest
- Naïve Bayes
- SVM Linear
- Logistic Regression

Result Analysis:-

The results of stand alone classifiers were compared withthe ensemble program by the author. By examining various sequences of bag-of-words (B-o-w), FeatureHashing (FH), and lexicons, the author evaluated the possible combos to enhance the certainity to classify sentiments. The best results defined in this report were recorded for analyzing purposes.

Future Scope:-

The future scope of this paper would be to study neutral twitter feeds, where the datasets are supplemented along with analog area data sets and comprises of distinct traits. As

weknow that diversity is a crucial feature for the thriving use of ensembles. Hence, more time will be devoted to this area of research.

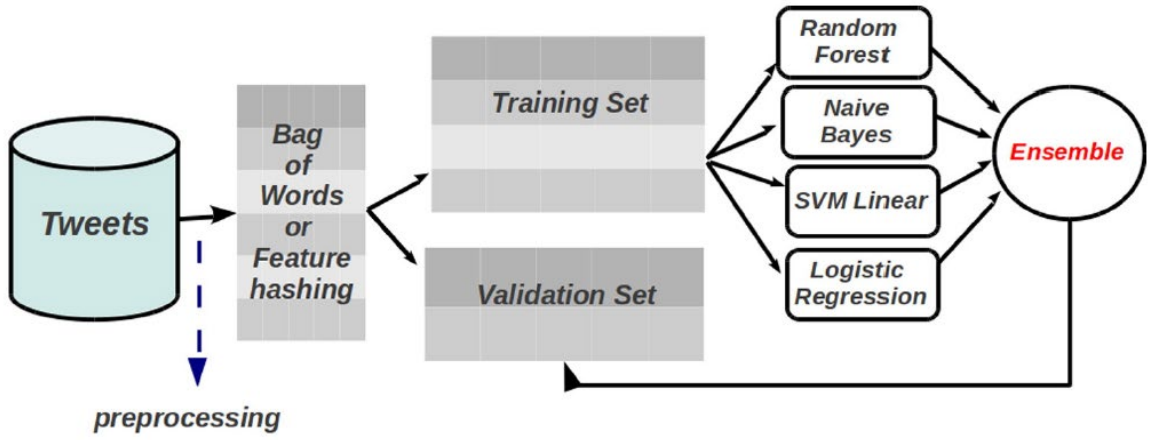


Fig 4: General Process of System

Paper – 6

Troussas, C., Virvou, M., Espinosa, K.J., Llaguno, K. and Caro, J., 2013, July. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In *IISA 2013* (pp. 1-6). IEEE.

Main Contribution:-

The first belief is the matter to know how people sense some specific topics which be acknowledged as a classification job. The sentiment is in subtle or advanced procedures in a given dataset. Any user can use a broad range of additional procedures to express their feelings over the internet. As data collected from any other website contains a lot of noise. As sentimental analysis is playing major role in social media because on social media everyone is free to express their emotion, thought or feelings. In order to extend this, the companies have therefore extended their customersatisfaction analysis over the internet such that they are able to collectlots of data. These studies are mainly targeted to twitter as in Twitter is restricted to only certain character length which tends user to use abbreviation and fragmented expressions.

Algorithm Used:-

- **Naive Bayes Classifier:** This classifier is twirling around Bayes rule that is looking at conditional probabilities that allow flipping the condition around. It provides us knowledge of event A will occur, given the evidence B. 222This helped in determining that the expectation of the counter result as to the given two components(A/B) independently:

$$P(A|B)=P(A)P(B|A)/P(B) \quad (1)$$

- **Corpus Creation:** It is an extensive collection of articles or registered statements used for linguistic interpretation. This work has been pigeonholed into two sections based on user status updates. Some of the particular range of users are recorded for Corpus. The system is trained according to the sentiments that the people share.

- **Corpus Creation:** It is an extensive collection of articles or registered statements used for linguistic interpretation. This work has been pigeonholed into two sections based on user status updates. Some of the particular range of users are recorded for Corpus. The system is trained according to the sentiments that the people share.

- **Classification:**Given the evidence, It is a conditional probability under which the event(A) will occur. Therefore the primary formula is :--

$$P(sentiment\_tokens)=P(sentiment)P(tokens\_sentiment)/P(tokens) \quad (2)$$

Result Analysis:-

In this study, we used three classifiers which were required in comparing the performance of fb status. The information gathered is about seven thousand status uploads from about ninty users which were physically marked as positive or negative. Samples are shown in this following table:

Table 1: Sample of Status updates

Sample of negative status updates:	Sample of positive status updates:
• Freaking full of doubt.	• Just finished making pancakes for breakfast oh and the yummiest part, it comes with a free strawberry syrup!
• I really don't like to shave my hair but i have to. frustrated :(	• inspired by you <3
• Can't sleep...	• 11 days to go before Christmas :)

Table 2: Data Distribution

	Training	Testing
Positive	1142	1142
Negative	1142	1142

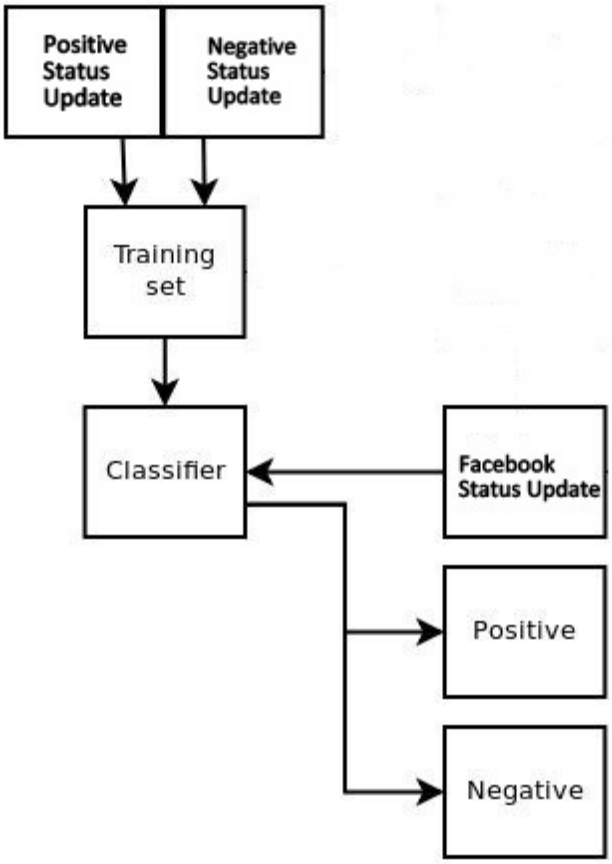


Fig 5: Tweets Processing using Esembler

Paper – 7

Kumar, A. and Sebastian, T.M., 2012. Sentiment analysis on twitter. ***International Journal of Computer Science Issues (IJCSI)*****, *9*(4), p.372.**

Main Contribution:-

As there is a rise in social networking, there has been user-generated content. Micro-blogging is a site that has millions of users sharing their thoughts daily. In this paper,

popular real-time micro-blogging takes into consideration real-time posts to their opinions about everything. This work provided to illustrate productiveness of the proposed system.

Algorithm Used:-

- **Pre-processing of tweets:-** The first step includes the removal all the URLs, hash tags and special character or twitter and calculate the percentage of the tweets in caps. Correct the spellings, which are the sequence of characters tagged by weight and is done by regular usage. By replacing all emotions with their sentiments tag and removing all the punctuation after counting.

- **Scoring module: -** This module differentiates the opinion carriers as the adjectives, verbs, and adverbs. Then, they used the corpus based approach to label the semantic adjustment of adjectives and the dictionary based approach to obtain the semantic adjustment of verbs and adverbs.

Result Analysis:-

This is the novel approach used to determine the sentiments on twitter feeds by extracting adjectives along with verbs and adverbs. The total twitter sentiment was measured by applying a linear equation that included emotion intensifiers.

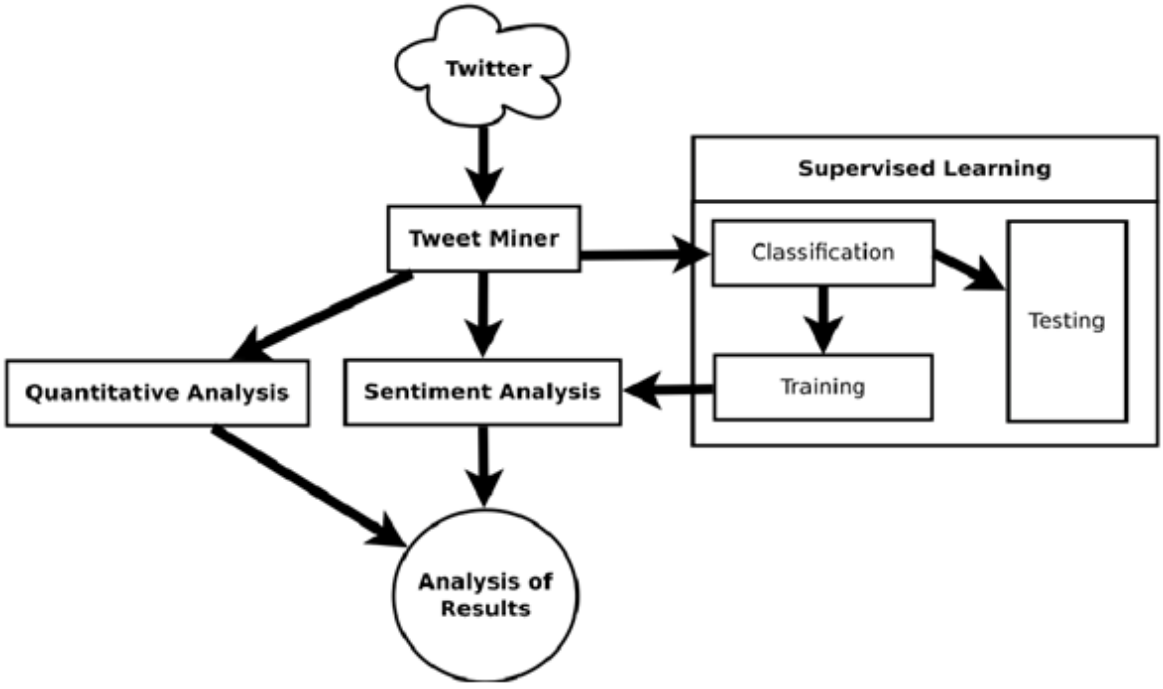


Fig 6:Twitter Processing

Paper – 8

Bharathi Bhaskaran, R., Prabhakaran, R., Saravanan, S. and Vinoth, M., 2018. TWITTER SENTIMENT ANALYSIS. *International Journal of Pure and Applied Mathematics*, *119*(10), pp.1785-1791.

Main Contribution:-

In this main aim is to erect a useful classifier for precise and programmed sentiment for foreign tweets. As there is no limit to the range of data transferred by texts, but short messages are used to share views and attitudes which are provided by the users. The foremost task is to give a note and classify it and depict the more robust sentiment to be chosen. As twitter has such a huge viewer, which draws the user to bear their ideas and panorama about any problem or any other attractive ongoing issues.

Algorithm Used:-

This article has been edited using Python language, which is a high level described language and is very attractive for its code readability and compactness.

1. Natural Language Processing
2. SCIKIT-LEARN
3. NumPy
4. Setting up environment for Sentiment Analysis
5. Data collection
6. Pre-processing
7. Feature Extraction

Result Analysis:-

The activity manifested in this document designates a unique design to determine sentiment analysis on twittter data. The corpus based method was used to find the denotative bearings of adjectives and the dictionary-based approach to find the semantic positioning of verbs and adverbs. A linear equation was used to determine the final sentiment of the given twitter dataset.

Future Scope:-

Sentiment Report is not merely a refined analytics means. It's an attractive area of research. Sentiment analysis courses till now have been used to identify the antithesis in the beliefs and ideas of all the users that access social media. Researchers and Systems are very interested to learn the thoughts of forms and how they respond to everything appearing around them. Groups use this to assess their promo operations and to promote their merchandise.

Paper – 9

Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), pp.1093-1113.

Main Contribution:-

This paper tackles a broad survey of the last update in this domain. Many recently introduced algorithms' improvements, and many Sentiment analysis applications are reviewed and introduced briefly in this survey. These chapters are classified according to their contributions to multiple Sentiment analysis techniques. The relevant areas to Sentiment analysis (transfer learning, emotion recognition, and constructing resources) that drew researchers recently are discussed. The main aim of this paper is to give an almost complete image of Sentiment analysis techniques and the related areas with brief details.

Algorithm Used:-

The Work done by the author was on the following algorithms:-

- NBC-Naive Bayes Classifier
- BN-Bayesian Network
- ME-Maximum Entropy Classifier
- SVM-Support Vector Machines Classifiers
- NN-Neural Network
- Decision tree classifiers

Result Analysis:-

After examining these articles, inevitably, the improvements in SC and FS algorithms are still an open domain for research. Naı¨ve Bayes(NB) and Support Vector Machines(SVMs) are the most often used Machine Learning algorithms for solving SC problems. They are recognized as a reference model where several recommended algorithms are compared to each other.

Future Scope:-

In future work, the author plans to devote his time to research on content based SA.

Paper – 10

Danieel Gayo-Ayvello , Panagiiotis T. Metaxas and Eni Murstafaraji 2011, What are the limitations of Electoral Predictions Using Twitter

Main Contribution:-

Utilizing social media for political conversation is converting a common practice, especially around election time. One exciting feature of this drift is the hope of pulsing the public's view about the elections, and that has pulled the attention of many researchers and editors. Allegedly, predicting elective results from social media data can be available and even flat. Positive outcomes have been reported but without a study of what system enables them. Our business puts to review the purported auspicious knowledge of social media metrics toward the 2010 US congressional polls. Unluckily, we find no association between the analysis results and the electoral results, disclaiming prior articles.

Algorithm Used:-

The Work done by the author was on the following algorithms:-

1. Twitter API
2. Polarity lexicon

Result Analysis:-

Polarity lexicons were used to discover positive, negative, and neutral verses. The results show us that the complexity of expert polling system can never be simulated by experimenting social-media data. The researchers have concluded that Twitter dataset was not much dependable while predictingthe result of the United States elections.

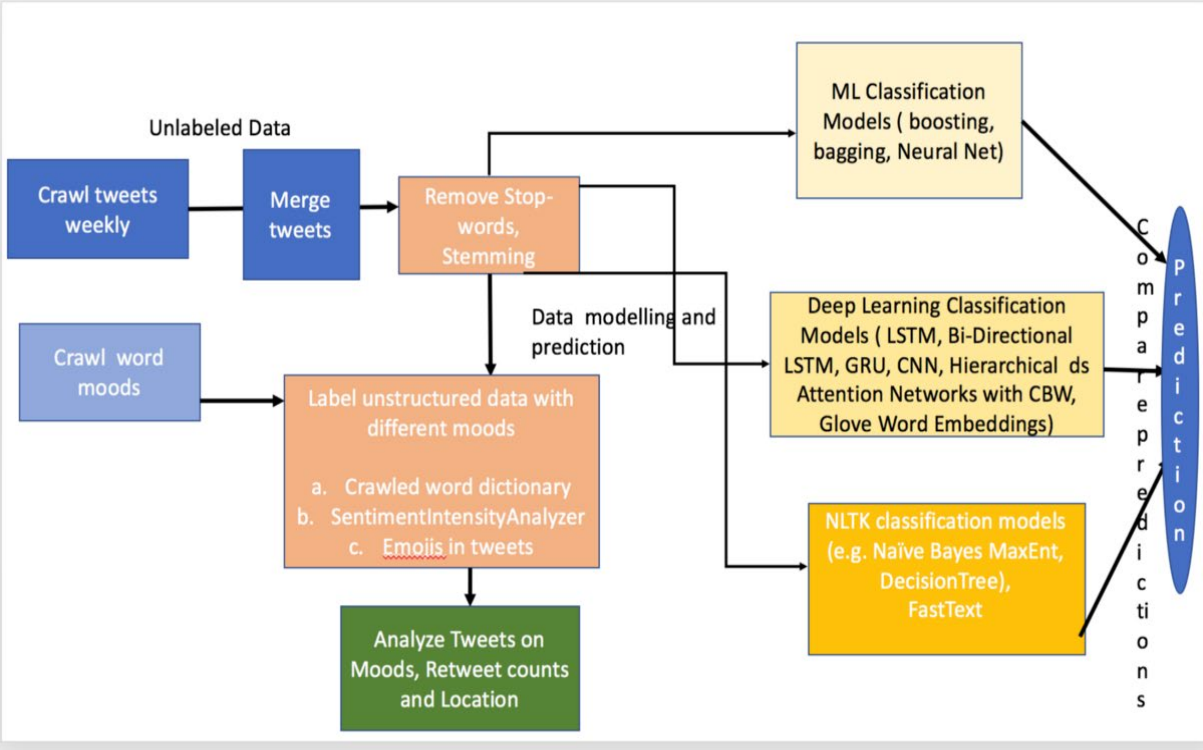


Fig 8: Tweets processing[yy]

Paper – 11

Bollen, J., Mao, H. and Pepe, A., 2011, July. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Fifth International AAAI Conference on Weblogs and Social Media.

Main Contribution:-

We use a psychometric tool to derive six mood elements anxiety, crisis, violence, vigor, weakness, trouble from Twitter content. We relate our results to a background of mass events associated with media and references. We observe that improvements in the fields such as(social, political, cultural, and economic) do have a meaningful, critical, andhighly precise effect on the various dimensions of the civil mood. We hypothesize that broad-scale investigations of atmosphere can provide a robust platform to form collective emotive trends in terms of their auspicious value with concerns to enduring social as well as economic symbols.

Algorithm Used:-

Profile of mood states

Result Analysis:-

They researched how public mood patterns relate to variations in macroscopic social and economical displays. They presented results for sentiment analysis on Twitter. They utilized beforehand proposed best in class unigram display as their model and revealed a general pick up of more than 4% for two arrangement errands.

Paper – 12

Altrabsheh, N., Cocea, M. and Fallahkhair, S., 2014, November. Sentiment analysis: towards a tool for analysing real-time students feedback. In *2014 IEEE 26th international conference on tools with artificial intelligence* (pp. 419-423). IEEE.

Main Contribution:-

To discuss this dilemma, we intend to examine feedback automatically using sentiment summary. Sentiment study is domaindependent, and although it has remained achieved to the instructional domain before, it has not used for real-time feedback. To find the most suitable model for an automated summary, we look at four aspects: preprocessing, characteristics, machine learning techniques, and the use of the inactive class. We found that the most leading result for the four characters is Support Vector Machines with the highest level and no neutral quality, which gave some percent efficiency.

Algorithm Used:-

The Machine Learning Techniques used:-

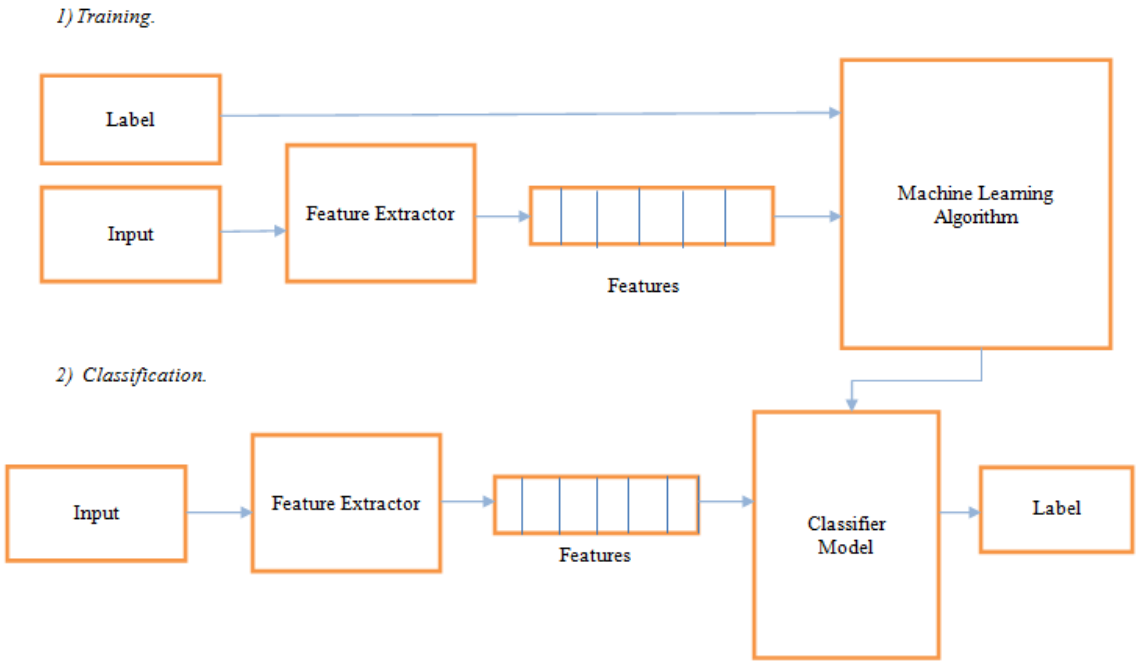
1. Support Vector Machines
2. Naive Bayes

Result Analysis:-

In this article, we examined different series of machine learning techniques, opinions, preprocessing levels, and the use of neutral class for analyzing real-time learners' feedback. However, interestingly, we found that in some models with the dull category, using no preprocessing gave the most dominant performance. Symbols and punctuation for our attention may hold some significance, therefore reducing these through preprocessing may point to a loss of reliable data which, in turn, drives to diminished production of the figures.

Future Scope:-

The expected task covers testing pos tagging as a background, including several preprocessing methods such as contradiction, and keeping the numbers and punctuation. We also intend to broaden our investigation area into identifying distinct sensations related to knowledge**.**



Paper – 13

Psomakelis, E., Tserpes, K., Anagnostopoulos, D. and Varvarigou, T., 2015.Comparing methods for twitter sentiment analysis.arXiv preprint arXiv:1505.02973.

Main Contribution:-

The significant participation of this work is the extensive example of opinion duality classification methods for Twitter text, the embodiment of a mixture of classifiers in the related set, and the collection and use of several manually explained tweets for the evaluation of the processes. Particularly concerning the latter, we contemplate it to be a foremost contribution in the knowledge that from experience the computerized explanation of tweets based on the exposure of pieces like the emoticons has been uncertain since it does not always reflect the case about the overall sentiment manifested by the author, especially when one considers the declaration of emotions through the text.

Algorithm Used:-

1. Logistic Regression
2. BestFirst Trees
3. Functional Trees
4. SVM
5. Naïve Bayesian

Result Analysis:-

In this activity, we performed a study and a summary of the most leading methods for sentiment analysis on Twitter. The emphasis was on the different models and the successions of many classifiers. The outcomes confirmed the advantage of graphs in capturing the expressed attitude in a document and especially in tweets. They also showed the developments that various alliances of methods and machine learning algorithms can produce in the belief rates of some sentiment analysis methods.

Future Scope:-

The discovery of this product collected in the careful evaluation of the performance of various sentiment analysis devices using manually explained datasets, as well as in the demonstration of the possibility to combine methods, creating new techniques for enhancing the quality of the outcome.

Paper-14

Martínez-Cámara, E., Martín-Valdivia, M.T., Urena-López, L.A. and Montejo-Ráez, A.R., 2014. Sentiment analysis in Twitter. ***Natural Language Engineering*****,** ***20*****(1), pp.1-28.**

Main Contribution:-

The investigation space that is interested in sentiment analysis has mature pretty much at a fast speed. Official papers and conferences or connected leagues are obligatory to grasp that this can be a course with sensible views for the prospect. The twitter growth has expanded analysis during this space due radically to its native applications in areas cherish business intelligence, recommender systems, graphical, though, to perceive this concern, a severe revision of state of the art is initially necessary. It's for this reason that this writing aims to represent a place to begin for those researches involved with the tardiest attributing to Twitter in Sentimental analysis.

Algorithm Used:-

1. Political opinion mining in Twitter
2. Polarity classification
3. Temporal prediction of events

Result Analysis:-

Sentiment Analysis could be a fleetly developing space within which investigation has been brought come in many various regions and on various issues connected with the task. Besides, from our purpose of reading, the high interest this discipline presently excites is

that the doable applications of the technology in domains reminiscent of business statistics, sentiment analysis, and recommender systems. Thus, it's necessary to develop systems that may extract the basic data scattered through the social network of Twitter.

Future scope:-

Sentiment analysis may be an unambiguously valuable tool for businesses that are wanting to live emotions, opinions, and sensations concerning their trademark. To date, the majority of sentiment analysis plans are accompanied virtually solely by businesses and makes through social media information, survey acknowledgments, and alternative centers of user-generated content. By reviewing and examining client sentiments, these brands will get interior to cross-check client practices and, finally, higher serve their readers with the product, services, and experiences.

CHAPTER 3: SYSTEM DEVELOPMENT

3.1 Software Requirement Specification:

- Processor: Intel Core i5 (or above).
- Data Bus: 64 bit.
- 8 GB RAM
- Operating System: Windows 10
- Python version: 3.6+
- MS Excel

3.2 Model Development:

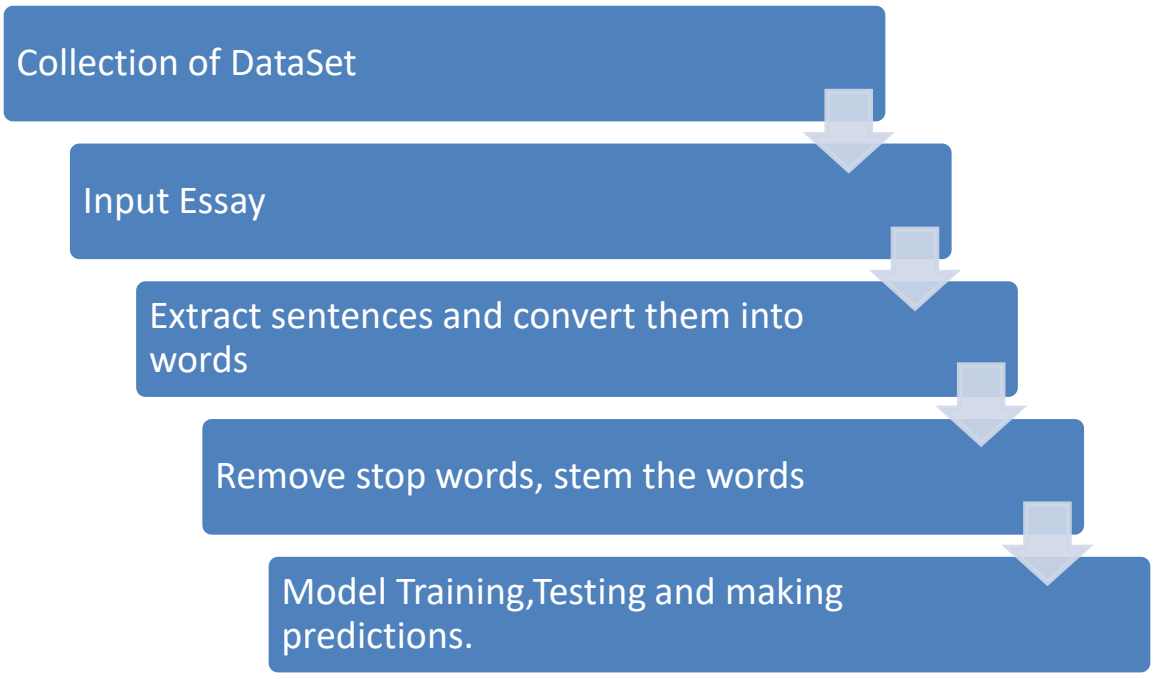


Fig 7:Model Development

Emotion is one kind of result, a unique, reasonably being mood, temperament, and sensation. Emotions are widely studied in science and behavior sciences, as they're a vital part of the attribute. Nowadays, they need conjointly attracted the eye of engineering science researchers, particularly within the field of AI. With recent advances within the field of matter analysis, the realm of feeling detection has become a favorite in process linguistic. Since feeling detection is that the newer space of matter analysis, it's weaker customary strategies, the feeling will be expressed as happiness, sadness, anger, disgust, fear, surprise, so forth. The tool developed as a district of this educational program permits a user to transfer a computer file and predicts the tone of the uploaded text with concerning 85%accuracy—the model trains at fifty,50000 random sentences out of the sentences within the dataset obtained from Kaggle. The rule accustomed predict the feeling of the sentences within the computer file is Logical Regression. When with success determinant the sentiment of every sentence, the tool returns the foremost oft occurring feeling all told the sentences. It conjointly computes a pie chart depicting the relative distribution of other emotions in the input file**.**

3.3 Text Mining:

Text mining introduces the interpretation of data included in the original language text. It can be described as the method of deriving significant data from tangled text references. The reinforcement area of text mining differs from biomedical applications to marketing applications and sentiment analysis. In retailing, text mining is described as the outline of consumer bond executives. This means a firm can intensify its predictive analytics models for customer turnover and keep track of customer views. The main motive behind text mining that is used to process data into a structured manner, ready for analysis, via the purpose of natural language processing and other scientific methods. Albeit, there are many perspectives within the field of study of text mining, information gathering (IE) is suitable for this project. Consequently, the following material points to define the difficulties and technology associated with information extraction and succeeding processing.

3.4 Text Analytics:

Natural Language ToolKit is a robust Python package that presents a set of various natural language algorithms. It is available, opensource, straightforward to use, big alliance, and well presented. Natural Language ToolKit consists of the most prevalent algorithms, such as tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation, and named entity recognition. Natural Language ToolKit assists the computer to examine, preprocess, and learn the reproduced text.

Each one of the above mentioned data processing stages are mentioned below in detail.

1 Tokenization:

It is the procedure of splitting a stream of textual content up to words, figures, and different purposeful factors brought up tokens. Tokens are often separated by the method of whitespace symbols and/or punctuation characters. The primary task that has got to be achieved before any process will transpire is to separate the matter knowledge into smaller elements. This is often a primary step in natural processing language, referred to as tokenization. At a better stage, the text is split at the start of passages and sentences. It's performed so we are able to appear to be at tokens as person factors that form up a tweet. As associate degree outcome of the length restriction of one hundred forty characters required by twitter, it's seldom the actual fact that twitter feeds can contain over a paragraph. In these concerns, the design set up at this step is to spot sentences properly. This may be done by playacting the punctuation marks resembling an amount mark. When the the text is analyzed. Emoticons and abbreviations are recognized because of the part of the tokenization method and forbidden a person or girl tokens. The successive step is to extract the tokens from sentences. The hurdle at this step is to manage the writing system among a sentence. Consequently, synchronic linguistics mistakes got to be fastened, URLs and punctuation shall be eliminated from the ensuing kit of tokens.

```
text: "Want to boost Twitter followers ?! http://bit.ly/8Ua""
tokens: ["want", "to", "boost", "twitter", "followers"]
```

FIG x

2 Stopwords:

The process of changing information to another thing is said as pre-processing. One in each of the foremost kinds of pre-processing is to separate useless information. In the language process, useless words (data) are said as stop words. A stop word could be an unremarkably used word that an inquiry engine has been programmed to ignore, each once categorization entries for looking and when retrieving them because of the results of an inquiry question. We might not wish these words to require up area in our info or taking over the precious time interval. For this, we are able to take away them simply by storing an inventory of words that you just deliberate to stop words. Natural language Toolkit in python incorporates a list of stopwords hold on in sixteen completely different languages. Stopwords are considered as noise in the text. Text may contain stop words such is,a, am, are, this, a, an, the, etc. Therefore, one needs to remove the stopwords.

3 Normalization:

The presence of abbreviations inside a tweet is declared, and then the abbreviations are replaced with their real definition under the normalization process. (for e.g., BRB − > be right back). We additionally perceive informal intensifiers such as all-caps (e.g., I LOVE this show!!! and personality repetitions their presence in the tweet. All-caps phrases are made into decrease case, and instances of repeated characters are replaced through a single character. Finally, the presence of any distinct Twitter tokens is referred to (e.g., #hashtags, usertags, and URLs) and placeholders indicating the token kind are substituted. Our hope is that this normalization improves the performance of the POS tagger, which is the final preprocessing step.

4 Part-of-speech (POS):

Parts of speech are the process of designating a check to every phrase in the order as to which grammatical phase of speech applies to,(noun, verb, adjective, adverb, coordinating conjunction, etc.) The relationship among its words has to be confirmed. This can be done by allowing every word, a class that acknowledges the syntactic functionality of that word. For each tweet, we have points for counts of the number of verbs, adverbs, adjectives, nouns, and any different parts of speech. Also identified as a section of speech tagging this action can be seen as an auxiliary requirement for n-grams assortment and lemmatization. The table given below includes the part of speech arithmetics used in the project..

Table 4:Tokenization

ADJ : adjective	**PART :** particle
ADV : adverb	**PRON :** pronoun
AUX : adjective	**PROPN :** proper noun
CONJ: conjunction	**PUNCT :** punctuation
DET: determiner	**SYM :** symbol
NOUN: noun	**VERB:** verb
NUM: numeral	**X :** other

5 Stemming and Lemmatization:

Stemming and lemmatization are basically used to decrease inflectional models and derivations of a term to a standard base model. The words like cry, cries, cried, crying will have an identical base, that is, cry. Stemming is a simple approach that cuts off the words edges so that the base model is considered. Lemmatization is the structural research of the words, reflecting their meanings, normally introduced as the lemma. Though for the English language is more structured, rich language, this method relies on their meanings rather than on the tenses used. Besides, a lemmatizer can import uncertainty offering all feasible lemma for word form or a word having different meanings. For performing this task as analyzing the reason, we have used Poter's stemming algo. It contains five stagings, where word compressions are offered. For the respective steps, rules, and

protocols to utilize them are described. Given below figure explains the rules of the first stage of the algorithm:

Table 5:Rules

Rules	Examples
SSES -> SS IES -> I SS -> SS S -> /	caresses -> caress ponies -> poni caress -> caress cats -> cat

3.5 Libraries Used:

Tweepy: -

The principal difference between Basic and OAuth authentication is the user and access codes. By using Basic Authentication, it was feasible to implement a username and password and obtain the API, but since 2010 when Twitter began demanding OAuth, the method is a bit more complicated.OAuth is a little complex initially than Basic Auth as it needs more work, but the advantages it awards are very productive:

- Twitter feeds can be customized to produce a series of strings that recognize the app, which was adopted before.
- It does not expose a username or the corresponding key, making it safer.
- It is more comfortable to handle the authorities, for example, a set of tokens and keys can be produced that only enables reading from an external source, so in case someone gets those access tokens, that person will not be able to draft or communicate through direct messages, reducing the chances of any attacks.
- If the user changes the password, the application you use the token at will still work.

- Methods utilized:
    i) Set_access_token
    ii) Items ()
    iii) Cursor

Textblob:-

TextBlob is a python module that allows the use of a simple API to access its methods and for processing textual data.

It includes the following functionalities:

- Tokenization
- Normalization
- Part-of-speech
- Stemming and Lemming
- StopWords

Matplotlib: -

Matplotlib is a Python 2D plotting library which provides publication feature designs in a kind of hardcopy arrangements and interactive conditions over programs. Matplotlib works to create easy tasks, easy and difficult tasks achievable. You can create plots, scatterplots, histograms, bar charts, or any type of charts, etc., with merely a few chunks lines of code. For simplistic plotting, the pyplot module presents a MATLAB-like interface, especially when united with IPython. For the potential user, you have complete command of line forms, font attributes, axes attributes, etc., using an object-oriented GUI or a set of functions that are commonly known to MATLAB users**.** These are some samples: -

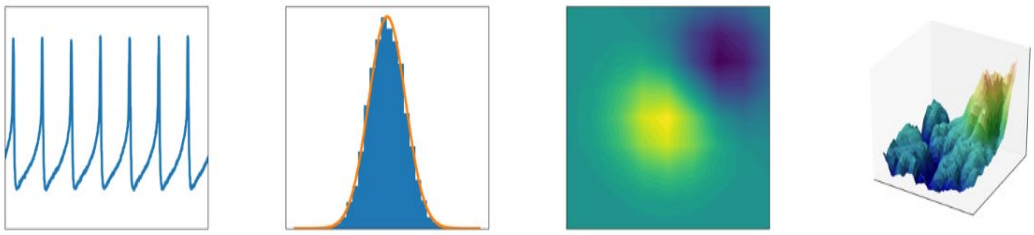


Fig 8:Graphs for explaination

- Methods Utilized:

i) axis()

ii) light_layout()

iii) appear()

iv) pie

v) legend ()

vi) title()

For stemming Natural processing language as well as natural processing toolkit (External libraries) is used which helps us to break the sentence so that all the required words would be stored in list as stop words are deleted and all the tenses used in the paragraph are converted into single word which implies the same meaning which is required to specify the emotions of the statement.

3.6) Machine learning Algorithm used:

Logistic Regression: This algorithm is used to predict the tone of each and every sentence to get the required output. It is statistical model which uses logics to generate a model on a binary dependent value and in this type of model many complex expression can be solved using regression method by estimating the value and comparing it with the required output. As classification is one of the methods used to classify the amount of data to be used as training and testing data. Under classification there comes logistic regression technique. As in this algorithm we have two type of variables as dependent and independent variables which tell us about the output or response of the model.

In this Algorithm with the help of **SPARSE MATRIX** we have converted string into numbers**.** As logistic regression gives us continuous values as output but we want our

output to be in discrete form so with the help of sparse matrix function we convert the string values in binary form so that our output detects the required amount of values as emotions.

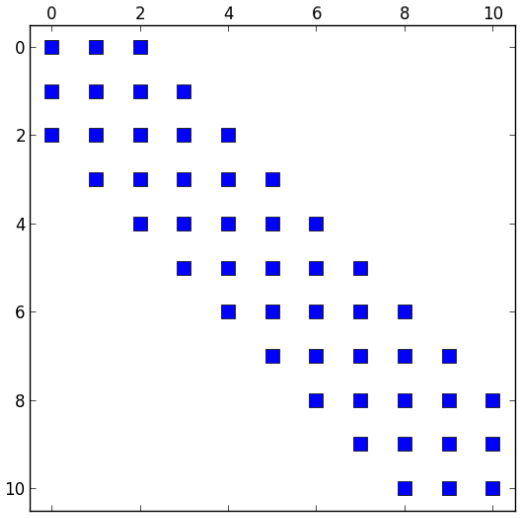


Fig 9:Example of Sparse Matrix

CHAPTER 4:
IMPLEMENTATION AND RESULT

4.1 Test Plan:

1 Any Set:

The tag variable and the range of sequence of variable are manually given an input.

2 Same Keyword/tag however completely different range of tweets:

By giving the same tag variable, person is asked to input a new range of sequence of variables to be searched for using the API.

3 Different Keyword/tag however same range of sequence:

Keeping the range of sequence same as before but changing/using a different keywords/ tags for the given test case.

4.2 Model training:

	A	B	C	D
1		text	emotions	
2	27383	i feel awful about it too because it s my job to get him in a posi	sadness	
3	110083	im alone i feel awful	sadness	
4	140764	ive probably mentioned this before but i really do feel proud o	joy	
5	100071	i was feeling a little low few days back	sadness	
6	2837	i beleive that i am much more sensitive to other peoples feelin	love	
7	18231	i find myself frustrated with christians because i feel that there	love	
8	10714	i am one of those people who feels like going to the gym is onl	joy	
9	35177	i feel especially pleased about this as this has been a long time	joy	
10	122177	i was struggling with these awful feelings and was saying such s	joy	
11	26723	i feel so enraged but helpless at the same time	anger	
12	41979	i said feeling a bit rebellious	anger	
13	2046	i also feel disillusioned that someone who claimed to value the	sadness	
14	98659	i mean is on this stupid trip of making the great album when th	joy	
15	50434	i woke up feeling particularly vile tried to ignore it but it got wo	anger	
16	9280	i could feel the vile moth burrowing its way into my brain seeki	anger	
17	92846	i know its just doing its job and doesnt actually have thoughts c	joy	
18	106363	i wish you knew every word i write i write for you and i think it	sadness	
19	23395	i feel weird knowing mine died when i wasn t around	fear	
20	31583	i feel assured that there is no such thing as ultimate forgetting t	joy	

input_data

Fig 10:Used dataset

1. The tool learns on 50,000 sentences picked randomly out of 4,16,809 sentences in the dataset.

2. The sentences are transformed into a matrix where the column consists of all the processed words occurring in the file and rows contain each sentence.

3. The algorithm used is logical regression.

4.3 Model testing:

1. Each processed sentence was the testing sentence.

2. Accuracy of the model is 82.89% .

4.4 Custom Input:

All the python libraries/files used:

```
from tkinter import *
from tkinter import ttk as ttk
import tkinter as tk
import tkinter.filedialog as fd
import tkinter.messagebox as mb
import pandas as pd
import numpy as np
import re
from nltk.tokenize import sent_tokenize
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from nltk.stem.porter import PorterStemmer
import nltk
import random
from collections import Counter
import matplotlib.pyplot as plt
```

How to tokenize or remove the stop words and to display the information:

```
ps = PorterStemmer()
stop_words=set(stopwords.words("english"))
cv = CountVectorizer(max_features = 4000) #to select top 4000 words most used
reg=LogisticRegression(solver='lbfgs',multi_class='auto',max_iter=1001)
lab=LabelEncoder()

def info():
    mb.showinfo("Info","Please browse a file first")
```

How to generate the model by coping the csv file and fitting the datasets into the file:

```
lef model():
    fh=open("input_data.csv")
    fh2=open("random_data.csv","w+")
    fh2.write("id,text,emotions\n")
    contents=[]
    for line in fh:
        contents.append(line)
    for i in range(0,50000):
        i=random.randint(1,416809)
        fh2.write(contents[i])
    fh.close()
    fh2.close()
    dataset=pd.read_csv("random_data.csv",encoding='cp1252')
    processed_list = []


    for i in range(50000):
        contents=re.sub('@[\w]*',' ',dataset['text'][i])
        contents = re.sub('[^a-zA-Z]', ' ', contents)
        contents = contents.lower()
        contents = contents.split()
        filtered_sent=[]
        for w in contents:
            if w not in stop_words:
                filtered_sent.append(ps.stem(w))

        filtered_sent = ' '.join(filtered_sent)
        processed_list.append(filtered_sent)


    X = cv.fit_transform(processed_list) #convert it in string and store data in X


    y=dataset["emotions"]
    y=y[0:50000]

    y=lab.fit_transform(y) #to  make y as interger type label
    reg.fit(X,y)
```

How to convert the words:

```
def convert_into_words(contents):

    tokenized_text=sent_tokenize(contents)

    processed_list=[]
    for i in tokenized_text:
        con=re.sub('@[\w]*',' ',i)
        con = re.sub('[^a-zA-Z]', ' ', con)
        con = con.lower()
        con = con.split()
        filtered_sent=[]
        for w in con:
            if w not in stop_words:
                filtered_sent.append(ps.stem(w))

        filtered_sent = ' '.join(filtered_sent)
        processed_list.append(filtered_sent)

    X_test = cv.transform(processed_list) #convert it in string and store data in X
    return X_test
```

4.5 Test Cases: -

Tools used :

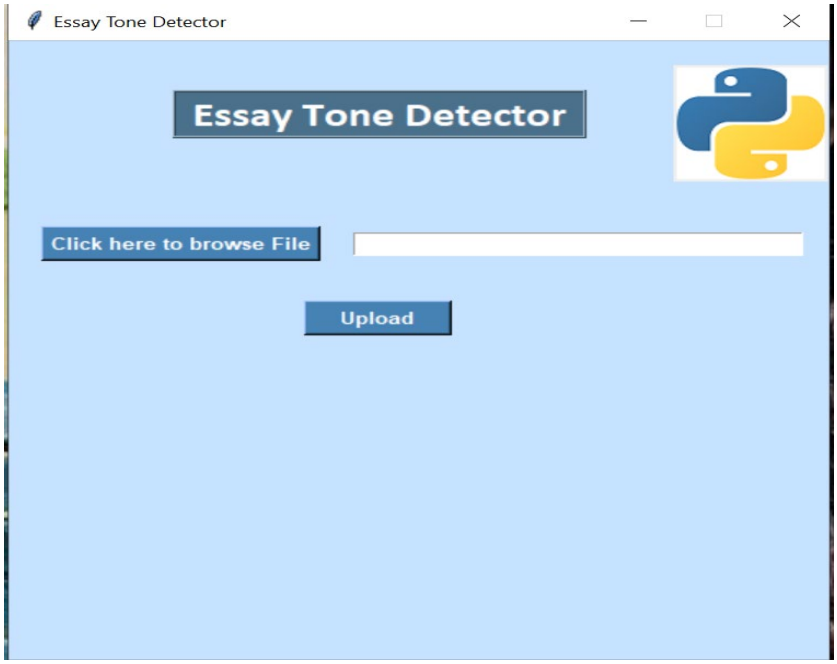
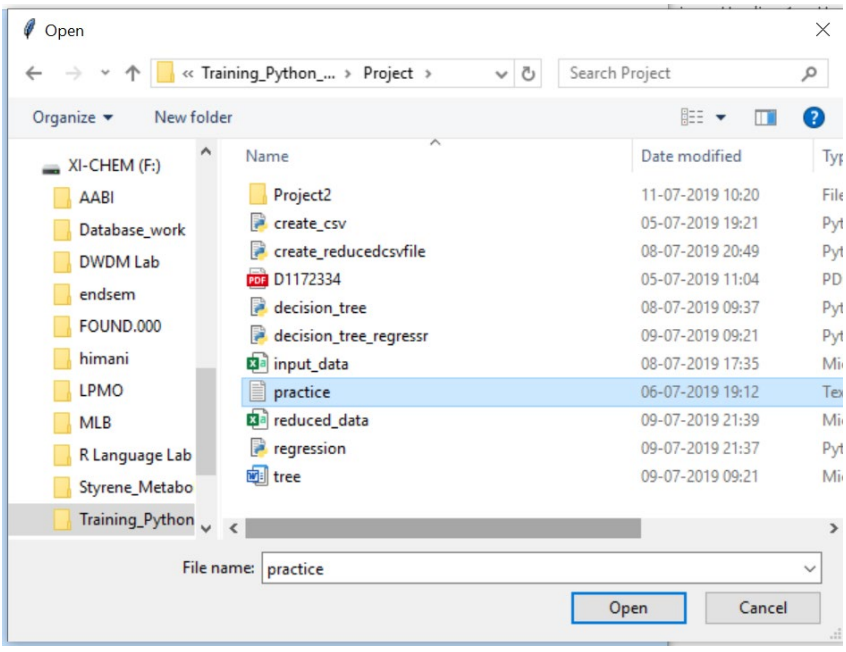


Fig 11: Main Window



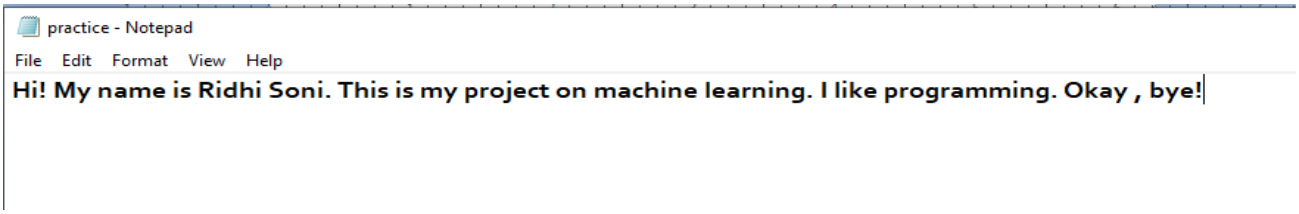


Fig 12: Input

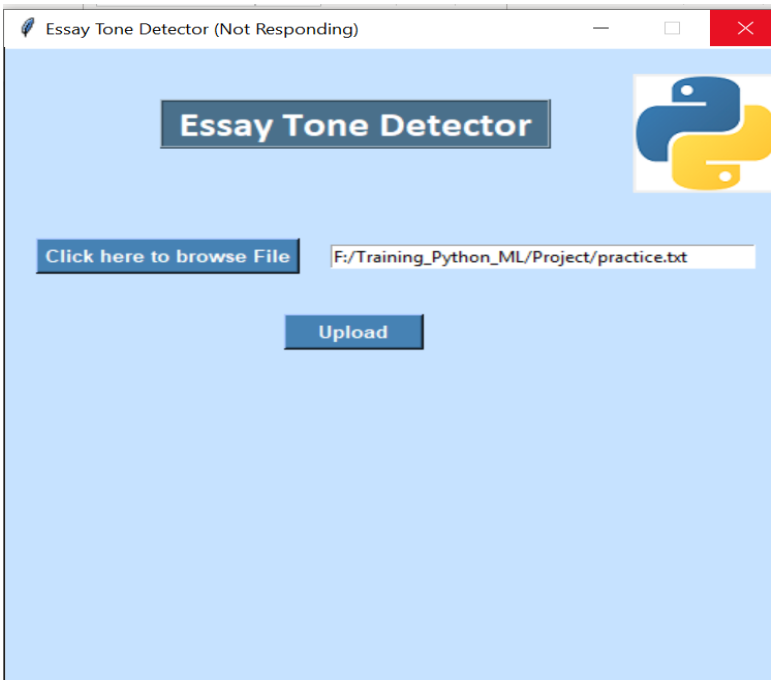


Fig 13: Upload files

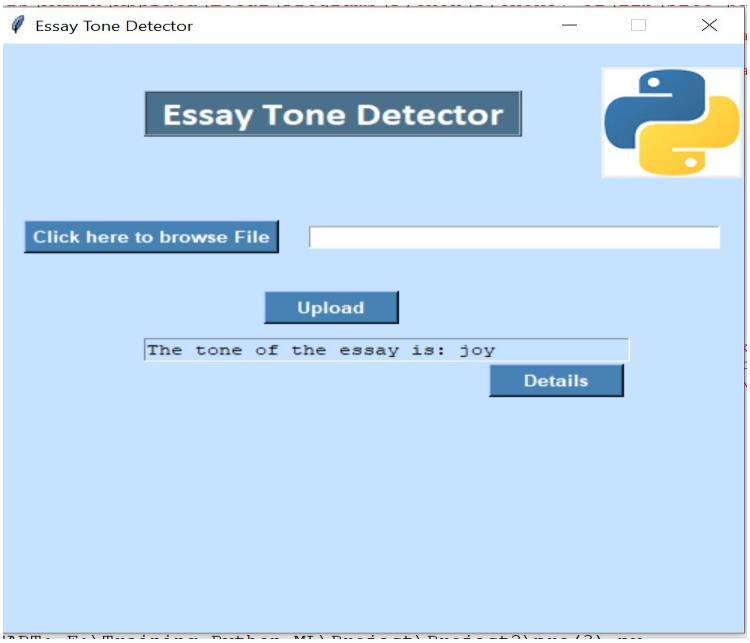


Fig 14: Output

CHAPTER 5:
CONCLUSION

5.1) Conclusions:

Sentiment analysis is a course of research that separates people's feelings, opinions, or emotions towards particular existences. This article throws a critical dilemma of sentiment analysis, sentiment duality categorization — online result reviews selected as data utilized for this research. A sentiment contradiction categorization method has aimed along with detailed explanations of each step. Operations for both sentence-level categorization and review-level categorization have completed

We used sequence API to fetch the tweets and store them in an excel file which can be used for future work. The module called textblob was used to detect the polarity of the tweets and thus providing a very powerful source for sentiment analysis. Sentiment analysis can be used to detect the ongoing trends in the market or even detect the on-going drift about a political party. In this we have done analysis on sequence of word whose limit can be anything by using logistic regression as our tool so as to get the accuracy and getting discrete values as output. Data-processing is the main task of this project which means we have to eliminate all the non required words by doing stemming, tokenization and removing all the stopwords in the sequence of words.

REFERENCES

[1]:Martínez-Cámara, E., Martín-Valdivia, M.T., Urena-López, L.A. and Montejo-Ráez, A.R., 2014. Sentiment analysis in Twitter. *Natural Language Engineering*, *20*(1), pp.1-28.

[2]: Suresh, H., 2016, October. An unsupervised fuzzy clustering method for twitter sentiment analysis. In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 80-85). IEEE.

[3]: Sarlan, A., Nadam, C. and Basri, S., 2014, November. Twitter sentiment analysis. In *Proceedings of the 6th International Conference on Information Technology and Multimedia* (pp. 212-216). IEEE.

[4]: IEEE. Pandarachalil, R., Sendhilkumar, S. and Mahalakshmi, G.S., 2015. Twitter sentiment analysis for large-scale data: an unsupervised approach. *Cognitive computation*, *7*(2), pp.254-262.

[5]: Ramadhan, W.P., Novianty, S.A. and Setianingsih, S.C., 2017, September. Sentiment analysis using multinomial logistic regression. In *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)* (pp. 46-49). IEEE.

[6]: Da Silva, N.F., Hruschka, E.R. and Hruschka Jr, E.R., 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, *66*, pp.170-179

[7]: Troussas, C., Virvou, M., Espinosa, K.J., Llaguno, K. and Caro, J., 2013, July. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In *IISA 2013* (pp. 1-6). IEEE.

[8]: Kumar, A. and Sebastian, T.M., 2012. Sentiment analysis on twitter. *International Journal of Computer Science Issues (IJCSI)*, *9*(4), p.372.

[9]: Bharathi Bhaskaran, R., Prabhakaran, R., Saravanan, S. and Vinoth, M., 2018. TWITTER SENTIMENT ANALYSIS. *International Journal of Pure and Applied Mathematics*, *119*(10), pp.1785-1791.

[10]: Saif, H., He, Y. and Alani, H., 2012, November. Semantic sentiment analysis of twitter. In *International semantic web conference* (pp. 508-524). Springer, Berlin, Heidelberg.

[11]: Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), pp.1093-1113.

[12]: Rosenthal, S., Farra, N. and Nakov, P., 2017, August. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)* (pp. 502-518).

[13]: Danieel Gayo-Ayvello, Panagiiotis T. Metaxas and Eni Murstafaraji 2011, What are the limitations of Electoral Predictions Using Twitter

[15]: Bollen, J., Mao, H. and Pepe, A., 2011, July. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In Fifth International AAAI Conference on Weblogs and Social Media.

[16]: Altrabsheh, N., Cocea, M. and Fallahkhair, S., 2014, November. Sentiment analysis: towards a tool for analysing real-time students feedback. In *2014 IEEE 26th international conference on tools with artificial intelligence* (pp. 419-423). IEEE.

[17]: https://www.nielsen.com/in/en/insights/news/2012/social-media-report-2012-social-mediacomes-of-age.html,[Accessed: May 10'19]

[18]: https://monkeylearn.com/sentiment-analysis, [Accessed: May 10'19]