Social Media Sentiment Analysis- A Relative Study on Twitter Dataset

Siddhartha Sangwan
Computer Science Dept.
KIET Group of Institutions
Delhi NCR, Ghaziabad
2018sidd@gmai.com

Swati Jain
Assistant Professor, CS Dept.
KIET Group of Institutions
Delhi NCR, Ghaziabad
swati.jain@kiet.edu

Keshav Gupta
Computer Science Dept.
KIET Group of Institutions
Delhi NCR, Ghaziabad
guptakeshav59@gmail.com

Sandeep Rao
Computer Science Dept.
KIET Group of Institutions
Delhi NCR, Ghaziabad
sandeep.1822co1051@kiet.edu

2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-8328-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOEI53556.2022.9777217

***Abstract*— Sentiment is an opinion of the person which a person expresses on the particular social media site. To get to know the intentions of the person on a particular thing, calculating the sentimental value of the person's review is must. TSA- Twitter Sentiment Analysis is tool which will be calculating the subjective and polarized value of the particular tweet apart from it, it will also fetch the recently posted tweets on the particular topic and will also calculate their sentimental value which will help to identify on which topic positive or negative tweets are being made along side it will give graphical representation of tweets distribution as well. TSA will also tell the sentimental value of the user enter text regardless of that being a real time tweet on twitter.**

Keywords: Twitter, TSA, Sentiment, Subjective, Polarized

1. Introduction

A tweet which is posted by a person on twitter is a sentiment which will be further classified as a positive sentiment , a negative sentiment or a neutral sentiment. Two types of sentimental values are calculated First is polarized value which ranges from -1 to +1, if the sentence polarized value is negative than there must be some presence of negative words in the sentence, on the contrary if polarized value is positive than sentence must have positive words and if value is 0 than it is a neutral word. Second is subjective value which ranges from 0 to +1, if value is closure to 0 than the tweet is fact otherwise it is personal opinion.

This comes under the NLP which is Natural Language Processing which is done using nltk that is Natural Language Processing toolkit. First the dataset will be processed using the regrex python library which is use to find the pattern in a particular statement, than twitter handles and special characters will be removed from the tweets as it will not add any value to the sentiment than short words will be removed which are less than 2 alphabets as they will also not add any value in sentiment.

Tokenization will be done after above mentioned process. Tokenization is a process of considering each word as a single token as it will form a list of words present in the tweet which will further help to stem the words , stemming refers to reduce a particular word to its stem word after that we will combine the statement of the stem words and form wordcloud which will be positive and negative and than top 10 hashtags of each aspect will be extracted and graph of each will be plotted.

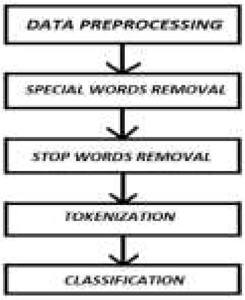


Figure 1: Sentiment Analysis Process

Figure 1 describes the process through which the TextBlob analysis will be done. First of all Data pre processing will be done that is the dataset CSV file will be analysed, than the special words will be removed from the tweet, than stop words will be removed generally stop words are the that do not affect sentiment value these words are less than or equal to 2 in length, Tokenization list is formed each words is considered as single token than classification is done on basis of word.

The main aim of the study is to find Polarized and Subjective values of the tweet posted by a particular person on twitter. Secondly to fetch a limited user entered number of tweets, preprocess them and finding the polarized and subjective value of each tweet using the tweepy library by creating a app and getting the confidential keys from the developer.twitter.com .

2. Related Works

The Dataset used to train the model is labelled Dataset where Label 1 signifies that the tweet is racial and contains hate speech on the contrary Label 0 signifies that the tweet is not racial. To train the model, Dataset is provided in form of csv file which contains more than 30000 tweets along with its id and label. We have taken help from several research papers as well to accomplish the project which are mentioned underneath.

The paper explains that the character limit of a particular tweet on twitter is limited to just 280 characters therefore it becomes very difficult to give a plot of a particular thing in just 280 words therefore people use some essential words and hashtags to express the main motive of the tweet. The main task of the work is to fetch those essential words and to get most out of it using the Natural Language Processing and the final performance is calculated using the f_1 value [1].

The paper gives a brief idea that Twitter is one of the biggest social media site which signifies that there is a ton of data which is posted on twitter everyday as twitter has over 290.5 million users. The data which is posted on twitter is unstructured data that is it is not in specific format and to deal with a huge amount of unstructured data we require machine learning and deep learning techniques [2].

The paper explains that Sentiment is an opinion of person which he expresses on a particular social media site, main motive of the social media sentiment analysis is to fetch the actual expression of the person which is part of the Natural Procession Language . There are various techniques to get the sentimental analysis done such as SVM technique, NB technique, NB Classifier and KNN technique a comparative study is done on all the methods to get the optimal one [3].

The main aim of the algorithm mentioned in the paper process is to get the most pertinent sentences from the native paper and sum up them to form a optimal summary which will further on decrease the redundancy in the data and will increase the efficiency and productivity of the data [6].

The paper explains that there is a fast brodening in the field of mining of ideas and emotion analysis aimed at finding text or ideas that exist in a variety of contexts a machine-learning platform with polarity statistics, emotional analysis and objectivity analysis. Sentimental analysis shows a text organization used to classify feelings expressed in categories in various ways such as confession, admiration, good, unpleasant, negative, etc [4] .

This paper focuses on Sentiment analysis is a broad field of research in the field of education and business. The word sentiment refers to opinion of individuals towards a specific domain. It is therefore also regarded as the idea mining. It leads to straight forward ideas about the domain, not facts. It can be displayed depending on the polarity, update or front with thumbs up and down to. show advantageous as nicely negative feeli ngs respectively. Emotions may be analyzed by the use of NLP, math or system learning techniques [5].

The paper explains that Retailers who sell their product on the online sites asks their customers to review their products in order to get to know about the product if it is good or bad. There can be ample of reviews on a single product. It makes it harder for product manufacturer to keep track of all customers review therefore to tackle this the main work is done in set of three steps: firstly, extracting the key features from review done on products by the customers, secondly, identify the sentences of opinion in each and every opinion also determines whether each sentence statement is true good or bad, thirdly, summarizes results [7].

This paper explains when an earthquake happens, people do a lot of related Twitter posts regarding earthquakes, which makes easy detection of earthquakes occurrence by looking at the tweets As an app, research paper writers developed earthquake clock in system in state of Japan as there is a huge number of twitter users in Japan who tweets when there is a earthquake through which probability of detecting earthquake becomes high by analysing the tweets. The sapplication detects earthquakes quickly and sends emails to subscribers users faster than Japan Meteorological Agency [9].

The main idea of the paper is to dispute that other more formal and for events like sports it is best to use high quality methods to summarize relevant tweets. Research Paper writers officially created the problem of summarizing event tweets and provided a learning-based solution subtle status of event representation via Hidden Markov models [10].

Asmaa Mountass, etal have taken numerous methods for sentiment evaluation with the aid of system gaining knowledge of techniques like Naïve Bayes . Research has made a summary of methods, the discovery of a timeline over the emotional category based entirely on accuracy and ingenuity. The Naive Bayes classifier does not respond to flexible records that provide a more accurate by domino effect. [13].

The paper mentions that at the times of presidential debates twitter becomes the biggest platform for the common people to share their opinion, therefore basically two tasks are performed, firstly extract the subject matter and tweets. Secondly, to fragment the event for which qualitative model called ETLDA was designed which will do the above mentioned tasks [12].

Swati Jain, etal proposed an effective measurement of the ranked pages using the Markov Chain Principle [8]. Akanksha Chaudhary, etal proposed a calculation of Reliability Prediction of Component-Based Software during Interaction [11]. Authors gave complete details regarding all the approaches to social media sentimental analysis explaining their advantage and disadvantages as well[14].

Shilpa C p.c, etal mentioned in research paper that feature extraction produces a vector provided as input to a classification model. Their model showed the best results in the task of classifying emotional Twitter messages[15].

3. Methodology

Algorithm used in the whole process is Logistic Regression. Logistic regression is a Classification approach which uses the Supervised Learning Technique. Supervised Learning technique works on the labelled data that has got a tag or class name to it. Classification refers to the machine learning approach to train with the discreate data For Example: If student is present in the class or not therefore two discrete classes are absent or present.

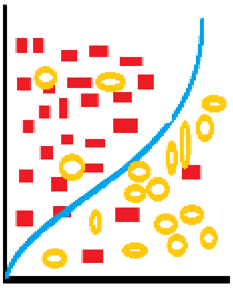


Figure 2: Classification

Figure 2 describes the classification which is a supervised learning technique in which the data is labelled and identifying the category of result is done on the basis of the testing data such as red rectangles and yellow ovals in the above figure are labelled categories.

Logistic Regression is the linear classification technique which works on predicted weight and coordinates

$$f(c)=z_0+z_1c_1+\ldots+z_nc_r \qquad (1)$$

Equation (1) shows the equation for the logistic regression where $z_0,z_1,\ldots,z_n$ are the coefficient where r >=1and c is the function coefficient along which the graph is plotted.

A. Flow Chart

TextBlob is the library of the python language which is use to process the data which in this case is tweets on twitter it is NLP Natural Language Processing which is use to calculate the Sentimental value in terms of Polarity and Subjectivity not just calculating the sentimental value TextBlob is also use for noun extraction, classification and many more things related to nltk libraries. To calculate the sentiment of a particular tweet using TextBlob library we will be using tweet. Sentiment which will give the sentiment value in terms of polarity and Subjectivity.

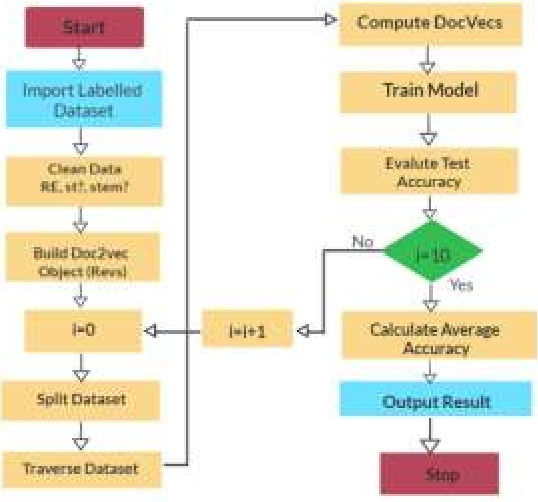


Figure 3: Sentimental Analysis FlowChart

Figure 3 describes the Sentimental Analysis Flow Chart, The first step is to import the dataset by using the pandas library and getting the information regarding it than the data will be preprocessed which is removing of stop words, removing of twitter username and removal of special words will be done.

Initially the value of i is initialized to 0 which is use to calculate the Accuracy of the model after the 10th iteration, then Dataset will be split that is tokenization of the dataset will be done then it will be traversed.

After stemming and evaluating the positive and negative hashtags Model Training will be done and the accuracy score is calculated for 10 times after which the average value of accuracy is taken and the output is obtained.

B. *Step By Step Procedure*

Implementation of code is done in python using jupyter notebook

Step 1: Cleaning Data – Removing all the special characters (!,@,#,$,%,^,&,*) as they will not add value to Sentimental Result.

Step 2: Removing Stop Words - Removing words like he, she, the, was etc.as they will not add any value to Sentimental Result.

Step 3: Tokenization – Separating the words and consider each word as a token.

Step 4: Stemming the words basically getting the base word of a respective word by using the porter stemmer library of python

Step 5: Getting word cloud of the particular labelled data . word cloud might be of positive words, negative words or all words .

Step 6: Extracting the hashtags and getting both positive and negative hashtags.

Step 7: Plotting the graphs of both top 15 used positive and negative words using the matplotlib library and seaborn library in the jupyter notebooks.

4. Result and Analysis

The below mentioned figures shows the result and analysis done on the TSA- Twitter Sentiment Analysis Web Application which is developed using streamlit which gives Bar Graph representation, Pie Chart representation and fetched tweets and there respective sentimental value of the word or sentence entered by user. Here it shows the result of word ‘white’.

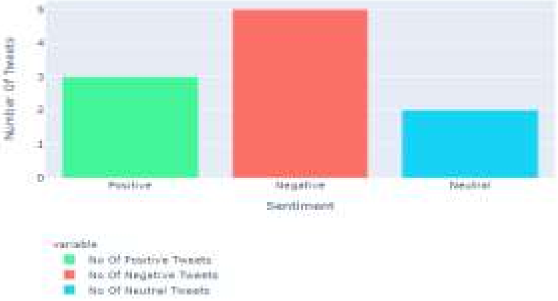


Figure 4: Tweet Distribution

Figure 4 signifies tweet distribution of the word ‘white’, it shows that out of last 10 tweets which has got white in user name or the tweet 5 words which include white in user name or the tweet has a negative polarized value on the contrary 3 has positive polarized and 2 tweets carries neutral value which signifies they are neither positive nor negative.

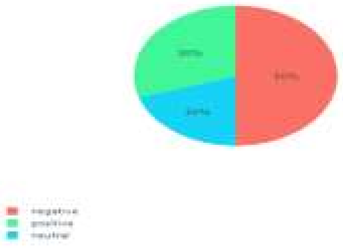
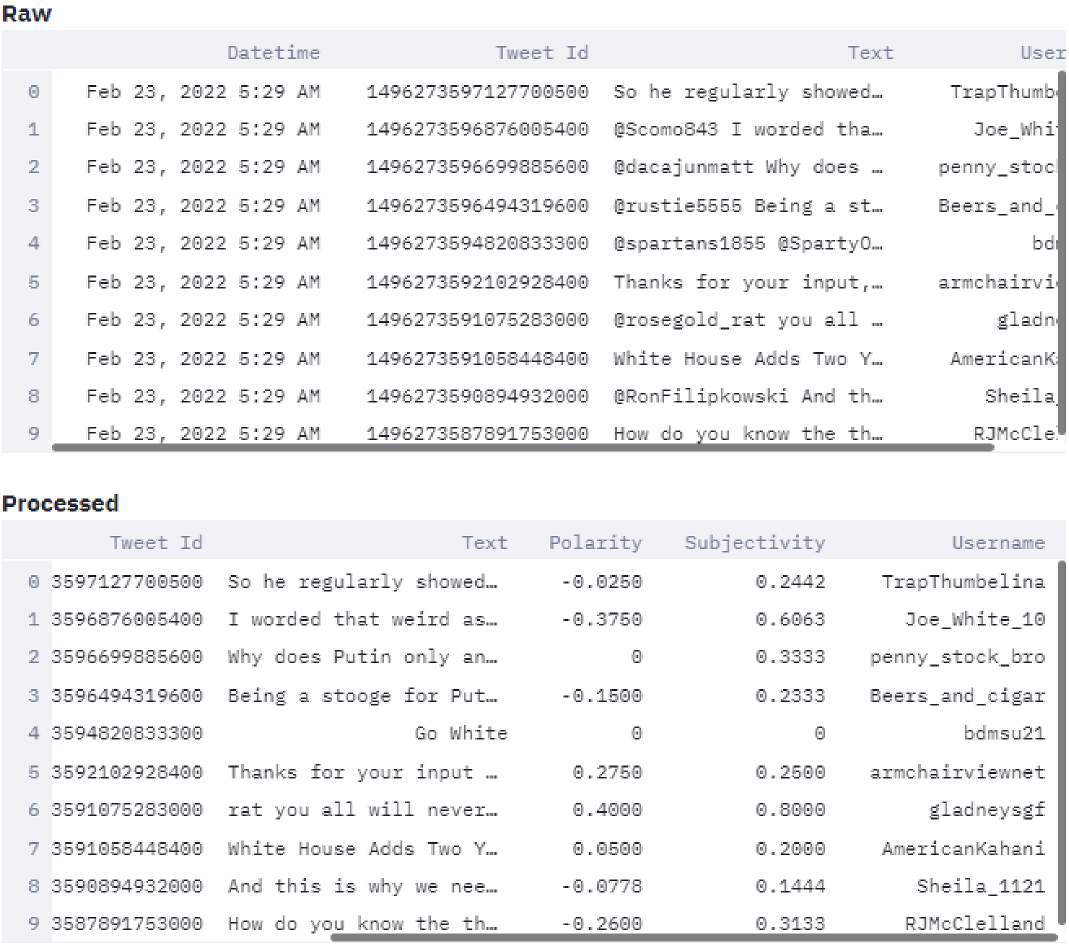


Figure 5: Tweet Propotion

Figure 5 signifies tweet proportion using pie chart of the word ‘white’. It shows 50% tweets have negative sentimental value, 20% have neutral sentimental value and 30% have positive

sentimental value.



Raw

	Datetime	Tweet Id	Text	User
0	Feb 23, 2022 5:29 AM	1496273597127700500	So he regularly showed…	TrapThumb
1	Feb 23, 2022 5:29 AM	1496273596876005400	@Scomo843 I worded tha…	Joe_Whi
2	Feb 23, 2022 5:29 AM	1496273596699885600	@dacajunmatt Why does …	penny_stoc
3	Feb 23, 2022 5:29 AM	1496273596494319600	@rustie5555 Being a st…	Beers_and_
4	Feb 23, 2022 5:29 AM	1496273594820833300	@spartans1855 @SpartyO…	bd
5	Feb 23, 2022 5:29 AM	1496273592102928400	Thanks for your input,…	armchairvi
6	Feb 23, 2022 5:29 AM	1496273591075283000	@rosegold_rat you all …	gladn
7	Feb 23, 2022 5:29 AM	1496273591058448400	White House Adds Two Y…	AmericanK
8	Feb 23, 2022 5:29 AM	1496273590894932000	@RonFilipkowski And th…	Sheila_
9	Feb 23, 2022 5:29 AM	1496273587891753000	How do you know the th…	RJMcCle

Processed

	Tweet Id	Text	Polarity	Subjectivity	Username
0	3597127700500	So he regularly showed…	-0.0250	0.2442	TrapThumbelina
1	3596876005400	I worded that weird as…	-0.3750	0.6063	Joe_White_10
2	3596699885600	Why does Putin only an…	0	0.3333	penny_stock_bro
3	3596494319600	Being a stooge for Put…	-0.1500	0.2333	Beers_and_cigar
4	3594820833300	Go White	0	0	bdmsu21
5	3592102928400	Thanks for your input …	0.2750	0.2500	armchairviewnet
6	3591075283000	rat you all will never…	0.4000	0.8000	gladneysgf
7	3591058448400	White House Adds Two Y…	0.0500	0.2000	AmericanKahani
8	3590894932000	And this is why we nee…	-0.0778	0.1444	Sheila_1121
9	3587891753000	How do you know the th…	-0.2600	0.3133	RJMcClelland

Figure 6: Raw and Processed Data along with polarized and subjective sentimental value of tweets

Figure 6 shows two figures, first figure shows the fetched raw actual data that contains Datetime, tweet Id, Actual Tweet and Username of the user, second figure shows the processed data along with the Polarity and the Subjectivity Sentimental Value of the user. Processed data is the data from which tweet handles and special characters and words are removed.

Table 1: Accuracy Score of the Different Dataset Csv Files

SNo.	Dataset	Average Accuracy Score
1.	Dataset A CSV File	*0.9469403*
2.	Dataset B CSV File	*0.9487148*
3.	Dataset C CSV File	*0.9483999*

Table 1 shows the accuracy score of the several Dataset CSV files which is calculated after model training for 10 times and the average value of each Dataset is taken into consideration.

The Accuracy score of the model on the base Dataset A calculated using the sklearn library of python by importing Logistic Regression is 0.9469403.

In order to support our model we also calculated the Accuracy score of two more Dataset labelled as Dataset A and Dataset B whose Accuracy Score comes out to be 0.9487148 and 0.9483999 respectively which signifies the best model accuracy is ~95.

5. Conclusion

The Research Paper shows the Sentimental Analysis including polarity and subjectivity done on more than 30000 tweets of the dataset . The final result is calculated by using the Average Accuracy score, the score is a result of average of 10 accuracy scores done on three dataset Dataset A(base Dataset), Dataset B and Dataset C it signifies the best model accuracy is ~95. Monitoring the twitter defines that how a particular person is thinking about several thing which gets us to know more about the outcomes of several things like election, products, equipment and etc.

TSA will be applicable to analyse the Polarity and Sentimental values of a tweet or custom written tweet. Positive value will signify good tweets, negative value will indicate negative tweet and zero will show neutral tweet. It will also fetch the polarity of top tweets when user will enter a particular word or sentence in the user interface and will give a statical representation of it and it will also show the related tweets.

6. Future Work

In future ,after analysing the tweets social media sentiment analysis will be applicable of reporting the tweets and post, comments of FB that contains the hate speech regardless of them being racist and sexist and will also take actions to remove, suspend or permanently delete accounts of those users.

The social media sentiment analysis will be applicable to check the positive and negative reviews of the product on Amazon, Flipkart, Snapdeal etc. It will also get a check on hatred speech on Facebook and Instagram and analyse posts on other social media sites as well. It will also give analysis of YouTube comment section and will get hatred speech out of it.

7. References

[1] Tharindu Weerasooriya, Nandula Perera, S.R. Liyange. A method to extract essentialkeywords from tweet using NLP. 2016 16th International Conferences on Advances in ICT for Emerging Regions (ICTer).

[2] Adyan Marendra Ramadhani, Hong Soon Goo. Twitter Sentiment Analysis using Deep Learning Methods. 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017.

[3] Devika M D, Sunitha C, Amal Ganesh "Sentiment Analysis: A Comparative Study on Different Approaches", Procedia Computer Science 2016, vol.87, pp 44-49.

[4] Dr. Pamela Vinitha Eric, Anu Priya KR. Twitter Sentimental Analysis using Deep Learning Techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology ISSN : 2456-3307. 2020.

[5] Dattu, Bholane Savita, and Deipali V. Gore, "A survey on sentiment analysis on twitter data using different techniques", International Journal of Computer Science and Information Technologies 2015, vol. 6, no. 6, pp. 5358-5362.

[6] X. W. Meng Wang and C Xu. An approach to concept oriented text summarization, Proceedings of ISCTTS05, IEEE International Conference, China, 1290-1293" 2005.

[7] M. Hu, B. Liu,"Mining and summarizing customer reviews,"Proc. 10th ACM SIGKDD, Washington, DC, USA, (2004).

[8] Swati Jain and Mukesh Rawat. Efficiency measures for ranked pages by Markov Chain Principle. International Journal of Information Technology ISSN 2511-2104.

[9] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors, in Proc. 19thInt. Conf. WWW, Raleigh, NC, USA, 2010.

[10] D. Chakrabarti and K. Punera, Event summarization using tweets, in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[11] Akanksha Chaudhary and Pradeep Tomar. Reliability Prediction of Component-Based Software during Interaction. Journal of Software Engineering Tools & Technology Trends. 2016.

[12] Y. Hu, A. John, F. Wang, and D. D. Seligmann, Et-lda: Joint topic modeling for aligning events and their twitter feedback, in Proc.26th AAAI Conf. Artif. Intell. Vancouver, BC, Canada, 2012.

[13] Asmaa Mountassir , Houda Benbrahim, Ilham Berrada, An empirical study to address the problem of unbalanced data sets in sentiment classification,IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea.

[14] Prajval Sudhir, Varun Deshakulkarni Suresh. Comparative study of various approaches, applications and classifiers for sentiment analysis. Global Transitions Proceedings 2021.

[15] Shilpa C p.c, Rissa Shereen, Susmi Jacob, P.Vinod Sentiment Analysis Using Deep Learning. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). 2021.