Sentiment Analysis in Twitter using Machine Learning

S.Jegadeesan[1],S Kamalesh[2],Renuka.K[3],Nagavarshini.S[4],Shalini.P[5]

[1,2]Faculty,Department of Information Technology,Velammal College of Engineering and Technology,Madurai.

[3,4,5]Student, Department of Information Technology,Velammal College of Engineering and Technology,Madurai.

Abstract— Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the people. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text. In this paper, we try to analyze the twitter posts about electronic products like mobiles, laptops etc using Machine Learning approach. By doing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. We present a new feature vector for classifying the tweets as positive, negative and extract peoples' opinion about products. As an advanced step, the proposed research work attempts to find the sentiment of tweets using Logistic Regression sentiment analysis, VADER sentiment analysis and BERT sentiment analysis.

***Keywords*—**Sentiment Analysis; Machine Learning (ML); Classification;BERT;VADER.

I. Introduction

Sentiment is a term used to describe a topic that is subjective and objective and a factual or non-factual topic that transcends the difference between a positive or negative topic [1]. Sentiment analysis is an analysis carried out based on rumors or gossip circulating [2]. Sentiment analysis is an analytical approach used to analyze a text. The purpose of this sentiment analysis is to determine a subjectivity of opinion, the result of a review or a tweets. Based on sentiment analysis, opinions from someone can be classified into various categories based on data size and document type [3].

Nowadays, the community often provides responses and criticisms of leaders, both political figures and public figures through social media such as twitter. Twitter is one of the social media that has a retweet feature that can be used by every user to re-upload information or tweets which allows the dissemination of information on social twitter media to be faster [2]. Twitter is also a social media that can be used to sentiment analysis using data tweets obtained by doing crawler data. Data crawler is a method used to collect data. In this study, we aim to analyze the level of sentiment from the community towards the 2019 presidential candidates of the Republic of Indonesia obtained from the public on Twitter social media, by doing crawler data.

Furthermore, the author will make a comparison of the accuracy of the Naïve Bayes method, with other classification methods such as SVM and KNN. Naïve Bayes method [3] is a method used to group data according to the categories that already exist.

The paper is organized as follows, part II will be explained about related research work and information related to sentimental analysis sentiment on twitter. Part III describes the methods that present the formulas used for classification on sentimental analysis sentiments on twitter. Part IV describes the performance evaluation that contains the results of research that has been done and section V concludes the results of the research that has been done.

II. Related Research Work

The word Sentiment has three layers of meaning:

- Opinion layer.
- Emotion layer.
- Idea colored by emotion layer.

So from the meaning of Sentiments analysis we can say that it is the study of emotions of the user or people. If we go by the definition of Liu then, we say sentiment analysis is the field of research that analyzes the thoughts, feelings, perceptions, behaviors, and emotions of individuals towards things like such as goods, products, political parites, people, problems, incidents, issues, and their attributes.

Farzindar distinguishes the study of feelings and the study of emotions to accentuate the slight difference. Examination of the emotions is more precisely categorized into minor details. Emotion is divided into six classes: rage, frustration, anxiety,happiness, and sorrow, excitement, most widely used in the literature [2]. There is presently no agreement about how many emotional groups should be included. Examination of the emotion is also called identification of moods. The distinction between the study of sentiments and the study of emotions goes over the range of this study. We can categorise Sense into various components like: holder, taget,dimension, and polarity. The growing part suits different tasks within a system. Holder denotes the person bearing theemotion.

A target defines the chosen person as the source of the emotion. A target determines the individual chosen as the origin of the sentiment. Polarity can be described in both positive and negative aspects, or can be represented in three positive, negative and neutral aspects. Feature defines the particular aspect or attribute of the objective to which the emotion is expressed. Take the following example1: Samsung Phone cost lower than ten thousand, does not gives a good performance. Aspect is also an essential part of sentiment analysis.

- Different levels of analysis are given below:

Analysis of sentiment can be classified according finer details of the text. Past research focuses primarily on the various levels given below levels:

- Document-level sentiments analysis:

In this level analysis is done to assess whether the sentiment conveyed in an entire text is positive, negative or neutral. Take an instance, provide product feedback, the system will be able to assess the overall polarity of sentiment. Sentiments of the level of the document imply that a fragment of text communicates feelings about a particular target. Although this is usually appropriate for products analysis, film reviews, hotel reviews, etc., it does not extend to circumstances where several targets are evaluated in a document [6].

- Sentence level sentiments analysis:

In this level we analyzes if the views expressed in a sentence are positive, negative or neutral. Sentence level sentiments analysis can be done in two ways. The first way to classify the sentiments in three different tags that are positive, negative or neutral. And the second method is to find the subjective of the sentence to differentiate between the text which has already been classified and which haven'tbeen classified and then mark them with tags such as positive, negative or neutral. The problems in determining the sentence level is that each sentence is related to semantically and syntactically to some other part of the text. This role, therefore, requires contextual knowledge, local aswell as global.

- Aspect level sentiments analysis:

Evaluation of sentiments at the aspect-level is when the aspects are extracted from the text using various mechanismsand then the sentiment is analyzed for each subject. This canalso be defined as a study of emotions at the feature level. You may decide the feelings of more than one individuals present in one sentence. These can be of three types which are given in brief below:

- Extracting aspects of target,
- Determining aspects-wise polarity,
- Summarizing the overall analysis.

III. Proposed Work

This paper attempts to describe the proposed work in detail.As discussed earlier, tweets related to COVID hashtags will be extracted and then sentiment analysis algorithms like Logistic Regression, BERT and VADERwill be applied.

A. Sentiment Analysis Using Logistic Regression

There are different types of classification models. They are discriminative and generative model. Logistic regression is a model of discriminative classification.So in this model, first all the stopwords should be removed from the dataset. Then it is tokenized and vectorized for further process.

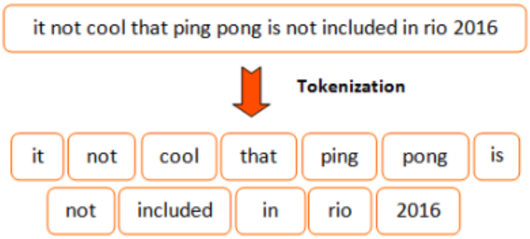


Fig 1: Tokenization in Sentences

For the vectorization purpose, IDs are assigned to each token. Then, the vectors are passed to the CountVectorization method to find the total number of times the word occurs in the dataset i.e. TF-IDF. Then it will be passed to the logistic regression model for sentiment analysis.

B. BERT

BERT stands for Bidirectional Encoder Representation from Transformers. This algorithm is trained to consider the aspect or concept from both the direction that is from left to right or from right to left simultaneously. As a result, we will be able to get the result more accurately. BERT extract more features compared to other **sentiment** analysis algorithms.

They also make use of MLM methods in some cases. The purpose of MLM is to mask a random word in a sentence with a small probability. When the model masks a word it replaces the word with a token [MASK] [13]. BERT is generally using Recurrent Neural Networknamed Long Short -Term Memory (LSTM).to train right to left and left to right simultaneously and concatenate them later. BERT uses Transformers instead of LSTM to get the context of the words because the transformer is more attention based algorithm.

At first, we have to tokenize our data and then vectorize it to get the corresponding ids of the word used. While tokenizing we have to add [CLS] token to the beginning of the sentence and [SEP] token to the end of each sentence.

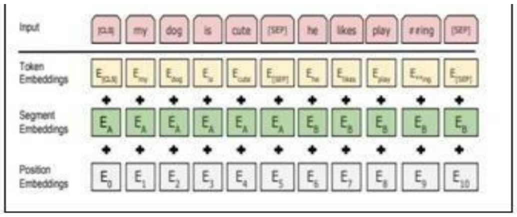


Fig 2: Tokenization And Vectorization in BERT

Then we have to give a small probability to each token by adding a small MASK value to it and finally perform an activation function on it to see if the score or the polarity value is exceeding the threshold set. If so then it is positive else it may be negative or neutral.

A. VADER

VADER is Valence Aware Dictionary and sentiment Reasoner. This VADER makes use of lexicons for getting valence i.e. strength of each word and polarity of each word. This VADER not only gives us whether a statement is positive, negative or neutral but also how much we will get it is positive, negative or neutral.So we will take the polarity of each word separately and apply average in it by normalizing it.

Normalized score = (score) / (sqrt ((score ∗score) + alpha))

The data is first taken for preprocessing and in this stage, we are removing all the stop words, punctuations and performing tokenization in it. This tokenization is done by taking each sentence and then splitting it into words and assigning the token id to all later on these ids will be used by the sentiment analysis algorithm to understand and increase the efficiency of the algorithm.

These tokens are then taken forboosting by using bi-gram, tri-gram methods. These boosting is done in order to reduce the bias and variance to minimal length so that there will be less error while predicting. And finally, the valence score is calculated for each token separately.As we get the set of tweets with the sentiment analysis applied to it we could go for the feature selection process in our work we will be making use of the key concept or the opinions of the users from there tweets to extract the features. After feature extraction, we will be making use of this data to classify the tweets into positive, negative or neutral and thus the sentiment polarity is derived. This takes for one single tweet. We will be collectively doing the same for all tweets in the dataset and then we will be finding out the accuracy and all for the above three algorithms.

Iv. Experiment Analysis

We have taken dataset related to tweets with different hashtags related to COVID -19. This may include #COVID- 19, #WUHAN, #Corona Virus. We have taken a dataset with more than 25000 data in it.Initially we have started our work by finding out the sentiment of a large group of people towards this pandemic situation.We applied different algorithm in our dataset to see the accuracy of each algorithm with respect to same dataset. We have divided our data set into 75 and 25 per cent for training and testing purposes respectively. Which is 18,750 for training and 6,250 for testing. At first we applied sentiment analysis with logistic regression to get the polarity score of the dataset.We are taking 75 and 25 percent because it is good to give more dataset for training as it will give more precise result while doing testing with the testing dataset.

Dataset	positive tweet	negative tweet	neutral tweet
Covid-19	15340	9620	40

Table 1: To Get How Much Feature Is There In Each Polarity

The above table has the number of tweets that is actually positive, negative and neutral .This is done manually as we have to get a reference to check is the predicted value is similar to that of the actual value We can see that from 25,000 data 15,340 were having a positive, 9,620 having negative and 40 having neutral feelings or sentiment towards COVID time.As we get the sentiment classification of each tweets from three different algorithms we could make a collective analysis to get the accuracy,precision,recall and f1-score on three different algorithms.And the datas are as follows:

Algorithm Used	Logistic Regression	BERT	VADER
Accuracy	0.77	0.80	0.85
Precision	0.74	0.85	0.87
Recall	0.70	0.83	0.85
F1 SCORE	0.75	0.81	0.84

Table 2: Accuracy, Precision, Recall And F1 Score For 3 Different Algorithms

In order to get a clear idea on the matric score of the three different algorithm we will go for the following bar graph representation.

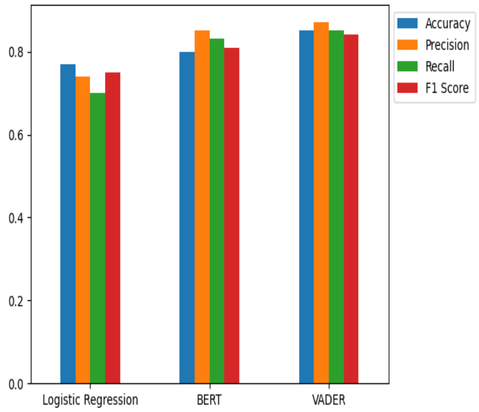
From the figure 5 we could clearly see that applying Logistic regression on COVID tweets is having very low accuracy as compared to the other two algorithms so now let’s concentrate on the

In the case of precision VADER is having more value. As VADER is not taking care of all the aspects of a sentence we cannot consider it as the best algorithm

remaining two algorithm mainly VADER and BERT.Here, we can see that the accuracy of BERT is good accuracy (0.92) compared to the other two algorithms as they are doing the checking of context forward and backward. BERT algorithm on a large dataset consumes more time to give the output so it is better to use along with GPU.

for sentiment analysis. VADER is considered to be a rule- based algorithm and BERT to be an Aspect-based rule.

Fig 4:Accuracy,Precision,Recall and f1-score of three different algorithm.



V. Conclusion

In this paper we have applied three different algorithms say Logistic Regression, BERT, VADER sentiment analysis algorithm. We have normalized the score of the three algorithm with in a range of -1 and 1. This is done so that the comparison will be fair and easy to identify. On comparing we can see that BERT is more accurate (92%) than VADER and Logistic Regression. BERT is more accurate the another algorithms because they look for the aspect of the sentences. VADER will look for the valence and polarity score which will reduce the performance of the feature selection. Logistic regression is not looking for the polarity strength or aspect of the tweets.Moreover we have applied this algorithms on large data and hence it take lot of time to process and get the output.The limitation of our work is it is domain based and we have not looked towards the mood of the user. So we have kept the concept of mood based sentiment analysis for future work.

References

[1] Lokmanyathilak Govindan Sankar Selvan, Teng-Sheng Moh “A Framework for Fast-Feedback Opinion Mining on Twitter Data Streams”, Collaboration Technologies and Systems (CTS), 2015 International Conference, 1-5 June 2015.

[2] Kashika Manocha, Harshita Gupta, Pankaj Kumar “Enterprise Analysis through Opinion Mining” International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

[3] Sunny Kumar, Shaveta Rani, Paramjeet Singh “Sentimental Analysis of Social Media Using R Language and Hadoop: Rhadoop”, Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 5th International Conference.

[4] Anukarsh G Prasad; Sanjana S, Skanda M Bhat, B S Harish, “Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data”, Knowledge Engineering and Applications (ICKEA), 2nd International Conference, Oct. 2017.

[5] Umeshwar Dayal Daniel A. Keim, Lars-Erik Haug, Ming Hao, Christian Rohrdantz, Halldór Janetzko, Mei-Chun Hsu “Visual Sentiment Analysis on Twitter Data Streams”Hewlett-Packard Labs and University of Konstanz. Published in Visual Analytics Science and Technology (VAST), IEEE Conference on 23-28 Oct. 2011 USA.

[6] Sonia Anastasia, Indra Budi, “Twitter Sentiment Analysis of Online Transportation Service Providers”, Advanced Computer Science and Information Systems (ICACSIS) International Conference, Oct 2016.

[7] Alessandra De Paola, Federico Concone, Giuseppe Lo Re, and Marco Morana, “Twitter Analysis for Real-Time Malware Discovery”, AEIT International Annual Conference on 20-22 Sept. 2017.

[8] Yeqing Yan, Hui Yang, Hui-ming Wang, “Two Simple and Effective Ensemble Classifiers for Twitter Sentiment Analysis”, Computing Conference, 18-20 July 2017.

[9] Hase Sudeep Kisan, Hase Anand Kisan, Aher Priyanka Suresh, “Collective Intelligence & Sentimental Analysis of Twitter Data by Using StandfordNLP Libraries with Software as a Service (SaaS)”, Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, 15-17 December. 2016.

[10] M.Trupthi, Suresh Pabboju, G.Narasimha, “Sentiment Analysis on Twitter using Streaming API”, Advance Computing Conference (IACC),

and 2017 IEEE 7th International Conference, Jan 2017.

[11] Saki Kitaoka, Takashi Hasuike, "Where is Safe: Analyzing the Relationship between the Area and Emotion Using Twitter Data", Published in Computational Intelligence (SSCI), 2017 IEEE Symposium Series, 27 Nov to 1 Dec 2017.

[12] Kaustav Roy, Disha Kohli, Rakeshkumar Kathirvel Senthil Kumar, Rupaksh Sahgal, Wen-Bin Yu, "Sentiment Analysis Of Twitter Data For Demonetization In India – A Text Mining Approach", Issues in Information Systems, Volume 18.

[13] Vishal A. Kharde, S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications 139(11): 5-15, April 2016