iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.58962

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Hate Speech Detection Using NLP and Machine Learning

Sasanka Boothati[1], Prof. Humera Khanam M[2], MD. A. Khudhus[3]
Department of Computer Science and Engineering, Sri Venkateswara University College of Engineering, Tirupati-517501

Abstract: ***The use of social media has been growing in an eccentric fashion making it a medium for sharing opinions, ideas, and thoughts of an individual with others. This has made things complex with what is considered a genuine comment or rather a hypocritic deliberative nuance to damage or incite hatred on an individual or a group belonging to a community, race, gender, nationality, etc, In this paper, the detection of hate speech with the use of sentiment polarity scores and the Term Frequency Inverse Document Frequency(TFIDF) scores with machine learning algorithms is to decrease the true negatives and false positives by the use of Natural Language Processing. The Machine Learning algorithms used are Logistic Regression and Random Forest Classifier. The phases of NLP are done to preprocess the tweets that are available on the Kaggle with about 25 thousand tweets from the social media giant "Twitter". The processed tweets are then with the use of two ML Algorithms trained for vaderSentiment polarity scores and TFIDF scores from which metrics are obtained. The results of sentiment polarity scores(7 points) are less accurate in the detection of hate speech as compared to TFIDF scores(8 points).***

Keywords: Natural Language Processing, Machine Learning, Sentiment Polarity Scores, TFIDF vectorization, Random Forest Classifier, Logistic Regression, Confusion Matrix.

I. INTRODUCTION

Hate Speech has been a growing problem for social media users as the ill effects are not only affecting an individual but also disturbing the harmony in society. The deliberate incitement has far more repercussions leading the nations to look into their Social Media guidelines to reduce them.

The constitution of India entitles its citizens with the Freedom of speech and expression as their fundamental right which also provides for an aggrieved person to directly file a case in the Supreme Court or High Court. Nowadays the triggers of hate speech are reverberating through the world with a frequency that is fast becoming intolerable. This calls for a better model for hate speech detection because often hate speech is disguised as offensive speech which causes all the damage. Nevertheless, this calls for an approach that can be used to provide better results.

This approach can be able to include the features of Natural Language Processing which can be efficient in understanding the nuances posed in the hate speech in better training of the model and the addition of Machine Learning algorithms can trigger hate speech with accuracy. The use of sentiment polarity scores and TFIDF scores complement bringing the better model

II. RELATED WORK

Hate Speech Detection has become a growing problem since the rapid increase in globalization. The First work on this was done in *T-Davidson's Experiment.* Many papers were published in the Hate Speech detection area. The major work was increasing the accuracy and efficiency of the importance of detection. Many methods are used along with different techniques to achieve the objective. The use of a keyword-based approach has been a primitive one and has had many true negatives and false positives leading to decreased confidence. Then Machine Learning Algorithms were used which developed a better insight into hate speech detection which was better but wasn't sufficient to reach the goal. The use of Natural Language Processing has been a breakthrough for text analysis and especially for Hate Speech Detection because hate speech is not just words but the expression as a whole which wasn't possible in computer language models and algorithms. Then using Natural Language Processing, there are many algorithms and methods used. Such as the doc2Vec method, deep learning method, bi-long Term Short Memory Recurrent Neural Networks, Genetic Programming, and Supervised and Unsupervised Machine Learning Algorithms. From all the supervised and unsupervised models Naïve Bayes classifier with TFIDF features performed best with an F-score of 0.719. These are the related work on Hate Speech detection and the following topic shows us the design and analysis part of this model.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

III.METHODOLOGY

A. Design

1) Step 1: Preprocessing of tweets for better interpretation and detection as shown in Figure 1.

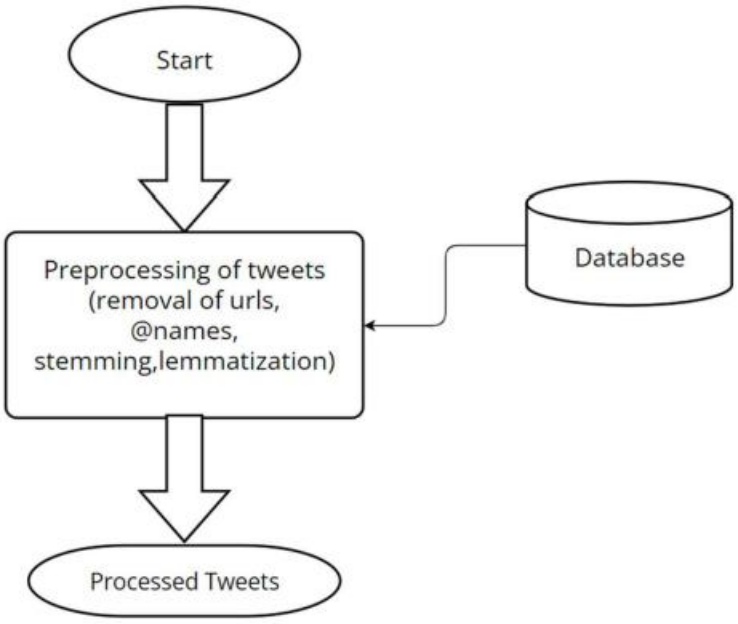


Figure 1: Preprocessing of tweets from Twitter dataset.

2) Step 2: Feature Engineering with both methods of TFIDF Vectorization and Sentiment Polarity Analysis and get evaluation matrices for both methods to compare as shown in Figure 2.

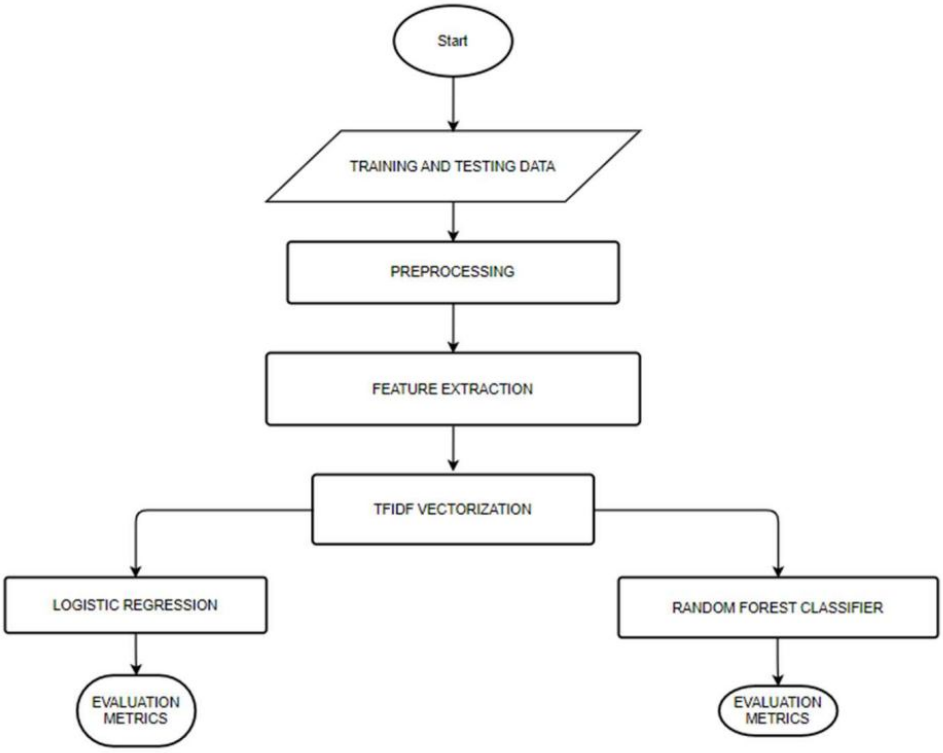


Figure 2: Feature Extraction for TFIDF Vectorization for Logistic Regression and Random Forest Classifier.

B. Methods and Algorithms

The methodology includes two methods for the detection of Hate Speech Detection using Sentiment Polarised Analysis(SPA) and TFIDF Vectorization for two Machine Learning (ML) Algorithms i.e, Logistic Regression(mostly used for Text Analysis) and Random Forest (given the big data associated with the training and testing data along with a desirable algorithm for Text Classification). These ML Algorithms are compared using evaluation metrics such as Accuracy, F1 score, Precision, and Recall. The confusion Matrices are also evaluated for algorithms in each of the methods.

C. Analysis

The major drawback of Hate Speech Detection is the scope for instances of hate speech is less leading to greater nonhate speech to hate speech which is contrary to the reality that we often notice in Social Media Platforms.

The TFIDF Vectorization takes the regular expression into consideration which can be able to increase the triggers of hate speech as we can see in the findings part. Natural Language Processing has been a great tool to increase the understandability of the model to notice the sarcastic comments and also feed the model with enough fuel to equip and detect the hate speech accurately.

IV. RESULTS

A. Findings

Evaluation metrics for Sentiment Polarity Scores for ML Algorithms:

1) Accuracy of Logistic Regression using Sentiment Analysis is shown in Figure 3:

Accuracy of Logistic Regression using Sentiment Polarity scores: 0.7609441194270729
Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	290
1	0.78	0.97	0.86	3832
2	0.33	0.08	0.13	835
accuracy			0.76	4957
macro avg	0.37	0.35	0.33	4957
weighted avg	0.66	0.76	0.69	4957

Figure 3: Evaluation Metrics for Logistic Regression using Sentiment Analysis

2) The accuracy of the Random Forest Classifier using Sentiment Analysis is shown in Figure 4:

Accuracy of Random Forest Classifier using Sentiment Polarity scores: 0.7694169860802905
Classification Report:

	precision	recall	f1-score	support
0	0.07	0.00	0.01	290
1	0.78	0.98	0.87	3832
2	0.44	0.05	0.10	835
accuracy			0.77	4957
macro avg	0.43	0.35	0.32	4957
weighted avg	0.68	0.77	0.69	4957

Figure 4: Evaluation metrics using Sentiment Analysis

3) Accuracy of Logistic Regression using TFIDF Vectorization is shown in Figure 5:

```
Accuracy of Logistic Regression using TFIDF Vectorization: 0.8892475287472261
Classification Report:
              precision    recall  f1-score   support

           0       0.48      0.19      0.27       290
           1       0.91      0.96      0.94      3832
           2       0.83      0.81      0.82       835

    accuracy                           0.89      4957
   macro avg       0.74      0.65      0.67      4957
weighted avg       0.87      0.89      0.88      4957
```

Figure 5: Evaluation metrics of Logistic Regression using TFIDF Vectorization

4) The accuracy of the Random Forest Classifier using TFIDF Vectorization is shown in Figure 6:

```
Accuracy of Random Forest Classifier using TFIDF Vectorization: 0.886624974783135
Classification Report:
              precision    recall  f1-score   support

           0       0.43      0.10      0.16       290
           1       0.91      0.96      0.93      3832
           2       0.83      0.80      0.82       835

    accuracy                           0.89      4957
   macro avg       0.72      0.62      0.64      4957
weighted avg       0.86      0.89      0.87      4957
```

Figure 6: Evaluation metrics of Random Forest Classifier using TFIDF Vectorization

B. Comparison of Algorithms with each Method

Method/Algorithm	Random Forest Classifier	Logistic Regression
1. TFIDF vectorization	0.8866	0.8892
2. Sentiment Polarised Analysis	0.7694	0.7609

Table: Accuracy of different methods corresponding to different ML algorithms

Figure 1 is a bar graph on the F1 scores of Logistic Regression and Random Forest Classifier which are associated with a slight difference.

The F1 score of Logistic Regression is 0.76 and that of the Random Forest Classifier is 0.7694 for the method of SPA.

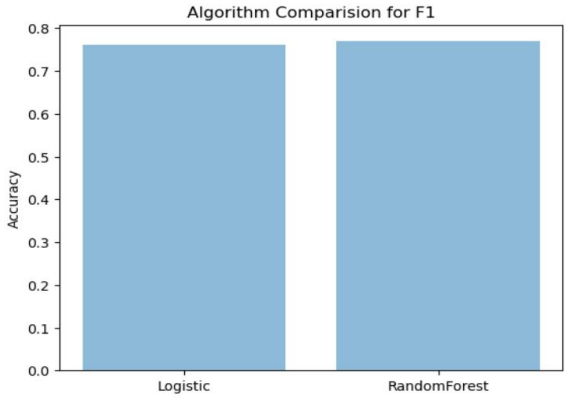


Figure 1: Bar plot on Algorithm Comparison for F1 score

Figure 2 is a barplot with subplots giving the algorithm comparison with the different evaluation metrics.

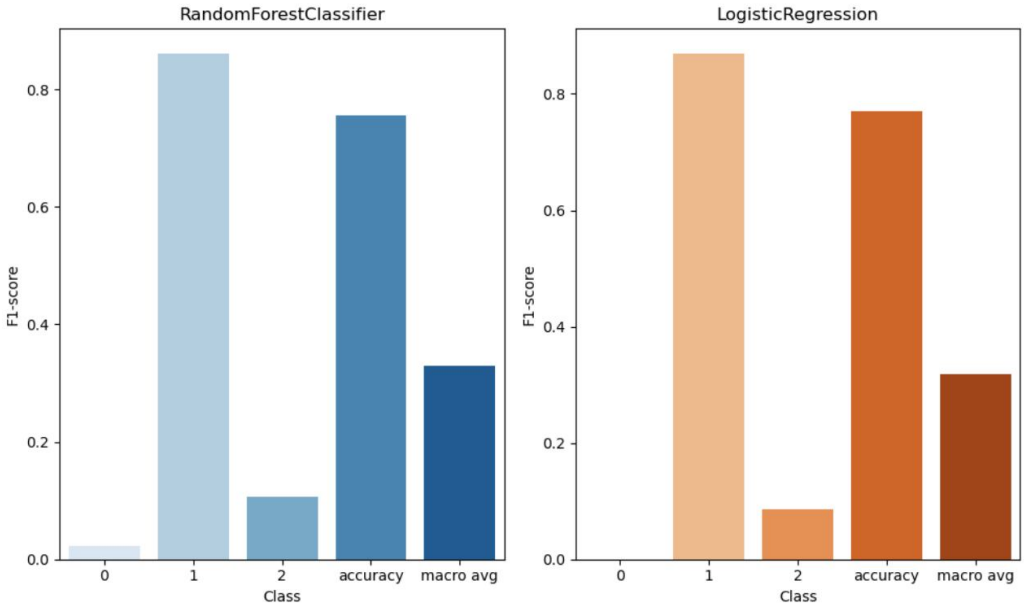


Figure 2: Algorithm Comparison of Different Evaluation Metrics

C. Discussion

The Findings of the model give the insights that the accuracy with the use of TFIDF Vectorization has a 100-point edge over that of the Sentiment Polarised Analysis. Concerning TFIDF vectorization we can see that the Random Forest Classifier is outperformed by Logistic Regression. In the same way concerning the Sentiment Polarised Analysis the Logistic Regression is outperformed by Random Forest Classifier.

We can also find that there is a very narrow variation in terms of Accuracy evaluation because of the size of the dataset that we have chosen. It also depends on the ratio with which the training and testing datasets are fed to the model.

V. CONCLUSION

The Hate Speech Detection Model has shown us that the use of TFIDF Vectorization can bring about greater accuracy in which the use of a Random Forest Classifier is highly recommended. It is so because the major complication of Logistic Regression is that it cannot handle the Big Data that has been exponentially rising with the users of social media increasing rapidly. This also gives scope for stakeholders to express their opinions rather more openly and this can be a cause of contention. This much amount of data processing can be a hurdle for Hate Speech Detection using Logistic Regression.

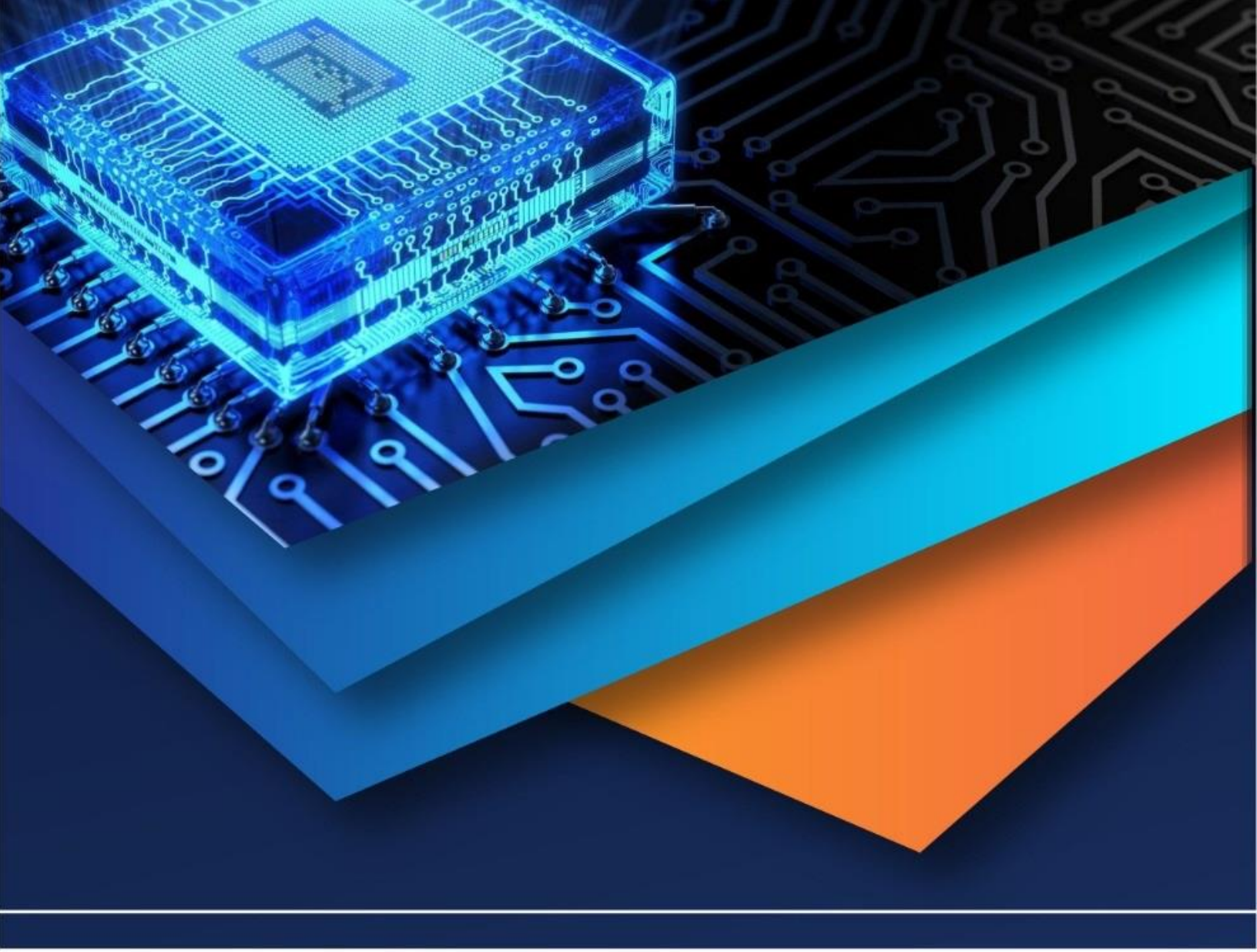
VI. FUTURE SCOPE

The Future Scope of Hate Speech is brighter with the use of Random Forest Classifier with TFIDF Vectorization for better results as Hate Speech can be a major hurdle for growth and development in the Globalising world that we are in. Worldwide nations are in search of the best technique for Hate Speech Detection because the harmony in this rapidly growing digitalized world can make it impossible to peacefully exist in a digital space that is proving to be entwined with an individual's personal and professional life.

REFERENCES

[1] Hate speech detection: Challenges and solutions by Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, Ophir Frieder
[2] M. H. Khanam, M. A. Khudhus and M. S. P. Babu, "Named Entity Recognition using Machine learning techniques for Telugu language," 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2016, pp. 940-944, doi: 10.1109/ICSESS.2016.7883220
[3] MACHINE LEARNING AND DEEP LEARNING TECHNIQUES: Sentiment Analysis Using Machine Learning and Deep Learning Techniques by M Humera Khanam.
[4] P. P. Jemima, B. R. Majumder, B. K. Ghosh and F. Hoda, "Hate Speech Detection using Machine Learning," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 1274-1277, doi:10.1109/ICCES54183.2022.9835776.
[5] C. Paul, "Hate Speech in Social Networks and Detection using Machine Learning Based Approaches," 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC), Silchar, India, 2023, pp. 1-7, doi:10.1109/ISACC56298.2023.10084222

[6] P. Patil, S. Raul, D. Raut and T. Nagarhalli, "Hate Speech Detection using Deep Learning and Text Analysis," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 322-330, doi: 10.1109/ICICCS56967.2023.10142895.
[7] N. D. T. Ruwandika and A. R. Weerasinghe, "Identification of Hate Speech in Social Media," 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2018, pp. 273-278, doi: 10.1109/ICTER.2018.8615517.
[8] M. K. A. Aljero and N. Dimililer, "Hate Speech Detection Using Genetic Programming," 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 2020, pp. 1-5, doi: 10.1109/ICOASE51841.2020.9436621.
[9] Speech and Language Processing by Daniel Jurafsky and James H.Martin
[10] https://vitalflux.com/hate-speech-detection-using-machine-learning/
[11] Hate Speech Detection Using Machine Learning Suraj Futane1, Twinkal Bandwal2, Dnyaneshwari Dhonde3, Sakshi Gudmewar4, Aishwarya kadam5
[12] https://en.wikipedia.org/wiki/Hate_speech



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY