ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse310942020.pdf

https://doi.org/10.30534/ijatcse/2020/310942020

Sentiment Analysis of Tweets on Social Issues using Machine Learning Approach

Chhinder Kaur[1], Dr. Anand Sharma[2]

[1]Department of Computer Applications, UCCA, Guru Kashi University, India,
chhinderkaur87@gmail.com

[2]Department of Computer Applications, UCCA, Guru Kashi University, India,
andz24@gmail.com

ABSTRACT

Sentiment analysis plays an important role in analyzing people's opinions, emotions, and feelings with the help of Natural Language Processing and Artificial Intelligence. This paper presents five major social issues of the world like Corruption, Women violence, Poverty, Child abuse, Illiteracy that are the most critical in the world. These evils are barriers to the social development of the peoples. Tweets from the years 2006 to July 2020 are collected to analyze the sentiments of peoples related to these social issues with Machine learning tools and techniques. In this proposed study different Classification algorithms, NLP techniques like tokenization, Stemming, stop words, Word Cloud are applied for effective enhancement of dataset. Sentiments are analyzed as positive, negative, and neutral from the twitter dataset. Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest Classifier, Decision Tree, and Stochastic Gradient Descent classifier applied to build the model and also measure the performance by precision, recall, and F1 score parameters.

Key words: Natural Language Processing, Word Cloud, Twitter Sentiment Analysis, Classification Algorithms.

1.INTRODUCTION

Social Media plays an important role to share people's opinions, feelings, emotions, and experiences in a few words. People are using several social media platforms like Facebook, WhatsApp, and Twitter, etc. to share their posts, twitter is one of these. Twitter is a mostly used social media platform where people can post their tweets within 280 characters.

In this paper, Twitter is used to collect people's posts related to major social issues like corruption, women violence, poverty, child abuse, illiteracy, etc. in the world. Twitter API is used to fetch the year wise tweets from the year 2006 to July 2020 using Python programming language. These tweets are collected from different countries of the world because most of the countries are affected by these evils. By the medium of Twitter people share their problems and opinions related to social issues

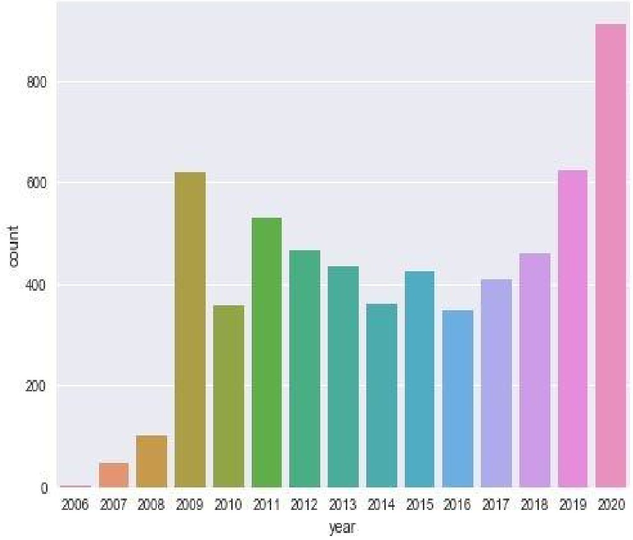

Figure 1: Data Description

In figure 1 shows the collection of tweets as per year from 2006 to the year 2020. It also shows that most of the tweets are collected from the year 2020.

After collecting the tweets from twitter using Twitter API, the dataset is saved as a CSV file in the database and then pre process the data to analyze the sentiments as positive, negative, and neutral using TextBlob in python. TextBlob is the best module in python to analyze the sentiments

Natural language processing (NLP) is the best way to train and improve the word extraction from twitter dataset**.** To enhance the efficiency of the twitter data, remove the unnecessary noise from the data like missing values, punctuation marks, etc., these are created a disturbance to analyze the sentiment efficiently. Preprocessing is done using the NLTK (Natural Language Toolkit).

In this paper, plot the most used hashtags to post their tweets by peoples and also plot the people's names who post most of the tweets are related to social issues

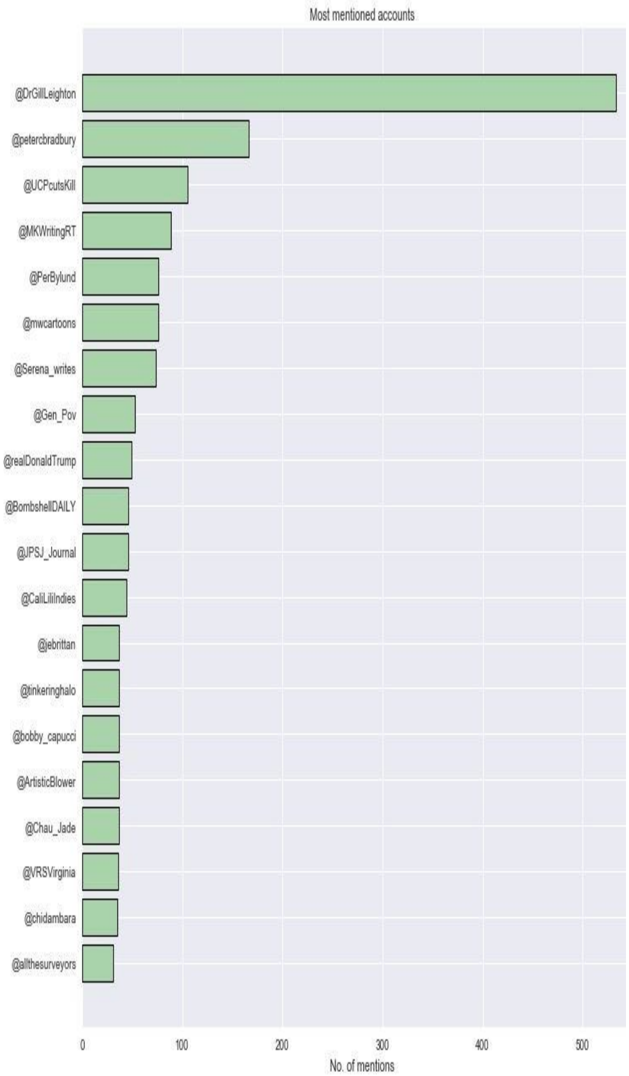

Figure 2: Most used hashtags

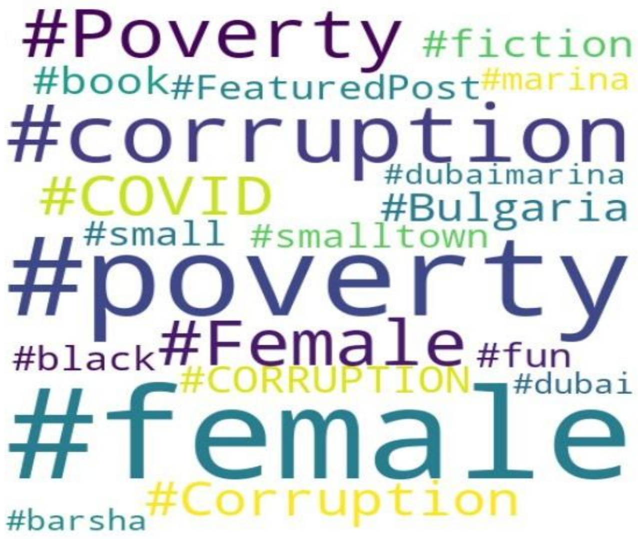

Figure 3: Word cloud of most used hashtags

Figure 2,3 shows that the hashtags used by people to post their tweets on twitter related to different social evils in the different countries in the world.

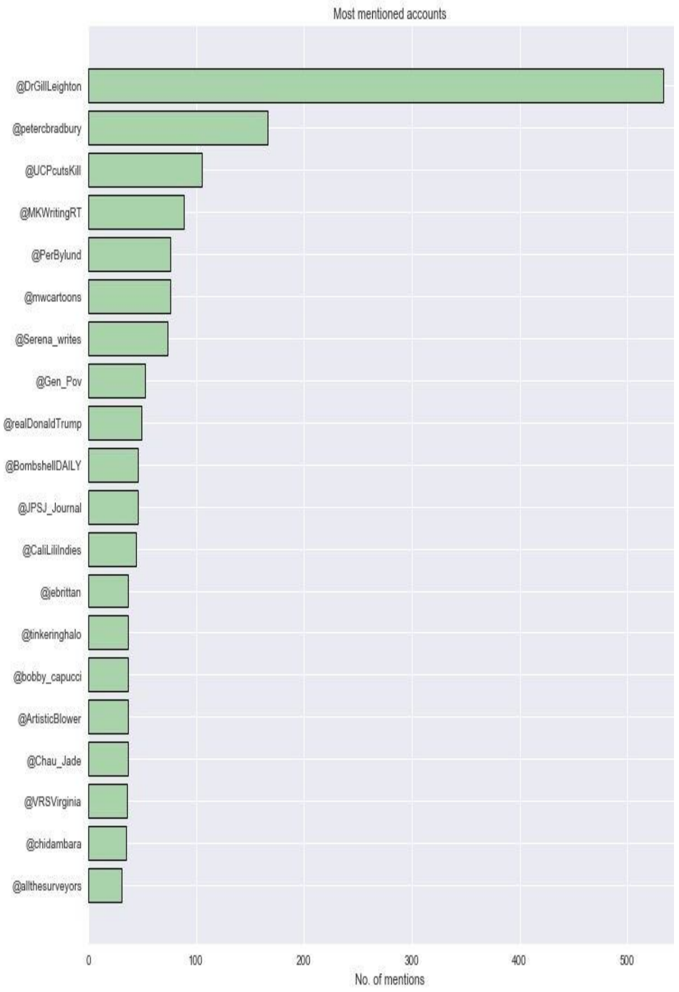

Figure 4: Most mentioned accounts

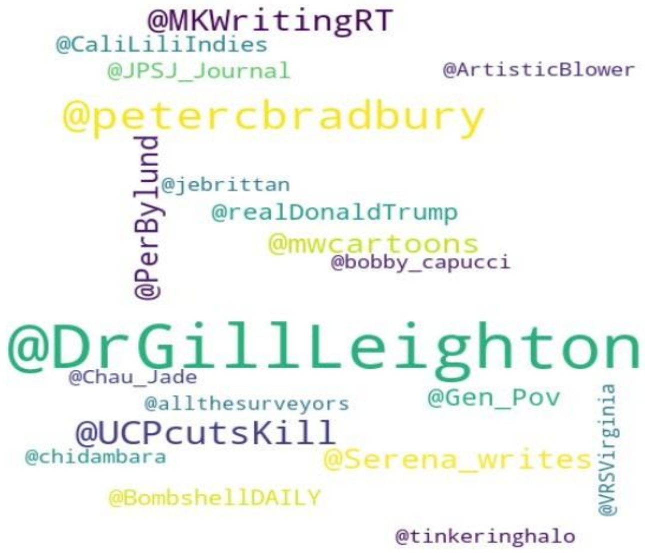

Figure 5: Word Cloud of most mentioned accounts

In Figure 4,5 shows that the most mentioned accounts. Dr. Gill Leighton posts most of the tweets related to social issues.

The different classification algorithms help to classify the people's opinions as positive, negative, and neutral sentiments. By using the different algorithms like naïve Bayes, Support Vector Machine, Random Forest Classifier, Logistic Regression, Decision Tree, and SGDC classifier summarize the result and predict the accuracy of the model. For applying machine learning algorithms, required different packages like pandas, NumPy, sklearn, seaborn, matplotlib, NLTK, word cloud, porter stemmer, etc.

Pandas are used to handling dataset operations like

- extracting Excel, CSV, text, reading a dataset
- selecting particular rows and columns
- converting dataset into a data frame
- removing unwanted data
- Checking with null values

NumPy package used for performing

- numerical
- calculus
- integration
- Array operation

Matplotlib and seaborn packages help to

- diagrammatic representation of data
- pie chart, Bar plot, correlation graph, etc.

Sklearn contains

- all the regression and classification algorithm supporting packages,
- accuracy, confusion matrix packages.
- data feature selection packages

Word cloud package helps us to

- clean the dataset and extract only keywords words from the dataset.

With the help of NLTK, we can process the feedback and extract the important features from customer reviews.

1.1 Literature Review

Bouazizi et al [1] Collected the twitter dataset from social media and analyzed the sentiments as positive, negative, and neutral and the SENTA tools used to build a model for multi-class classification. This experiment shows 60.2% accuracy on the multi-class classification and accuracy of the model after removing neutral tweets is 81.3% and in the latter case, the accuracy of classification equal to 70.1%. Hasan et al [2] this paper, analyzes the sentiment and opinion mining of social media through various machine-learning tools and techniques during elections.

Moreover, this paper also provides a comparison of techniques of sentiment analysis in the analysis of politics. Masdisornchote et al. [3] analyze the sentiments of people's feelings, opinions regarding particular product reviews, issues, topic, and particular situations. In this paper the author classify the text as Subjectivity and also analyze the sentiments of each text as positive and negative. This research focuses on the sentiment analysis of social issues.

Groot [4] in his thesis classifies the sentiments as three classes as positive, negative, and neutral using data mining techniques. Preprocess the data from text using a feature vector and then split the dataset into a train and test. supervised learning algorithms as support vector machines and naive Bayes are used to build a prediction model.

Shahjahan et al. [5] also analyzes the feelings, communications, and ideas using social media platforms. This paper focuses on newspapers, radio, and television where people share information and ideas on Twitter, Facebook, Flickr, and Blogs, etc. Xue, Jia Chen et al [6] in this paper analyze the sentiments of people related to domestic violence using data mining tools and techniques.

Bing [7] focuses on people's opinions as positive or negative sentiments. The main focus of this paper is text classification, text clustering, and other text mining techniques and natural language processing tasks used to enhance the accuracy of the classification.

Smailović [8] focuses on predicting stock market prices using the Support Vector Machine (SVM) classifier. Tweets related to the stock market are classified into three sentiment categories of positive, negative, and neutral (instead of positive and negative only). Ahmad [9] analyses sentiments of tweets related to corruption that are posted from peoples. In this paper analyze the sentiments and effects of corruption in Indian society Various tools and techniques are used to analyze tweets.

Shahid [10] reviews problem caused by social media that affects the children from 2 years to 18 years. Analyze the effects on children's mental health, behavior, and other problems. The main focus of this paper is to aware of the parents to protect their children from the negative effects of social media.

Medhat et al. [11] and Clavel et al. [18] analyzed the emotions, opinions. The main focus of this paper discusses the sentiments using various sentiment analysis techniques and also discusses the different applications of sentiment analysis.

Abraham et al [12] analyses the reviews of customers related to mobile phones using different machine learning techniques. O. Devi [13] used a fine-grained sentiment analysis using machine learning approaches in this paper. D'Andrea et al. discussed various Approaches, Tools, and Applications of Sentiment Analysis[14]. Masdisornchote analyzed the social issues in their research[15]. Jianqiang et al. compared the different pre-processing methods of text for analyzed sentiments[16] and Yu [17] refining the scores of sentiment analysis. Putra et al. [19] analyzed the user reviews regarding mobile applications used the Naïve Bayes algorithm. M. Bouazizi et al analyzed the multiclass twitter sentiments used Pattern-Based Approach[20].

2. METHODOLOGY

Workflow

In this paper, used the Natural language processing and Machine learning classification algorithm to analyze the sentiments as positive, negative, and neutral of people's opinions/emotions. This workflow diagram explains the complete process of the proposed study.

A. Collection of Tweets: Collected required dataset of tweets from different hashtags of social issues using Twitter API in Python.

B. Store Tweets: Stored all collected tweets into different CSV files as per hashtag and then combine into one CSV file.

C. Pre-processing: After storing the tweets into one CSV file, preprocessed of the dataset using NLP, pandas, NumPy, etc. module has been used for preprocessing and remove punctuation marks, missing values from the dataset to eliminate unwanted noise from the dataset.

D. Sentiment Labeling: Labeled all tweets as positive, negative, and neutral sentiments using Text Blob, which is a very effective module for tweets labeling in Python.

E. Classification Algorithms: Different Machine learning algorithms like naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree, and SGDC are applied to train the model and find the accuracy of the specific model.

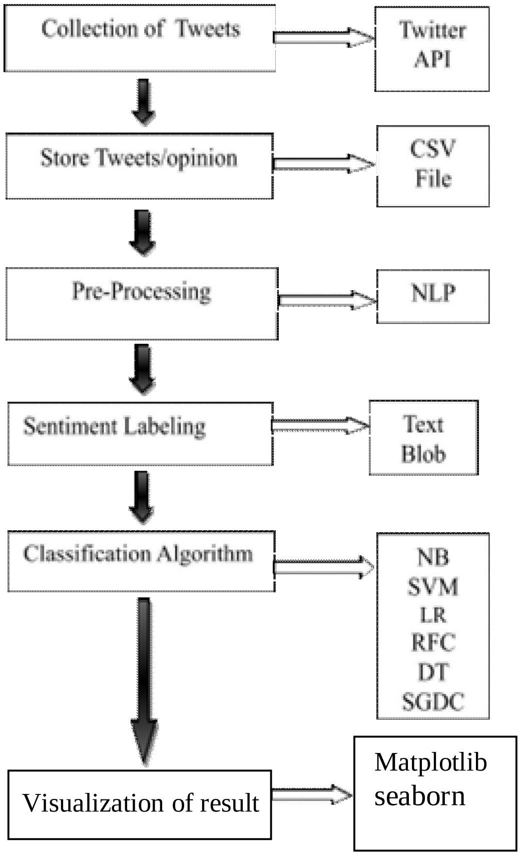

Figure 6: Flow Diagram for NLP and Classification

Twitter Analysis Procedure

1) **Sentiment Analysis:** Sentiment analysis is very helpful to analyze the people's opinions, emotions, feelings, reviews, and likes/dislikes. Sentiment analysis helps in the business for a product review, education sector, politics, social issues, etc.

2) **Natural Language Processing:** It is a technique for train the model so that it needs to break down, perceive, and comprehend the human language. NLP can be utilized for discourse acknowledgment, understanding client surveys, news arrangements, and so forth.

2.2.1. Steps for NLP processing

i. **Tokenization**: Tokenization is the method of breaking the tweets/sentences into little lumps or words. Sentence tokenize breaks each section into a sentence and Word tokenized breaks each section into words.

ii. **Stemming**: Stemming keeps only the root words. Ex: a word like Analyzed, Analyzing is reduced to a common word like Analyze.

iii. **Parts of Speech tagging (POS):** POS is utilized to distinguish things, pronouns, intensifiers, and action words. POS is utilized to distinguish the relationship within the sentences and appoint the labels to compare words.

iv. **Stop words**: Stop words are creating unwanted noise in the dataset, so it is compulsory to remove from the dataset for efficient analysis. is, am, as, they, was, at, it, for, etc. are stop words. NLTK module is used for removing these types of noise from the textual dataset

v. **Punctuation Marks:** Punctuation marks and special symbols like hashtags, mentions, question marks, etc. are created noise into the dataset, so it's compulsory to eliminate them into the textual dataset.

vi. **Word cloud:** Word Cloud is a Data perception method used to show the common words that are used by people to share their tweets on twitter or any other social media sites. The word cloud helps to identify the words.

Count Vectorization: Count vectorization package converts all the words int 0's and 1's and it will be represented as a count vector matrix. It also identifies the frequency of every word. Once all NLP processes are done then the dataset has to split into 80% of the training dataset and 20% of the testing dataset. By using a training dataset, we can train the machine and by using the testing dataset we can test the machines to find the accuracy of a model.

Machine Learning Algorithms

Classification Algorithms- It is used to train the machine by using past data and then use this learning method to classify new observations. In this paper, Decision tree, Random forest, K Neighbor, Logistic regression, SVM and Naïve Bayes, SGDC algorithm to train the machine.

i. **Decision Tree** is also called a Classification and Regression Tree (CART). It is a supervised algorithm used for both classification and regression purposes. It represents the data in the form of a tree, branches represent all possible choices and the leaf node represents the final decision. This model is best suited for the less noisy dataset.

ii. **Naive Bayes algorithm** It is a supervised classification algorithm that works based on the principle of the conditional probability of the occurrence of an event based on past knowledge that will help correlate to an event.

iii. **Random Forest algorithm** is applicable for both regression and classification problems. It is an ensemble method, for a continuous dataset it will consider the average of all the trees and for classification dataset, it will choose a result in which maximum trees are voting**.**

iv. **Logistic Regression**: It is a classification and supervised algorithm. It has the dependent variable 'y' and independent variable 'x' and it will consider only discrete values. This algorithm uses a sigmoid function that will map the results between 0 and 1 (Based on threshold value)

v. **Support Vector Machine** is a supervised machine learning algorithm. It is used for regression or classification problems. It is a very simple and effective algorithm to solve the problem. This algorithm is used in the two-group classification problem.

vi. **K Neighbors classifier** is a supervised classification algorithm and it is used to classifies the data depends on the nearest neighbors. Here K represents the number of nearest neighbors. The nearest Neighbors algorithm is also called a lazy algorithm and it is best suited if the dataset is small and noise-free.

vii. **Shastic Gradient Descent (SGD)** is a simple algorithm and also a more efficient approach. It is applied linear classifiers and regressions under convex loss functions such as Support Vector Machine and Logistic Regression.

3. RESULTS

This paper contains tweets of social issues from the year 2006 to year July 2020 and stored into a CSV file and after that preprocessing has been done using the NLTK module. These preprocessed tweets are labeled using the Text Blob library in Python. Tweets are labeled as positive, negative, and neutral. The details of tweets labeling are shown below in Table 1.

Table 1: No. of tweets as Neutral, Positive and Negative

Class label	Percentage of tweets (%)
Negative	16.6%
Neutral	45.5%
Positive	37.9%

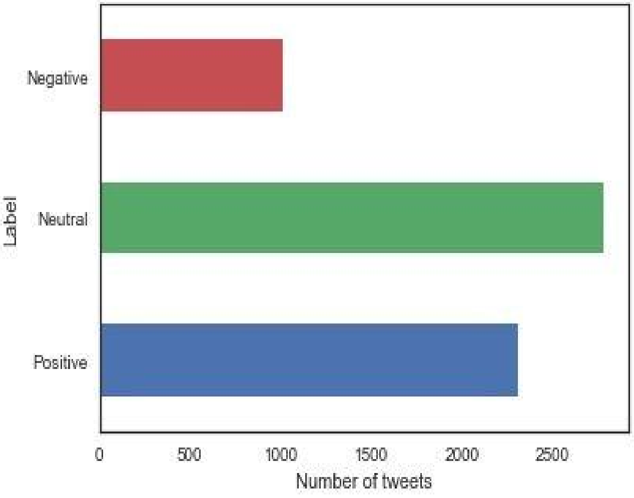

Figure 7: Sentiment Analysis

Figure 7 shows that tweets are labeled as positive, negative, and neutral. Neutral tweets are higher than positive and negative tweets. Negative tweets are less than positive tweets.

This paper evaluated the accuracy of different classifiers on the tweet dataset and compared them based on their performance.

In this paper, also calculated the Precision, Recall, and F1 score of each classifier to enhance the accuracy of each model.

Table 2: Precision, Recall, and F1 score of a Decision Tree Classifier.

Decision Tree Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.95	0.83	0.88
Neutral	0.63	0.83	0.72
Positive	0.86	0.92	0.89
Avg/total	0.88	0.86	0.86

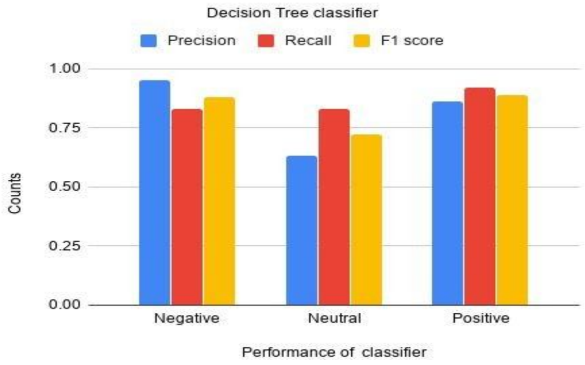

Figure 8: Performance of Decision Tree Classifier

Table 2 and figure 8, shows that the Precision, Recall, and F1 score of Decision Tree classifier, Precision is higher in negative sentiment, Recall is higher in neutral sentiment and F1 score in positive sentiment.

Table 3: Precision, Recall and F1 score of Naïve Bayes Classifier

Naïve Bayes Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.88	0.87	0.87
Neutral	0.67	0.92	0.78
Positive	0.92	0.82	0.87
Avg/total	0.87	0.85	0.86

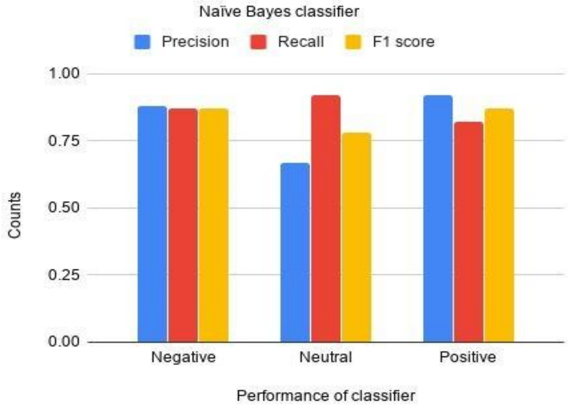

Figure 9: Performance of Naïve Bayes Classifier

Table 3 and Figure 9 shows that Precision, Recall, and F1 score of Naïve Bayes classifier, precision is higher in negative sentiment, Recall is higher in neutral sentiment and F1 score is higher in positive sentiments.

Table 4: Precision, Recall and F1 score of Random Forest Classifier

Random Forest Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.98	0.82	0.90
Neutral	0.70	0.94	0.80
Positive	0.86	0.96	0.90
Avg/total	0.90	0.88	0.89

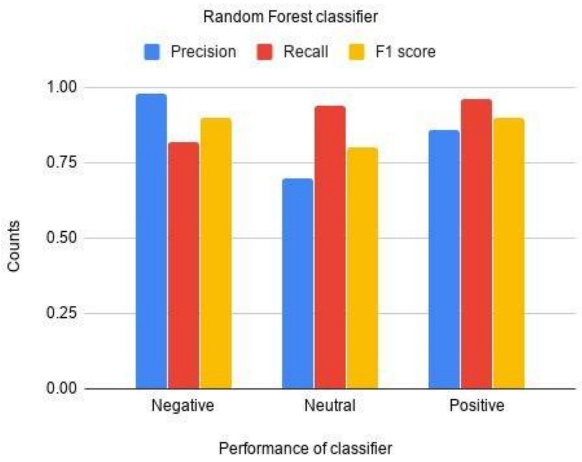

Figure 10: Performance of Random Forest Classifier

Table 4 and Figure10 shows that Precision, Recall, and F1 score of Random Forest Classifier. Precision is higher in negative sentiment, Recall is higher in neutral sentiment and F1 score is higher in positive sentiments.

Table 5: Precision, Recall and F1 score of Logistic Regression classifier

Logistic Regression Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.96	0.81	0.88
Neutral	0.60	0.99	0.75
Positive	0.88	0.91	0.89
Avg/total	0.89	0.87	0.87

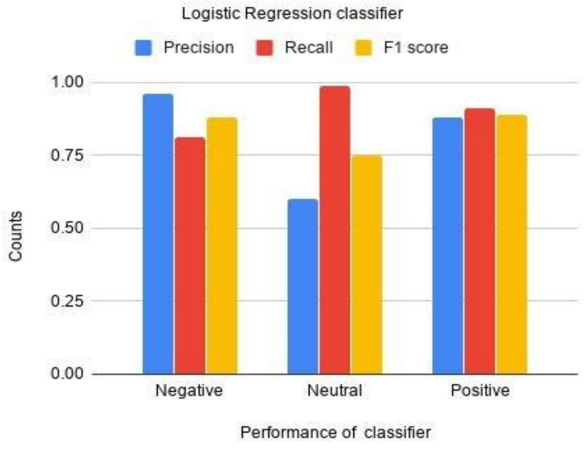

Figure 11: Performance of Logistic Regression Classifier

Table 5, Figure 11 shows that Precision, Recall, and F1 the score of Logistic Regression classifier, Precision is higher in negative sentiment, Recall is higher in neutral sentiment and the F1 score is higher in positive sentiments.

Table 6: Precision, Recall and F1 score of Support Vector Machine

Support Vector Machine Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.96	0.86	0.91
Neutral	0.73	0.96	0.83
Positive	091	0.92	0.91
Avg/total	0.91	0.90	0.90

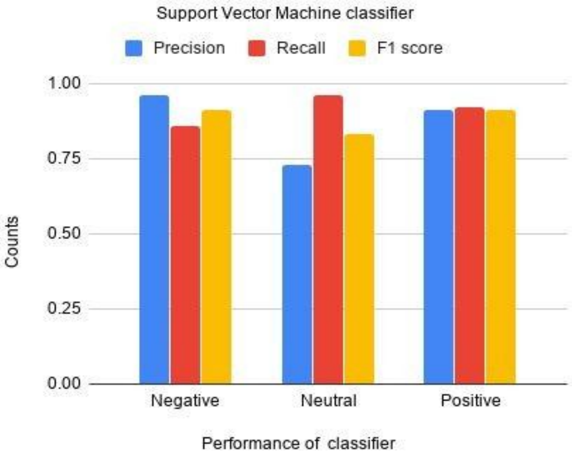

Figure 12: Performance of Support Vector Machine

Table 6, Figure12 shows that Precision, Recall, and F1 score of Support Vector Machine Classifier. Precision is higher in negative sentiment, Recall is higher in neutral sentiment and F1 score is higher in positive sentiments.

Table 7: Precision, Recall and F1 score of K Neighbors classifier

K Neighbors Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.86	0.71	0.78
Neutral	0.54	0.88	0.67
Positive	0.72	0.78	0.75
Avg/total	0.78	0.75	0.76

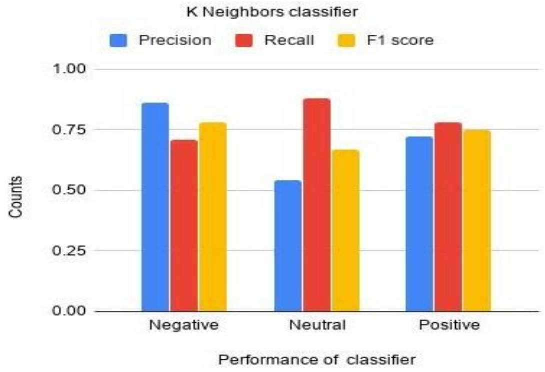

Figure 13: Performance of K Neighbors Classifier

Table 7, Figure 13 shows that Precision, Recall, and F1 score of K Neighbor Classifier. Precision is higher in negative sentiment, Recall is higher in neutral sentiment and F1 score is higher in positive sentiments.

Table 8: Precision, Recall and F1 score of SGD Classifier

SGD Classifier			
	Precision	**Recall**	**F1 score**
Negative	0.96	0.88	0.92
Neutral	0.75	0.98	0.85
Positive	092	0.91	0.92
Avg/total	0.91	0.91	0.91

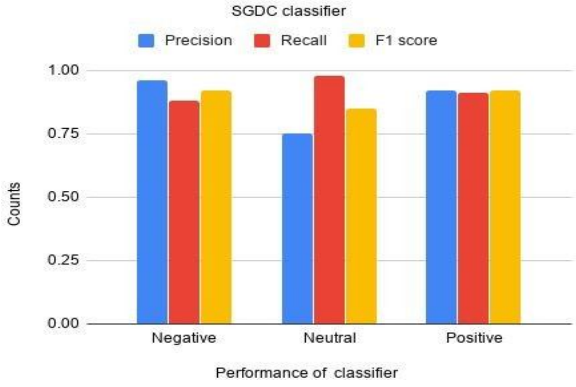

Figure 14: Performance of SGDC Classifier

Table 8 and Figure 14 show that Precision, Recall, and F1 score of SGD classifier. Precision is higher in negative sentiment, Recall is higher in neutral sentiment and F1 score is higher in positive sentiments.

Table 9: Accuracy of each model with the name of the classifier.

Classification Algorithm	**Accuracy (%)**
Naïve Bayes	85.49
Support Vector Machine	89.91
Decision Tree	86.14
Random Forest Classifier	88.36
K Neighbors Classifier	75.32
Logistic Regression	86.63
SGDClassifier	90.57

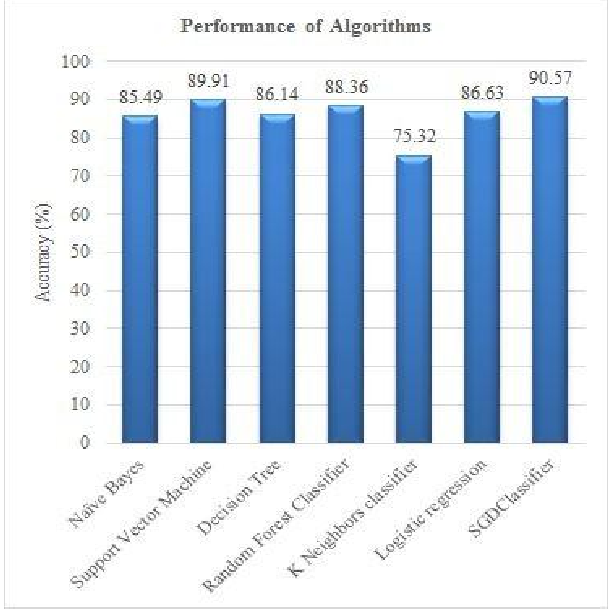

Figure 15: Performance of Classifiers

Table 9 and Figure 15 shows that the accuracy of each classifier, the accuracy of Naïve Bayes is 85.49%, the accuracy of support vector machine is 89.91%, the accuracy of Decision Tree is 86.14%, the accuracy of random forest Classifier is 88.36%, the accuracy of K Neighbors classifier is 75.32 %, the accuracy of logistic regression is 86.63% and accuracy of SGDC classifier is 90.57%.. The accuracy of the SGDC classifier 90.57 % is higher than other classifiers. So the SGDC classifier is more effective than other models.

4. CONCLUSION

In the conclusion of this paper, Sentiments of social issues are analyzed that are very critical issues of the globe. In this paper, 16.6% of tweets are negative 45.5% tweets are neutral, and 37.9% tweets are positive. It shows that the rate of negative sentiments is lower than neutral and positive sentiments. Various classifiers like Support Vector Machine, Logistic Regression, Random Forest Classifier, Naïve Bayes, Decision Tree, SGD classifier are applied to build an effective model and to enhance the accuracy of models. Accuracy of Naïve Bayes is 85.49%, the accuracy of support vector machine is 89.91%, the accuracy of Decision Tree is 86.14%, the accuracy of random forest Classifier is 88.36%, the accuracy of K Neighbors classifier is 75.32 %, the accuracy of logistic regression is 86.63% and accuracy of SGDC classifier is 90.57%.

Accuracy of SGDC classifier 90.57 % that is higher than other classifiers. Also measure the precision, recall, and F1- score of each classifier. This research shows that people are aware of their rights and self-security but still need more awareness to be secure and safe to live with freedom in society

REFERENCES

1. M. Bouazizi and T. Ohtsuki, "**A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter**," *IEEE Access*, vol. 5, pp. 20617–20639, 2017
2. A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "**Machine Learning-Based Sentiment Analysis for Twitter Accounts**," *Math. Comput. Appl.*, vol. 23, no. 1, p. 11, 2018 https://doi.org/10.3390/mca23010011
3. M. Masdisornchote, "**A Sentiment Analysis Framework for Social Issues**," *41st Annu. Conf. Ind. Electron. Soc. (IECON 2015)*, no. May, pp. 357–361,2015
4. R. De Groot, ***Data Mining for Tweet Sentiment Classification***. 2012.
5. A. T. M. Shahjahan and K. U. Chisty, "**Social Media Research and Its Effect on Our Society**," *Int.J. Soc. Behav. Educ. Econ. Bus. Ind. Eng.*, vol. 8, no. 6, pp. 2009–2013, 2014.
6. J. Xue, J. Chen, and R. Gelles, "**Using Data Mining Techniques to Examine Domestic Violence Topics on Twitter,**" *Violence Gend.*, vol. 6, no. 2, pp. 105–114, 2019
7. B. Liu, **"Sentiment Analysis and Subjectivity,"** *To Appear Handb. Nat. Lang. Process. Second Ed.*, vol. 2, pp. 1–38, 2010.
8. J. Smailović, "**Sentiment Analysis in Streams of Microblogging Posts,"** *Kt.Ijs.Si*, p. 140, 2014,
9. A. Shamshad, **"Corruption: A Social Evil in India,**" *Indian J. Public Adm.*, vol. 57, no. 3, pp. 806–817, 2011
10. A. Shahid and M. Sumbul, **"Social Evils in Media: Challenges and Solutions in 21St Century**," *PEOPLE Int. J. Soc. Sci.*, vol. 3, no. 3, pp. 854–875, 2017
11. W. Medhat, A. Hassan, and H. Korashy, **"Sentiment analysis algorithms and applications: A survey**," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014
12. M. P. Abraham and K. R. Udaya Kumar Reddy, **"Feature-based sentiment analysis of mobile product reviews using machine learning techniques**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 2289–2296, 2020, doi: 10.30534/ijatcse/2020/210922020.
13. O. R. Devi, **"A Deep Analysis on Aspect based Sentiment Text Classification Approaches,"** *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 1795–1801, 2019. https://doi.org/10.30534/ijatcse/2019/01852019
14. A. D'Andrea, F. Ferri, P. Grifoni, and T. Guzzo, **"Approaches, Tools and Applications for Sentiment Analysis Implementation,"** *Int. J. Comput. Appl.*, vol. 125, no. 3, pp. 26–33, 2015, doi: 10.5120/ijca2015905866.
15. M. Masdisornchote**, "A Sentiment Analysis Framework for Social Issues,"** *41st Annu. Conf. Ind. Electron. Soc. (IECON 2015)*, no. May, pp. 357–361, 2015, doi: 10.1038/s41598-017-14885-w.
16. Z. Jianqiang and G. Xiaolin, **"Comparison research on text pre-processing methods on twitter sentiment analysis,"** *IEEE Access*, vol. 5, pp. 2870–2879, 2017.
17. L. C. Yu, J. Wang, K. Robert Lai, and X. Zhang, **"Refining Word Embeddings Using Intensity Scores for Sentiment Analysis,"** *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 3, pp. 671–681, 2018, doi: 10.1109/TASLP.2017.2788182.
18. C. Clavel and Z. Callejas**, "Sentiment Analysis: From Opinion Mining to Human-Agent Interaction,"** *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 74–93, 2016, doi: 10.1109/TAFFC.2015.2444846.
19. R. R. Putra, M. E. Johan, and E. R. Kaburuan, **"A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia,"** *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 1856–1860, 2019. https://doi.org/10.30534/ijatcse/2019/07852019
20. M. Bouazizi and T. Ohtsuki, **"A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter,"** *IEEE Access*, vol. 5, pp. 20617–20639, 2017.