2023 International Conference on Computer Communication and Informatics (*ICCCI*), Jan. 23 – 25, 2023, Coimbatore, INDIA

EMOTION ANALYSIS OF TWEETS

Durga Prasad D
Electronics and Communication Engineering
Vishnu Institute of technology
Bhimavaram, INDIA
durgaprasad.d@vishnu.edu.in

Pardhu Guttula
Electronics and Communication Engineering
Vishnu Institute of technology
Bhimavaram, INDIA
19pa1a04c7@vishnu.edu.in

Tangellapalli Sai Ramya Manasa
Electronics and Communication Engineering
Vishnu Institute of technology
Bhimavaram, INDIA
19pa1a04f5@vishnu.edu.in

Vemulapalli Bavya Sri
Electronics and Communication Engineering
Vishnu Institute of Technology
Bhimavaram, INDIA
19pa1a04h3@vishnu.edu.in

Thote Prashanth
Electronics and Communication Engineering
Vishnu Institute of Technology
Bhimavaram, INDIA
19pa1a04g1@vishnu.edu.in

Pitani Niranjan Sai
Electronics and Communication Engineering
Vishnu Institute of Technology
Bhimavaram, INDIA
19pa1a04d2@vishnu.edu.i

***Abstract*—As technology is developing day by day messages, social media platforms became a platform of representing person's life, thoughts and their behaviour. We all know this developing technology has more freedom to share their thoughts, personal feelings, what they are feeling at the current moment within a single second, within a single sentence on social media. There will be plenty of data available to know about a person behavior and this data can be used for detecting various emotions. Our project Emotion Analysis of tweets itself says that to detect a particular emotion held by a tweet. That is done by analyzing expressed opinions, images, sentiments, and other activities. In our case we took mainly posted opinions or tweets as input to implement our NLP (Natural Language Processing) model. This project is about constructing machine learning models to find which model is predicting emotion of tweets with highest accuracy.**

***Index Terms*— Twitter, Sentiment, Social media, natural language processing**

I. INTRODUCTION

Twitter is a person to person communication site introduced to the world in 2006, without any doubts it became the most well known social medium stages accessible today, with nearly hundred millions of users active everyday and approximate five hundred tweets were sent day by day. Twitter can be used inconceivably. It tends to be utilized to get news, follow high-profile big names, or keep in contact with old secondary school companions. Twitter initially started as a SMS-based platform, so the particular message has character limit of one hundred and forty as a necessary condition. As day by day twitter platform developed to turn into a web stage, they kept the cutoff essentially in light of the fact that it lined up with Twitter's image. Twitter itself becomes a platform that plans to make exceptionally short content for tech-weighty, by considering deficiency present day in world. Twitter has developed dramatically over the course of last 15 years.The main motivation of twitter is to spread data at quicker rate irrespective of seriousness of content or information.

The tweets posted by people can be photos or videos or any links of particular website etc. People or users can control the content displayed on their screens by following their liked persons, liked topics and liked organizations etc. Users can also see the re-tweets from individuals they follow and advanced tweets which are paid notices. To assist with restricting phony records, Twitter made the confirmed record image, which shows the record is genuine and has a place with the individual or organization. For the particular account being checked, Twitter affirms the identification of particular individual or an organization. This feature in twitter keeps up with entrust with public or clients or users.

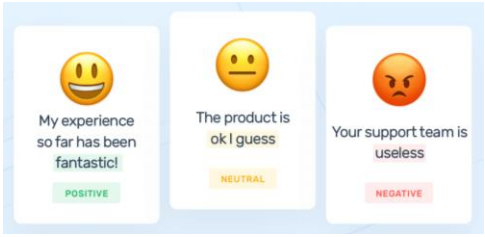
This twitter can easily figure out the thing or any particular content being highlighted by most of the people or get viewed by most of the people. In twitter application a search bar is available in that search bar user can find or search a person's account, particular trending topic etc. Twitter has an explore function to find trending subject around us. This can be done by adding "#" symbol in the starting of a word being searched then it provides the tweets or topics related to that #word. At the point when people post a tweet in their accounts these posts can be seen by their followers' feeds. These tweets consists of news, jokes or sharing articles and can likewise be looked through on Twitter. But the length of posting a tweet or any character content is restricted to one hundred and forty characters. The breaking point of character count is currently two hundred and eighty characters, including punctuations, whitespaces and special characters. To mesh tweets into a discussion string or interface them to an overall point, individuals can add hash tags to a watchword in their post. This hash tag behaves like a meta tag, is communicated and can be searched under "#keyword".

II. EMOTION ANALYSIS OF TWEETS

Emotion analysis is the most common way of recognizing and breaking down the hidden emotions or feelings communicated in literary information. This analytics is used to draw the text information available from different platform sources which helps to investigate the abstract data and figure out the feelings that a particular sentence is holding. High level machine learning procedures or techniques assists us in detecting emotions of a particular tweet.

Nowadays this emotion analysis has turned into a vital

component in progressing one's own business. Because if a person bought a product and he is fully satisfied with the working condition of a product and services provided by the company then he normally provides feedback to that organization regarding everything, then this feedback will be taken as input by the organization and concentrate on this feedback factors and try to develop some extra services and develop features to the product. Similarly for the case of a person not satisfied with product and services also, through the user's feedback organization provides implementations in services. Since this whole process is related to single person it will be easy to detect a person thoughts on the services provided by the organization. But what will be the process in case of thousands of reviews. So our project helps to find which type of emotion a particular review or tweet is being held.



1. INTRODUCTION TO PROBLEM

Though many software techniques available to know and extract data about a person's emotional opinion on a specific product or service through the feedback provided by them, still many organizations and data analysts facing issues for extracting the data from the provided feedback by people. If we want to detect or find a particular emotion is holding for a particular tweet it is very easy for a single tweet but with the increasing growth in the World Wide Web usage, people are habituated to use social media platforms such as Twitter and other platforms results in generating large volumes of data and various people opinions in form of tweets or comments then in this case it will be very difficult to know a particular emotion that a particular tweet holding. This problem requires an organized manner solution.

So our project converts this huge amount of data of tweets into organized manner that a machine can easily understand by using Natural Language Processing techniques and finally we are going to develop a machine learning model with best accuracy which is useful for calculating emotions of tweets.

2. TOOLS AND ALGORITHMS

A) Natural Language Processing (NLP)

NLP methods depend on machine learning and particularly factual realizing which utilizes a general learning calculation joined with an enormous example, a corpus, of information to get familiar with the standards. Detecting emotions of particular tweets has been taken care of a Characteristic Language Handling meant NLP, at many degrees of granularity. Beginning from being a report level grouping task, it has been dealt with at the sentence level and all the more as of late at the expression level. NLP is a field in software engineering which includes causing PCs to get significance from human language and contribution as an approach to collaborating with this present reality.

B) Support Vector Machine

Support vector machine is a directed learning framework and utilized for characterization and relapse issues. Support vector machine is very preferred by a lot of people as it produces prominent rightness with less calculation power. It is generally utilized in order issues. As we all know there are three kinds of machine learning algorithms. They are supervised algorithm, unsupervised algorithm and reinforcement learning algorithm. This Support Vector machine algorithm comes under supervised machine learning model. A help vector machine is a particular classifier officially characterized by partitioning the hyperplane.

Given marked preparing information the calculation yields best hyperplane which characterized new models. In two-layered space, this hyperplane is a line parting a plane into two sections where each class lies on one or the other side. The aim of this algorithm is it tries to fit the hyperplane between different classes that maximizes the distance from hyperplane line to points of classes.

C) Random Forest

In our project we use another machine learning technique is random forest. Random forest algorithm follows voting procedure. Multiple decision trees combine to form tree type structure and this is known as Random Forest. The main aim of this technique is to predict which type of data gets highest voting. The highest voted data is shown to us as an output. It uses outfit realizing, which is a strategy that joins numerous classifiers to give answers for complex issues. This calculation comprises of numerous choices of decision trees.

This calculation lays out the result in view of the forecasts of the choice of decision trees. It predicts by taking the normal or mean of the result from different trees. Expanding the quantity of decision trees builds the accuracy of the result in efficient manner.

III. CASE STUDY

The main aim of our project is to find the emotion of particular tweet and then we develop the best suitable machine learning algorithms test the module based on the accuracy. The flowchart is as follows:

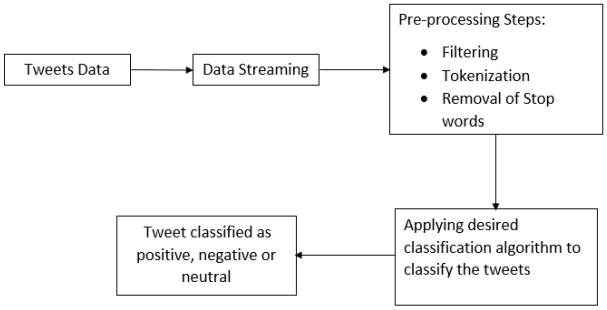


Figure.1 Flowchart for classification of tweets

1. Collection of Data

Our project requires huge amount of data consists of tweets. It will be difficult for us to collect or gather data consists of tweets through manual work. It may be found as time consuming process and also it may be found as collection of data by our own is treated as not that much important in this present generation. As the present world have lots of technologies which are developing or improving day by day it doesn't make sense to collect data by our own waste our time on it. So to make our work easier this data is taken from available online resources and tried to observe the above few tweets corresponding to each labels, it is safe to conclude that label 0 is for non-offensive tweets and label 1 is for offensive tweets. After observing data we applied data preprocessing and cleaning techniques used in NLP. This data is raw data and it not in organized manner. We apply some data filtering techniques to delete the unnecessary content from our data in the following steps.

2. Data Streaming

It is the process making data in form of organized manner which means in this process the derived words or sentences would diminish to their root forms. For example, a sentence having a word giving now this stemmer will reduce the phrases "give", "given", "giving" to root word "give". The main benefit of stemming is that it makes examination between the words which are less complex because we don't have to manage complex syntactic changes of the word. For our project we make use of Porter Stemmer algorithm on our data consists of tweets. In this data streaming processing of data takes place. This processing of data includes tokenization which is the method involved with dividing the tweets into individual words which are called as tokens. These tokens are split when a sentence contains a whitespace or any punctuations. For example, the tweet is like "Such an amazing climax!!" is split into different words from each whitespace as follows:

{
Such
an
amazing
climax

!!
}

3. Filtering of Data

The tweets obtained after processing of data still the data contains information which is not in an organized manner or it may contain noise in the data i.e raw data. This data found to be not useful to our project. So to remove the unnecessary information from our tweets our data need to be operated by further more improving techniques to obtain the data in desired and simple manner by processing it. In this filtering process the tweets are filtered for the removal of noise in it by using techniques of removing of stop words, characters which does not make any value to our data and any other punctuations, numbers which are not necessary.

This filtering process contains removal of any special characters, words which does not add any value to sentence etc.

Stop words: After tokenization and stemming processes tweets may contain some stop words which looks like incredibly familiar words like "the", "is", "a", etc and these words does not hold any extra data. So these words are treated as unimportant and finalize them as they are not serving any purpose and this component is carries out utilizing a rundown put away in stopfile.dat. Then we contrast each and every word in tweets and this rundown can erase the words which are matching the list of words which does not add any value to a tweet.

Removing non-alphabetical characters: The symbols for example "!#^&*$%@" and numbers which does not holds any kind of information are treated as useless and we try to remove these symbols from our data so it makes our work easer to perform further more operations on data to find emotion of tweets. Ordinary expressions like we use in our daily life and add some importance to the sentence due to their presence will be remained without removing and remaining characters or numbers or symbols are removed. This assists with diminishing the messiness in our tweets data.

Stemming: It is the process making data in form of organized manner which means in this process the derived words or sentences would diminish to their root forms. For example, a sentence having a word "fishing" now this stemmer will reduce the phrases "fishes", "fishing" to root word "fish". The main benefit of stemming is that it makes examination between the words which are less complex because we don't have to manage complex syntactic changes of the word. For our project we make use of Porter Stemmer algorithm on our data consists of tweets. In this data streaming processing of data takes place

4. Applying Algorithms

After analyzing data we will implement some machine learning algorithms such as Support Vector Machines (SVM) and Random Forest algorithms. By using these techniques we would attempt to fit our model in the best suitable algorithm. This best algorithm can be identified by observing F1 score of each algorithm. We will come to conclusion by observing the maximum F1 score of a particular algorithm means if an algorithm has higher F1 score compared to another algorithm, we can say that algorithm provide very accurate result.

The data is in need of adding labels to unlabeled information for the detailed study of emotion of tweets, this can be done by using "vader lexicon" model. This model treated as one of the best approaches for the detailed analysis of emotions of tweets**.** The vader lexicon can be accessed using Natural Language Toolkit library in python. The necessary libraries in python and the unlabeled data which is in need of adding labels imported to our working environment. Since our taken dataset contains two columns, we try to add labels for the unlabeled data. After adding labels to our tweet information then we try to add four more new columns to the labeled data. These four columns are "Positive", "Negative", "Neutral" and "Compound" by ascertaining

sentimental scores of particular tweet. The below figure shows the code for adding 4 columns of Positive, Negative, Neutral and Compound to the dataset.

Fig 2 Code snippet for calculating Compound score

As indicated by the business principles, in the event that the compound score of emotion of a sentence is more than 0.05, it is treated as Good, and if the score is lesser than -0.05, it is classified as Bad, if none of the above cases are satisfied then, it's unbiased. With the help of above information, another section is added in our labelled dataset which consists of emotion scores. The below figure shows classification of emotion of tweet by comparing compound score.

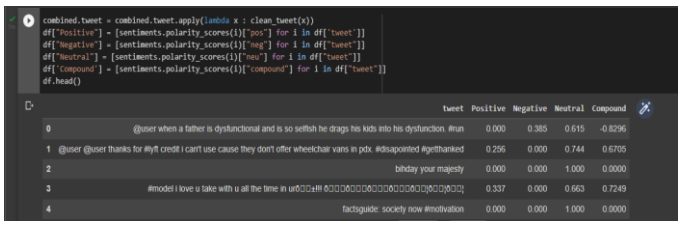


Fig 3 Code snippet to classify emotion of tweet

From the above values we will draw a pie chart to have a better visualization of data. The below figure shows the count of “Positive tweets”, “Negative tweets” and “Neutral tweets” present in the given dataset.

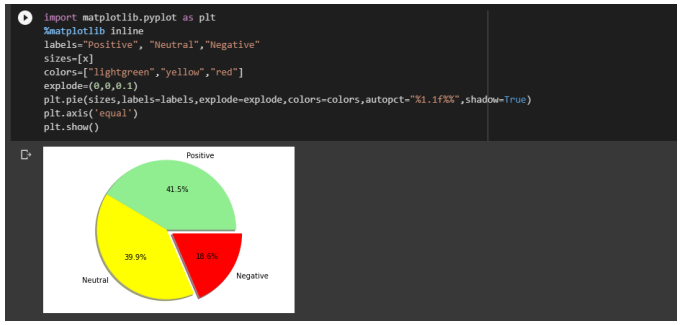


Fig 4 Code snippet for pie chart visualization of count of tweets based on emotion

IV. RESULTS – MODELS ACCURACY

After cleaning and clear analysis of data, we will implement machine learning algorithms such as Logistic Regression, Support vector machine, Random Forest etc. After implementing algorithms, we will try to find out F1 score of each model to check which model is giving best accuracy

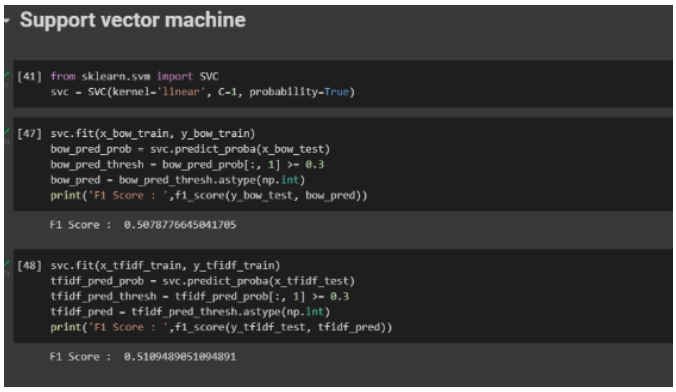


Fig 4 Code snippet for finding SVM algorithm accuracy.

Next, we will implement Random Forest model and we will find F1 score of this model. The below figure explains the code for Random Forest model and finding the F1 score of this model.

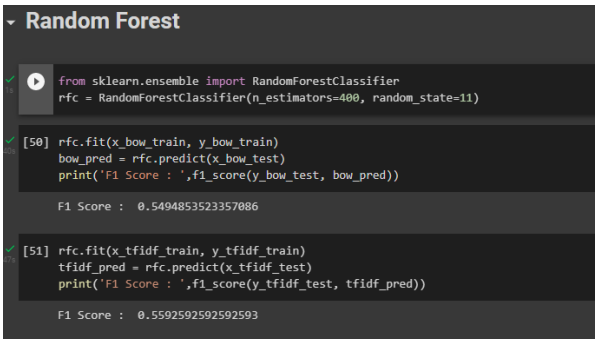


Fig 5 Code snippet for finding Random Forest model accuracy

From the above code we observe that F1 score of Random Forest algorithm is 0.55 which is near to SVM model. So we can conclude saying that this two works with the same accuracy and we can use any of these two models as F1 score is same.

V. FUTURE WORK

From future perspective, we would like to extend this project by implementing some machine learning algorithms for applications like election results, product ratings, movies' outcomes and running the project on clusters to expand its functionalities. Moreover, we wouldlike to make a web application for users to input keywords and get analyzed results. Computation of overall tweet score can be done for a singlekeyword which can provide an overall sentiment of publicregarding a topic.

VI. CONCLUSION

One of the wellspring platform like twitter consists of huge amount information in which some of the information is treated as important and some of the information may contain noisy and unwanted information. For our project this unwanted information may lead machine learning algorithms to work with less efficiency.

So to eliminate this extra information which does not add or create any importance to the sentences, we use few data filtering techniques. These filtering methods are also used in providing data which we want and remove unnecessary information. Generally, tweets represent feedback or person's individual opinion on particular thing or an event. These tweets may be offensive or non-offensive, good or bad. Irrespective of it we took the whole tweet dataset as input and try to detect emotion of a particular sentence. This can be done by using Natural language toolkit. After cleaning the data, we develop machine learning algorithms such as Support Vector Machine and Random Forest. The best model fit is calculated by finding F1 score of these two models and based on the F1 scores we took maximum value of F1 score algorithm and we treat that algorithm as our main priority model. By doing this project we have learnt how to find and calculate a particular tweet emotional analysis score for a large dataset within few minutes and divide tweets respective to their condition of feeling of a person in an organized manner.

REFERENCES

[1] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011 , pp. 30-38

[2] Pang, B.and Lee, L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

[3] Neethu M,S and Rajashree R," Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661

[4] A. Pak,and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Special Issue of International Journal of Computer Application, France:Universitede Paris-Sud, 2010

[5] Dasgupta, S. S., Natarajan, S., Kaipa, K. K., Bhattacherjee, S. K., & Viswanathan, A. (2015, October). Sentiment analysis of Facebook data using Hadoop based open source technologies. In Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on (pp. 1-3). IEEE.

[6] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using Twitter sentiment analysis," in Inventive Computation Technologies (ICICT), International Conference on, 2016, vol. 1, pp. 1–5.

[7] V. Sahayak, V. Shete, and A. Pathan, "Sentiment analysis on twitter data," International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 2, no. 1, pp. 178–183, 2015.

[8] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15. Germany: Springer,2010, pp. 1-15.

[9] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010.

[10] V. M. K. Peddinti and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter," in Analyzing Microtext Workshop, AAAI, 2011.

[11] Ortigosa, A., Martín, J. M., & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior, 31, 527-541.

[12] Trinh, S., Nguyen, L., Vo, M., & Do, P. (2016). Lexicon-based sentiment analysis of Facebook comments in Vietnamese language. In Recent developments in intelligent information and database systems (pp. 263-276). Springer International Publishing.