International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

ML TECHNIQUE-BASED SENTIMENT ANALYSIS FOR MEASURING IMPACT OF SOCIAL MEDIA DATA USING PYTHON TOOL

Preeti Mehrotra[1],Devashri Deoskar[2]
1Department of Computer Science &Engineering,TIEIT Collage Bhopal, India
2Professor,Department of Computer Science &Engineering, TIEIT Collage Bhopal, India

Abstract:

Sentiment analysis focuses on recognising and categorising the thoughts and feelings represented in a piece of writing. Sharing thoughts and feelings on social networking platforms has become a regular practise these days. As a result, a significant volume of data is created each day, from which valuable information may be gleaned via data mining. These data may be used for the sentiment analysis to get a consolidated view on certain items. Because of the prevalence of slang as well as misspellings, doing sentiment analysis on Twitter may be challenging. In addition, new words are continually being encountered, making it more difficult to interpret and calculate the emotion. A tweet on Twitter can only be 140 characters long. As a result, another hurdle to overcome is gleaning useful information from condensed texts. The analysis of feelings in tweets may greatly benefit from a "knowledge-based approach" as well as machine learning. In this study, we'll look at what individuals are saying about Covid-19 in their tweets. To better serve the views, a simple sentiment score might be computed and then classified as either good or negative by individuals all around the globe.

1. INTRODUCTION

Twitter has turned into a famous micro blogging administration that has a vast number of comments and client opinions to make status messages called tweets. Online users utilize these tweets for buying products to refreshing the tweet reviews to fix our idea. This yet additionally to express their tweets comments have sentiment towards items, products, instances and other process they are considered. Various tweet analyzing Framework be implemented in numerous situations defines the feedback about the Twitter information are valuable in certifiable circumstances.

Twitter, is a social blog of the tweet words discusses the relational association, gives collected feedbacks among customers' and hosts feeling rich data over a wide course of action of customers and subjects. Along these tweets sentimental analysis, knowledge mining from feedback and sentiments from Twitter will be particularly useful for selecting new products.

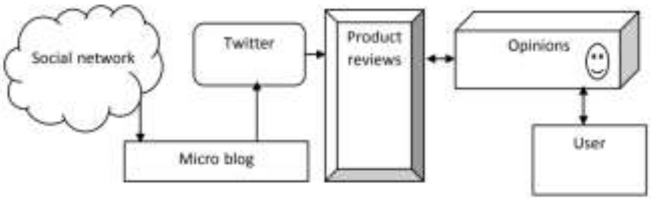


Figure 4.1 Microblogging Social Network

The primary purpose of this investigation has to divide the tweets in Twitter that can motivate the path towards characterizing the more relevant classification(s) that divides the different categories of the classification.

Sentiment analysis is a recent branch of study in "Natural Language Processing" (NLP) that aims to discover and categorise subjective views and feelings in text. There are many different types of sentiment analysis, each of which focuses on a certain aspect of a person's emotional response to a product or service.

Text is categorised based on the following criteria in the sentiment analysis:

- “The polarity of the sentiment expressed (into positive, negative, and neutral);
- The polarity of the outcome (e.g., improvement versus death in medical texts);
- Agree or disagree with a topic (e.g., political debates);
- Good or bad news;
- Support or opposition;
- Pros and cons”

1.1 Sentiment Analysis

There is a kind of hearing in the dialog that speaks of the perception of a certain item or point of view from all individuals. Sense investigation, or, in other words, involves building a gander by building and assembling the item on blog posts, comments, reviews or ratings made by tweets. A couple of emotional experiments can be useful in various ways. For example, in exhibiting, it helps in settling on a choice about the achievement of an advancement fight or new item dispatch, make sense of which types of an item or organization are popular and even perceive which economics like or severely dislike particular features.

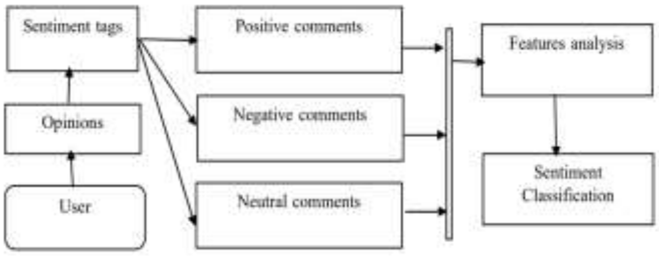


Figure 4.1.1 Process of Sentiment Analysis

In general, there are 3 approaches to sentiment classification: “Lexicon-based, machine learning-based, and hybrid”. ML employs well-known machine learning methods and language characteristics. In the "Lexicon-based Approach", a sentiment lexicon is required. A lexicon is a list of well-known and pre-compiled emotive phrases that may be consulted. For sentiment analysis, dictionary-based approaches may be subdivided into those that rely on semantics or statistical techniques, and those that employ a combination of both. Both techniques are prevalent, while sentiment lexicons play an important part in many ways.

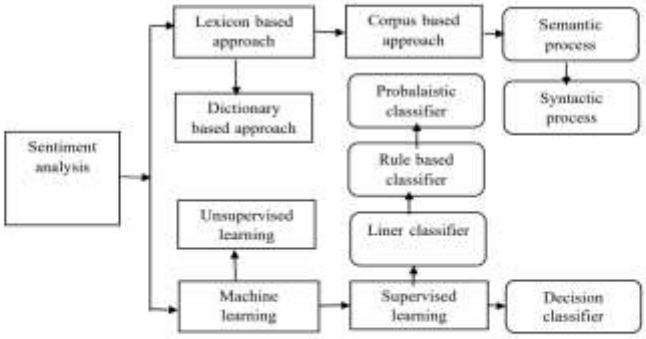


Figure 4.1.2 Sentiment Analysis Methods

1.2 Natural Language Processing (NLP)

Natural language processing (NLP) aims to give computers the capacity to interpret text as well as spoken speech like humans. NLP blends the “computational linguistics” with statistics, machine learning, as well as deep learning models. These techniques let the computers to 'understand' the human language in the text or the speech data, including intent & mood. Having pre-processed the data, the next stage is to create an NLP algorithm then train it to understand natural language as well as do certain tasks.
NLP issues may be solved using one of two algorithms:

1. **A rule-based approach.**

The grammatical rules for rule-based systems are handcrafted by linguistics specialists or knowledge engineers. In the early days of developing NLP algorithms, this was the first strategy used, and it is being used today.

2. **Machine learning algorithms.**

A machine-learning model is based on statistical approaches and learns to accomplish a task after being given examples (training data).

Among the greatest advantages of the "machine learning algorithms" is the fact that they are self-taught. Automated learning eliminates the need for pre-defined rules; instead, robots draw on their own experience to make educated guesses about the future

2. LITERATURE REVIEW

E. Boiy, P. Hens, proposed an automatic sentiment investigation in online content; the programmed examination of sentiments on data found on the Web is valuable for any organization or foundation thinking about quality control. This strategy can be made out of date by social event such data consequently from the World Wide Web, People share their encounters on-line, ventilate their assessments or essentially speak pretty much anything. One of the sources are blogs, a medium through which the blog proprietor makes discourses about a specific subject or discusses his or her own encounters, welcoming per users to give their very own remarks. Another source are the electronic discourse sheets, where individuals can examine a wide range of subjects, or request other individuals' sentiments.

B. Pang, L. Lee, and S. Vaithyanathan, talked about a Sentiment grouping utilizing machine taking in strategies. “Machine learning is a branch of artificial intelligence that focuses on the study and development of algorithms that are able to learn from and predict data, rather than following to a set of predetermined rules”. Manmade awareness may be used to allow frameworks to absorb and improve without being explicitly altered. For the machine learning, it's all about developing PC programmes that can access and use data to learn on their own. Man-made brainpower's central sub-zone, machine learning, allows PCs to enter a mode of the self-learning without being explicitly altered. Also examined a Sentiment grouping utilizing machine learning procedures; today, a lot of data are accessible in online archives. As a component of the push to more readily compose this data for clients, analysts have been effectively examining the issue of the programmed content classification. Connected a particular unsupervised learning procedure dependent on the shared data between archive phrases and the words "great" and "poor", where the common data is registered utilizing insights accumulated by a web search tool.

C. Hsu, C. Chang, and C. Lin, talked about a typical technique to isolate the data set into two sections, of which one is viewed as obscure. The expectation exactness got from the "obscure" set all the more decisively mirrors the execution on ordering an autonomous data set. An enhanced form of this system is known as cross-approval. A reasonable manual for support vector grouping used to Support Vector Machines are a helpful strategy for data order. Despite the fact that clients don't have to comprehend the fundamental hypothesis behind

Support Vector Machines, they quickly present the nuts and bolts vital for clarifying our method. A grouping assignment for the most part includes isolating data into preparing and testing sets.

X. Ni, X. Quan, Z. Lu, L. Wenyin, and B. Hua proposed a new "short text clustering strategy", Termcut. G = (V,E) where each vertex represents a small text fragment and every weighted-edge between the two vertices quantifies the connection between the vertices A novel criteria, RMcut, is used to find the core words. Quality of clusters is measured according to the clustering principles that minimise intercluster similarity while enhancing intra-cluster similarity, which is the purpose of the Rmcut criteria.

D. Pinto, P. Rosso, and H. Jimenez,introduced new metrics in order to identify whether or not a corpus is made up of short texts with a confined domain. Using an autonomously generated "lexical knowledge resource" drawn from the same target data set, they developed a technique for enriching baseline corpora by adding co-related words (not from an external resource). They used this approach to group scholarly abstracts from a certain field.

L. Ji, H. Shi, M. Li, M. Cai, and P. Feng, we can extract valuable insights from the product reviews using a sentiment mining as well as retrieval system. Visual representations of the system's sentiment orientation as well as comparison of positive & negative evaluations were also included. Research has revealed that the method is possible and successful when tested on a real dataset.

3. Research Methodology

The proposed study simply focuses twitter text mining to come up with overall sentiments of common people are positive or negative. The researcher aims to analyses pre-processing techniques, tokenization and term frequency calculations to form result classes. And also to represent these classes using visual techniques to identify and judge overall level of positive or negative impact of demonetization on common people through their tweet discussions.

The main aim of present research work is to apply ML technique-based sentiment analysis for measuring impact of social media data using Python tool that would help the analyst, businessmen, politicians, governments as well as students in the process of "sentiment analysis".

For the said research work the researcher has set following objectives:

To extract the tweet data set from kaggel on "Covid-19".

i. Preprocess Tweets (need to remove noise and preprocess tweets like convert the tweets to lower case)
ii. To study "machine learning based sentiment analysis" for opinion categorization, pattern detection and prediction.
iii. Analyzing Tweets for Sentiments.
iv. To apply grouping and sub grouping fuzzy classifications of social media data.
v. Compare our result with existing work.

The tweets are collected in first stage and tokenization process is applied for that data. The training set data is generated after preprocessing step and assigned as input to the LR, SVM. NaïveBayes, Random Forest, K-Nearest Neighbour algorithm for sentiment classification analysis.

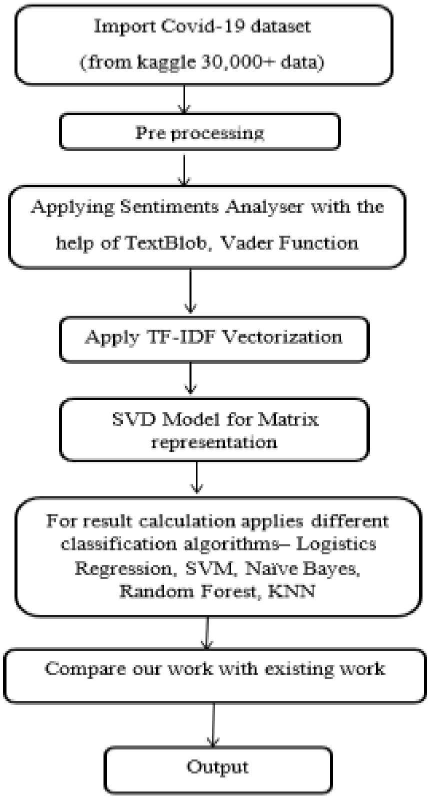
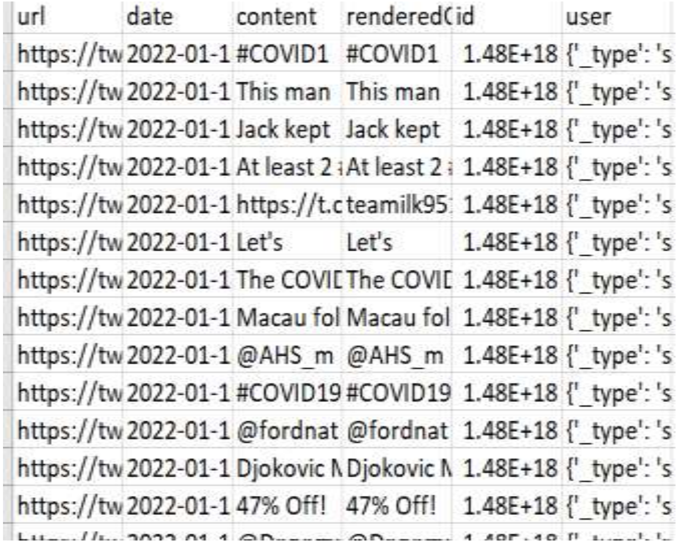

Import Covid-19 dataset
(from kaggle 30,000+ data)
Pre processing
Applying Sentiments Analyser with the help of TextBlob, Vader Function
Apply TF-IDF Vectorization
SVD Model for Matrix representation
For result calculation applies different classification algorithms– Logistics Regression, SVM, Naive Bayes, Random Forest, KNN
Compare our work with existing work
Output

Figure 4.2 Proposed Methodology

- Apply Pre-Processing:
- Removal of stop words, punchuation marks, whitespaces, numbers and noisy data
- Removal of stopwords, punctuation marks, whitespaces, numbers and noisy data
- Detect sentiment term in refined dataset terms.
- Determine term words, frequency and probability of occurrence i. e. Quantify Polarity.
- Determine positive and negative classes for entire data with positive attitude and negative attitude.
- Analyze the classes to get overall sentiment on tweets of covid-19.
- Create Visualization – plot graphs using gplot2 for generated dataset.

3.1 DATA COLLECTION

Twitter streaming API is used for collecting tweets with including URL's. Public streaming API is used for accessing real-time tweets in which only 1% of overall public tweets are granted to access due to restrictions over protected accounts. .CSV format is the form that collected tweets could be viewed and each line is pared as objects. Therefore, a total of 30,000+ tweets are collected and several attributes are presented in the collected tweets and mainly two types classified and these are tweet-based features and user-based features. For the proposed work, text attribute is used for sentiment classification process. The dataset parameters are described as “Positive”, “Neutral”, and “Negative”.



url	date	content	renderedC	id	user
https://tw	2022-01-1	#COVID1	#COVID1	1.48E+18	{'_type': 's
https://tw	2022-01-1	This man	This man	1.48E+18	{'_type': 's
https://tw	2022-01-1	Jack kept	Jack kept	1.48E+18	{'_type': 's
https://tw	2022-01-1	At least 2 :	At least 2 :	1.48E+18	{'_type': 's
https://tw	2022-01-1	https://t.c	teamilk95:	1.48E+18	{'_type': 's
https://tw	2022-01-1	Let's	Let's	1.48E+18	{'_type': 's
https://tw	2022-01-1	The COVII	The COVII	1.48E+18	{'_type': 's
https://tw	2022-01-1	Macau fol	Macau fol	1.48E+18	{'_type': 's
https://tw	2022-01-1	@AHS_m	@AHS_m	1.48E+18	{'_type': 's
https://tw	2022-01-1	#COVID19	#COVID19	1.48E+18	{'_type': 's
https://tw	2022-01-1	@fordnat	@fordnat	1.48E+18	{'_type': 's
https://tw	2022-01-1	Djokovic N	Djokovic N	1.48E+18	{'_type': 's
https://tw	2022-01-1	47% Off!	47% Off!	1.48E+18	{'_type': 's

Figure 3.1 Datasets sample

3.2 Data Preprocessing

Web based life locales have numerous dialects that utilized which are not quite the same as predominant press found and words in the lexicon. An uncommon ”slang”, emojis are utilized in web-based social networking stages to stress words by rehashing a portion of their letters. Furthermore, particular attributes like markup tweets are utilized for dialects in twitter that were reposted by different clients with ”RT” and furthermore clients signs ”@” and markup of subjects utilizing ”#” is utilized. The preprocessing of tweets contains following stages as appeared in Figure 4.2.

Figure 3.2 demonstrates the preprocessing steps that incorporate expulsion of stop words, contraction development, amending incorrectly spells in the content, stemming of words, recognizable proof of labels, positive and negative word arrangements of each tweet.

Tokenization: In this process, the texts are converted into meaningful words, phrases or symbols that are known to be tokens. The tokens are used for text parsing or mining. During the process of tokenization, if any error occurred then some problem can be caused during classification process. The first step includes in segmenting text in into Word boundaries are located which helps in the formation of tokenization. Starting of the word and end of the statement is known as word boundary. Tokenization is a prerequisite for any kind of language processing. If the words are separated by spaces, tokenization will be a breeze. When tokenization happens, the punctuation sign is believed to represent white space if there is no white space present in the token. The tokens' concepts must be established before any processing can begin. The concept may be used in a variety of ways, including linguistic or methodological. As an example, consider the following definition of tokenization:

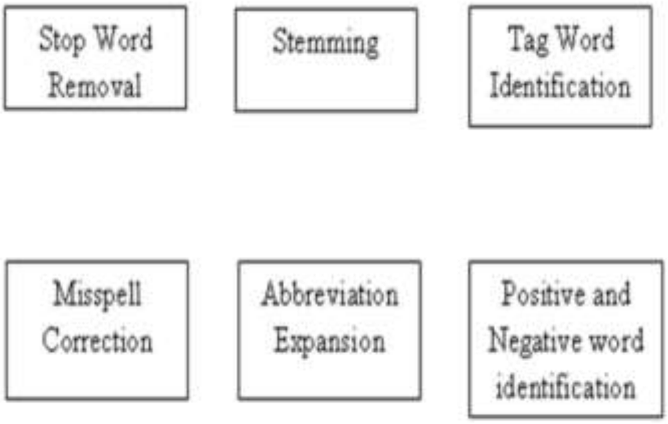
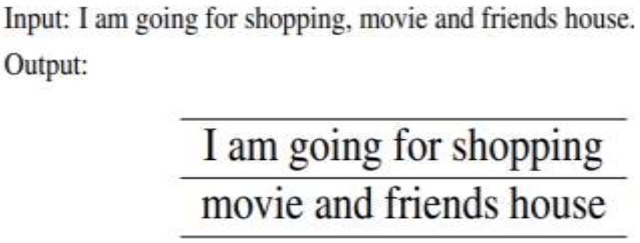


Figure 4.3 Preprocessing Steps



Input: I am going for shopping, movie and friends house.

Output:

I am going for shopping
movie and friends house

Feature Generation: The preprocessing steps are done for further feature generating process. The preprocessed tweets are forwarded for generating features. The frequency of negative and positive words, analyzing scores for the positive and negative, tag count in the text are analyzed in this section and overall scoring for the texts are presented. “Here, various features are used in the learning classifier and these are word count that defined as total words present in each tweet after preprocessing is done, tag count referred as total number of @ tags used in each tweet, negative word count is the total number of negative words present in each tweet, positive word count is the total positive words in each tweet, positive score is the total number of positive scores gained after adding the positive adjective, negative score is defined by total number of negative scores obtained when adding each negative adjective, and score is the final total outcome by subtracting negative score for each tweet with positive score”.

4. RESULTS AND DISCUSSION

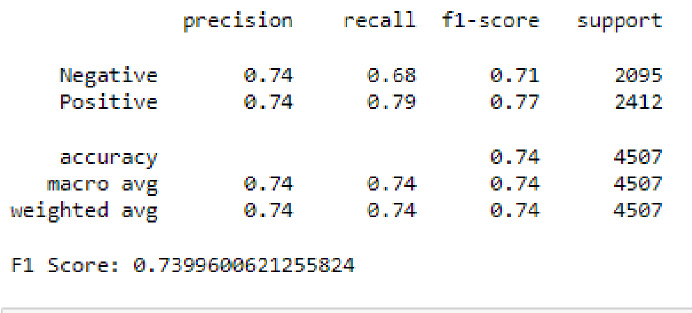
The results of this study demonstrate that the machine learning algorithms outperforms other models, in terms of accuracy. The use of these algorithms captures the context of the text in both directions, which can be crucial for understanding the sentiment in the tweets. Twitter Sentiment Analysis (TSA) has been used for several applications which include product reviews, political orientation extraction, stock market prediction etc. More over the real time tweets analysis on various issues has become a strong indicator to analyze the human behavior and reaction on various issues.
All packages mention in below paragraph which are used in this project.

Software & Tools

Language: Python
OS: Windows 10

Packages

- Numpy
- Pandas
- Matplotlib
- Nltk
- sklearn



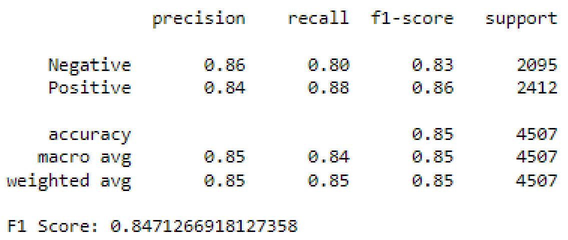
```
              precision    recall  f1-score   support

    Negative       0.74      0.68      0.71      2095
    Positive       0.74      0.79      0.77      2412

    accuracy                           0.74      4507
   macro avg       0.74      0.74      0.74      4507
weighted avg       0.74      0.74      0.74      4507

F1 Score: 0.7399600621255824
```

Figure 4.1 KNN Result



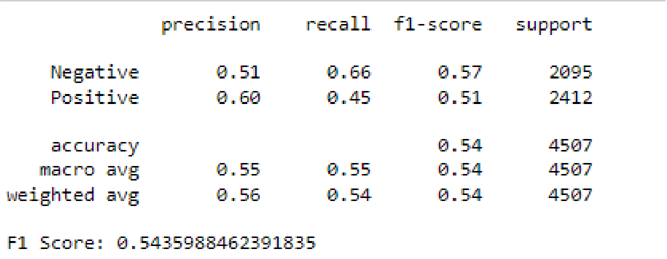
```
              precision    recall  f1-score   support

    Negative       0.86      0.80      0.83      2095
    Positive       0.84      0.88      0.86      2412

    accuracy                           0.85      4507
   macro avg       0.85      0.84      0.85      4507
weighted avg       0.85      0.85      0.85      4507

F1 Score: 0.8471266918127358
```

Figure 4.2 Logistic Regression result



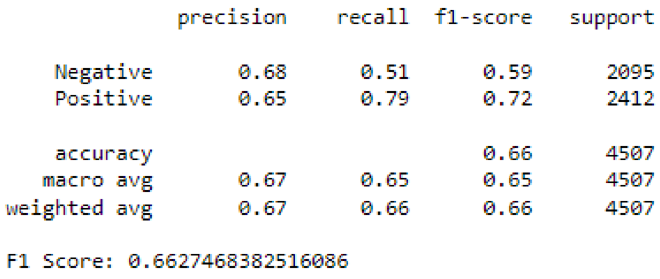
```
              precision    recall  f1-score   support

    Negative       0.51      0.66      0.57      2095
    Positive       0.60      0.45      0.51      2412

    accuracy                           0.54      4507
   macro avg       0.55      0.55      0.54      4507
weighted avg       0.56      0.54      0.54      4507

F1 Score: 0.5435988462391835
```

Figure 4.3 Naïve Bayes Result



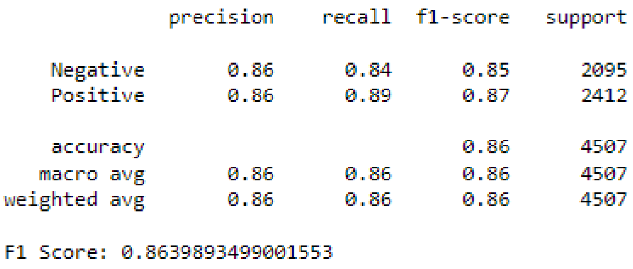
```
              precision    recall  f1-score   support

    Negative       0.68      0.51      0.59      2095
    Positive       0.65      0.79      0.72      2412

    accuracy                           0.66      4507
   macro avg       0.67      0.65      0.65      4507
weighted avg       0.67      0.66      0.66      4507

F1 Score: 0.6627468382516086
```

Figure 4.4 Random Forest Result



```
              precision    recall  f1-score   support

    Negative       0.86      0.84      0.85      2095
    Positive       0.86      0.89      0.87      2412

    accuracy                           0.86      4507
   macro avg       0.86      0.86      0.86      4507
weighted avg       0.86      0.86      0.86      4507

F1 Score: 0.8639893499001553
```

Figure 4.5 SVM Results

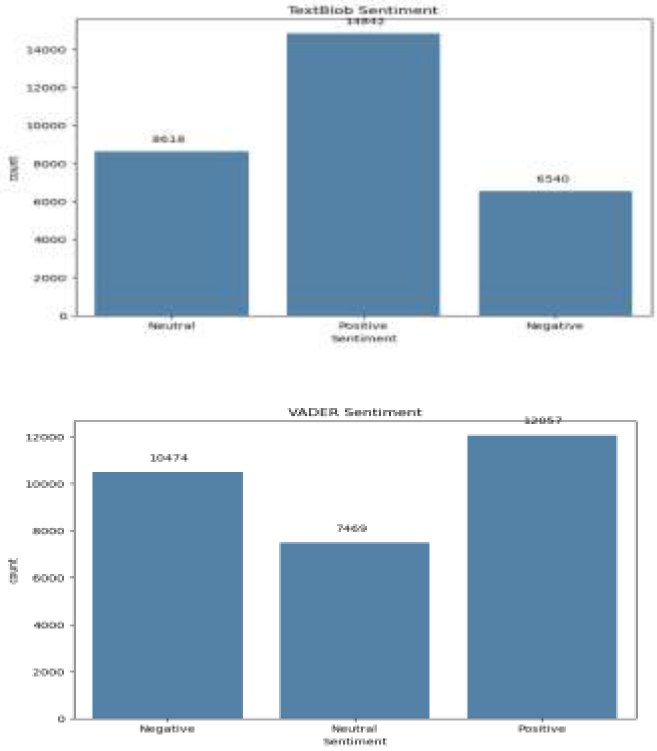


Figure 4.6 (a) TextBlob Sentiment (b)Vader Sentiment

5. CONCLUSION AND FUTURE WORK

The advent of online social networks has tremendously increased the online participation of users in various activities and hence produced large volume of contents. Most of these user generated contents are in the form of “Short Text”. Moreover, the cryptic and informal nature of these short texts produced in social media also poses new challenges. The analysis of this large volume of data requires multidisciplinary approaches which includes social media analytics, opinion mining, social network analysis etc.

In this proposed work, the results obtained can be seen a very huge difference in terms of accuracy when compared with other classifiers when using a large amount of data.

The future scope involves in combining other existing algorithms to obtain the performance for large number of datasets and ratio of spam to non-spam tweets should be considered for improving the performance metrics. Also, different extraction techniques can be combined to obtain new feature set which would be helpful in improving the classification performance.

REFERENCES

[1]. E. Boiy, P. Hens, K. Deschacht, and M. F. Moens, Automatic sentiment analysis in on-line text,” *Openness Digit. Publ. Awareness, Discov. Access - Proc. 11th Int. Conf. Electron. Publ. ELPUB 2007*, no. January 2016, pp. 349–360, 2007.

[2]. B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification Using Machine Learning Techniques,” *EMNLP*, vol. 10, Jun. 2002, doi: 10.3115/1118693.1118704.

[3]. C. Hsu, C. Chang, and C. Lin, “A practical guide to support vector classification,” *BJU Int.*, vol. 101, pp. 1396–1400, Jan. 2008.`

[4]. X. Ni, X. Quan, Z. Lu, L. Wenyin, and B. Hua, “Short text clustering by finding core terms,” *Knowl. Inf. Syst.*, vol. 27, pp. 345–365, Jun. 2011, doi: 10.1007/s10115-010-0299-7.

[5]. D. Pinto, P. Rosso, and H. Jimenez, “A Self-enriching Methodology for Clustering Narrow Domain Short Texts,” *Comput. J.*, vol. 54, pp. 1148–1165, Jul. 2011, doi: 10.1093/comjnl/bxq069

[6]. L. Ji, H. Shi, M. Li, M. Cai, and P. Feng, *Opinion mining of product reviews based on semantic role labeling*. 2010.