ISSN No:-2456-2165

Segregating Tweets Using Machine Learning

I. Deva Prasad [1], G. Hemanth Kumar [2],
Student, UG Student,
Dept. of Computer Science, Sathyabama University,
Chennai, India.

Dr. T. Sasikala., M.E, Ph.D.,[3]
Dean, Dept. of Computer Science,
Sathyabama University, Chennai, India.

Abstract:- The size of informal organization information is expanding rapidly. All the different types of issues and problems are communicating in online social media platforms. This project tells us how to find out different types of issues going in social media. We all know that twitter is one of the social media platforms. So, Twitter is picking up prominence these days and the vast majority are utilizing this stage to communicate their conclusions. Slant investigation on Twitter is the utilization of breaking down the supposition of twitter data passed on by the client. The examination on this issue proclamation has developed reliably. So, this project aims to collect the all twitter data which is posted by the public users with organization hashtags. By using Sentiment analysis we will find out the tweets which is a positive comment and which is a negative comment and which is neutral. Right now, plan to portray the systems embraced, the procedure and models applied, alongside a summed-up approach utilizing python. Conclusion examination expects to decide or quantify the frame of mind of the essayist regarding some subject.

Keywords:- *Informal Information, segregating Tweets, Evaluating data, Text analysis, Hashtag analysis.*

I. INTRODUCTION

The way we live, experience and express ourselves has changed dramatically over the past decade. Social mediums significantly changed how we communicate and talk to one another. We are in an era where we can communicate our ideas to the whole world in just one shot. This is possible because of social media. Majority of us, in fact almost all of us use social networks every day to let our family and friends know what we are up to and try to be in touch with them and talk to them through those media. We express all the time. We want others to know what we are doing in our lives and we want to know what others are up to. We are habituated to express so much on social media that sometimes we express our emotions and feelings without prior thinking of the consequences that it will bring upon us. There is every chance that sometimes what we express and post can be easily misunderstood and trigger other groups of people leading to social disturbances. Out of many available social networking sites, Twitter is one of the most used social media for such purposes.

Such tweets sometimes have the potential to cause legal issues and can cause reputation damage as well and no one can go through all the tweets of users to filter and sort them out as whether they are of any disturbance to anyone or not. So we use sentiment analysis on twitter data and find out tweets of users that can cause any social disturbance and can report directly to local police thereby trying to establish social stability.

II. EXISTING SYSTEM

Currently there is a certain framework for how the process of sentimental analysis takes place. Initially the static information that is available is separated from a web-based life stage. The information that is extricated is now stored in a csv document or excel sheet which is then used for the program or application. Once after the data stored in excel sheet is processed then the filtration takes place. They are classified into positive, negative and nonpartisan. An extremity number exists that extends from -1 to +1. Using the extremity number the program or application decides the feeling of the post or the announcement.

- The feeling is named positive, negative, neutral.
- If polarity>0 , it means positive,
- If polarity= 0, it means neutral ,
- If polarity<0, it means negative.

➢ *Disadvantages:*

- The client who is dissecting the announcements needs to experience the whole archive (csv record) to get by and large broad report.
- The feelings are arranged uniquely into three classifications.

III. PROPOSED SYSTEM

Proposed system manages all types of performing limits very effectively via online electronics i.e., TWITTER. AS twitter will be having different types of posts which forms a huge database. Tweets considering the small messages or posts will be in different types of languages, slang and also consists of wrong spells. Here, sentence formation level estimation assessment will be done. This has to be conceivable in 7 different phases.

➢ Phase 1: Data has to be given as input for the analysis process. Here, data indicates that will be username or it may be with hashtag.
➢ Phase 2: And then, data with a huge amount will be shown which are to be analysed.Those are the tweets which are recouped from the database of the twitter.
➢ Phase 3: By then in the third phase, those recouped data from twitter will be taken care in a data set.
➢ Phase 4: Tweets has to be dealt with it. This movement is performed before feature extraction. Getting ready advances consolidate clearing the URL, ousting stop-

ISSN No:-2456-2165

words, keeping up a vital good ways from mis-spells and different types of slang. Mis- spellings of data square will be avoided by supplanting constants with 2 occasions.

- Phase 5: Different of slang will contribute rich to tweet's opinion.
- Phase 6: Consequently, words of slang jargon will be kept to switch the slang words which occurs in tweets along with their related ramifications.
- Phase 7: Next phase is the extraction of the feature. A vector of segmentation is formed maltreatment significant other choices.

A. Default Out Come of Existing System:

- Having the resultant in the form of bar graph and pie chart is very understandable .
- There are the seven different categories which understands very well on the emotions of the tweets. They are 1. strongly positive, 2. positive, 3. weakly positive, 4. neutral, 5. strongly negative, 6. negative, 7. weakly negative.
- Instead of putting away the information preceding the examination, we can get constant information by giving the data of hashtag or username from twitter to break down the tweets of an individual or a predefined hashtag. The proposed framework conquers the disadvantages of the overarching framework Now, it will be on our instructing set at that point, it is must to remove supportive alternatives from data set which may be used in the strategy for order. Let it get the examine some of matter organization strategies which can help America in highlight extraction.
- **Tokenization**: It is the method to break up the flood of substance in to a different words, pictures as tokens. These tokens will be disengaged with the white-spaces and also character will be em-phased. Tokenization will perform the investigation of tokens to different parts to form a tweet structure.
- URLs and client references are evacuated if client is keen interested on just examining the content of the tweet.
- Punctuation of data will prints and number might expelled out.
- Conversion to Lowercase: Tweets might be standardized by changing over it to lowercase which it makes the examination with an English lexicon simpler.
- Rejection of stop words: In sentiment analysis some words are rejected when there is no meaning in it like is, it, they, like, was, there because there i no meaning in it.

IV. IMPLEMENTATION

Enter the data of the username or the hashtags and also the number of tweets which you want to dissect the tweets. Here are the stages of the implementation with appropriate result.

Stage 1: Tweet cursor will extract the tweets from the tweet bundle with the given inquiry word.
Stage 2: These extracted words will be given with the tokenization and then get cleaned where the accentuation mark, emoji's, URLs, extra stop-words will be expelled out.
Stage 3: The above stages are which have been sent to the Naive Bayes Classifier for the investigation purpose.
Stage 4: Thereafter, classifier will do the processing by highlighting and then pulled out the polarities to the each and every one which running from -1 to 1.
Stage 5: Now every extremity will be given with particular tweet net extremity will utilized to arrange tweets into its particular classification.

Result: The final result will be as expected as grouping of percentage of the given tweets in specific classification and also the pie chart will be delineate all classes that have arranged.

- *Implementing The Classifier:*

In implementing the classifier we have used is the NAIVE BAYES CLASSIFIER. It explains S is the sentiment and M is the messages. And also data set with video games audits are utilized. Separating all the surveys by the resultant score and isolating the models similarly between the scores -1 and 1.

- *Making Data Ready:*

It will set up the information by the tokenization, along with the cleaning of the information and then move to Bag-of-words. Then do the creation of the element vector for each of the record. All the words will tally with the recurrence of many of times token has been showed up in each archive. Here the number of segments will tokens that novel in assortment of archives. Number of lines are the complete reports of the entire assortment now changing.

- *Building:*

For the recognition of the patterns which are used for the classification and regression analysis, Naive Bayes Algorithm is used. The extracted data will get pre-processing and then classified into keyword related tweets. Groups will be classified and predicted by the keywords according to the user group. Under theese classification, same type of the clusters will be divided into groups. A negative classification will be formed for all the negative tweets by clustering. And positive classification will be formed for all the positive tweets.

ISSN No:-2456-2165

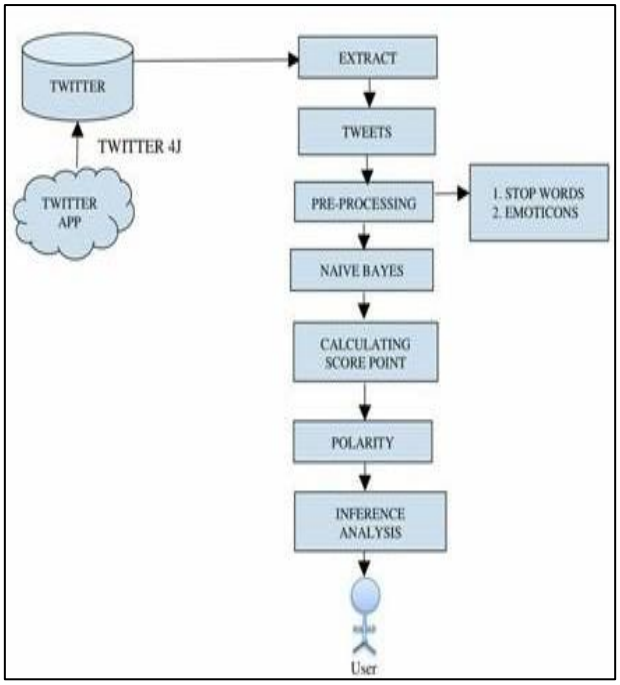

Fig 1:- architechture

V. RESULTS

When we enter the require organization Hashtag then we can see the tweets in the twitter and then we can calculate the feelings of the tweets by using text analysis. Finally, after calculating all the feelings of the tweets then tweets are displayed in the screen for the particular hashtag along with that at the end it will show this tweet is positive or negative. If positive or neutral it shows "0" if Negative it shows "1".after that all the data is stored in the database separately.

Overall score point: Topic #iphone10	
Overall Polarity	Calculation
Mixed	5.436573%
Polarity	Calculation
Positive Polarity	5.107084%
Negative Polarity	2.306425%
Neutral Polarity	1.8121911%

Table 1:- Different types of Polarity and Calculation

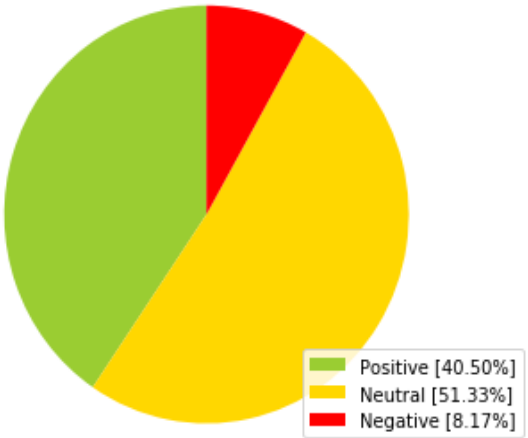

Fig 2:- The relation between the positive, negative and neutral tweets percentage of people's thoughts.

VI. CONCLUSION

If the public tweet anything in twitter by using the organization username or organization Hashtag we can easily find the tweets and we can easily extract the tweets. Before evaluating the tweets we need to check the internet connectivity because the entire project is totally based on internet and social media. We will apply Naive Bays Algorithm to classify the tweets into two categories i.e. which is positive and Neutral or Negative. After evaluating the tweets we can segregate into three categories and stored in data set like all the positive or neutral tweets in one block and Negative tweets in another block.

VII. FUTURE WORK

The way we live, experience and express ourselves has changed dramatically over the past decade. Social mediums significantly changed In Future I am expecting to do this Sentiment Analysis using Social Media Data like Instagram, Facebook, LinkedIn, Tiktok etc... With other languages except for English and also I will try to implement to find out the username along with the tweets and also I will try to find out the different types of tweets for different organizations at same time.

ACKNOWLEDGMENT

We are very thankful to our guide Dr. T. Sasikala M.E., Ph.D. Dean, school of computing for her valuable, suggestions and continues encouragement for this work. We also convey our sincere thanks to Dr. S. Vigneshwari M.E., Ph.d and Dr. L. Lakshmanan M.E., Ph.D, Head of the department, computer science and engineering, for providing full support at the time of reviews. And i also thank for all the teaching and non teaching staff of the department for their support.

REFERENCES

[1]. H. Zang, “The optimality of Naïve-Bayes”, Proc. FLAIRS, 2004
[2]. C.D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”,Cambridge University Press, pp. 234-265, 2008
[3]. Vigneshwari, S., Bharathi, B., Sasikala, T., Mukkamala, S. A study on the application of machine learning algorithms using R, Journal of Computational and Theoretical Nanoscience, 2019, Journal of Computational and Theoretical Nanoscience, 16(8), pp. 3466-3472
[4]. M. Schmidt, N. L. Roux and F. Bach, “Minimizing finite Sums with the Stochastic Average Gradient”, 2002
[5]. Y. LeCun, L. Bottou, G. Orr and K. Muller, “Efficient BackProp”, Proc. In Neural Networks: Tricks of the trade 1998.
[6]. T. Wu, C. Lin and R. Weng, “Probality estimates for multi-class classification by pairwise coupling”, Proc. JMLR-5, pp. 975-1005, 2004
[7]. “Support Vector Machines” [Online], http://scikit-learn.org/stable/modules/svm.html#svm-classification , Accessed Jan 2016
[8]. P. Pang and L. Lee, “Opinion Mining and Sentiment Analysis. Foundation and Trends in Information Retrieval”, vol. 2(1-2), pp.1-135, 2008
[9]. E. Loper and S. Bird, “NLTK: the Natural Language Toolkit”, Proc. ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics ,vol. 1,pp. 63-70,
[10]. H. Wang, D. Can, F. Bar and S. Narayana, “A system for real-time Twitter sentiment analysis of 2012 U.S.presidential election cycle”, Proc. ACL 2012 System Demostration, pp. 115-120, 2012
[11]. Hamid Bagheri, Md Johirul Islam,“Sentiment analysis of twitter data”, Annual International Conference “Dialogue”(2017) (pp. 14-28)
[12]. David Zimbra, M. Ghiassi and Sean Lee, “Brandconnected Twitter Sentiment Analysis mistreatment Feature Engineering and therefore the Dynamic design for Artificial Neural Networks”, IEEE 1530-1605, 2016.
[13]. Bhumika Gupta, Monika Negi, Kanika Vishwa, “Study of Twitter Sentiment Analysis mistreatment Machine Learning Algorithms on Python” International Journal of laptop Applications 0975-8887, 2017
[14]. Aliza Sarlan, Chayanit Nadam and Shuib Basri, “Twitter Sentiment Analysis”, 2014 International Conference on data Technology and Multimedia (ICIMU), Putrajaya, Malaya Gregorian calendar month eighteen – twenty, 2014.
[15]. Alexander Pak, Apostle Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, Proceedings of the International Conference on Language Resources and analysis,
[16]. LREC 2010, 17-23 could 2010, Valletta, Malta,2010 7. Mining Twitter information with Python (Part half-dozen – Sentiment Analysis Basics), marcobonzanini 2015.
[17]. Naive Bayes for Sentiment Analysis, MediumCorporation,2018.