Hate Speech in Social Networks and Detection using Machine Learning Based Approaches

Chayan Paul
Amity Institute of Information Technology
Amity University Kolkata
Kolkata,
chayan.aus@gmail.com

2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC) | 979-8-3503-9938-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISACC56298.2023.10084222

***Abstract*— The use of social networking sites has increased considerably in last few years and as a result the user generated contents in the web also increased manifold. These data are mostly present in unstructured and quasi-structured formats. Many social media platforms are being affected due to the presence of hate speech. It is present in many forms such as verbal aggression and through photos which we know as memes and so on. This study considers twitter data for detecting hate speech on the internet. From the past few years, machine learning and natural language processing approaches are being used to detect hateful content on the web. This study aims to deal with the problem of hate speech detection in text data using machine learning algorithms. Feature selection for the dataset was performed before passing the dataset to the machine learning models. Different machine learning algorithms are implemented on an open-source twitter dataset. Performance of different algorithms are compared using standard performance measure metrics and presented in this paper. The experimental result shows artificial neural network outperforms the other algorithms considered in this study.**

Keywords—: Machine Learning, Social Media, Hate Speech, Natural Language Processing

I. INTRODUCTION

In the growing world of online presence with all the social media applications like Instagram, Facebook, twitter etc., trying to connect everyone in the world with a device like smartphone or computer and internet is possible. Knowing the potential of internet connectivity, the internet users have grown a lot with time. Though it made our lives easier, it has also created a whole new dimension of problems. With the growing number of users and the user generated online contents, it has become difficult to control the flow of hateful contents in these platforms. From the past decade, the use of Machine Learning and Natural Language Processing techniques have become popular to detect the harmful content from this enormous amount of data which is created every second.

The term hate simply means "intense dislike" which could be towards a person or a thing or an idea. Humans tend to exhibit feelings like happiness, sadness, anger, love and hate. In this modern era of human civilization living in a global village, everyone is moving towards digitalization and all-time connectivity through the internet via social media platforms. Users may not define feelings in words but can understand them in a physical environment through one's tone, actions, posture and facial gestures. In the case of online social media platforms, the communication is done through messages and multimedia which conveys information. Whereas here people lack the basis of understanding of one's feelings like hate due to inadequate judgement since it is an online interface. Hate in social media is exhibited through verbal aggression where words forming a sentence play a major role in conveying hate.

Practice of freedom of speech often leads to a lot of hate speech in these online platforms. Hate speech and online bullying comes under one single term. Hate speech indicates any means verbal or nonverbal communication which assaults a person or a group of persons on the basis of their cast, creed, religion, sexual preferences, and other issues related to their identity, choices. Hate speech undermines the sense of assurance, kills the confidence, increases the anxiety and distress. These lead to the bad mental health of an individual who is a victim of hate speech. The hate speech of a particular community or minorities will lead to an imbalance in society by making the lives of a particular set of individuals difficult. In some cases, the mass genocide of communities and minorities happen and that all is due to hate speech.

With the growing user-generated content on social media applications like Instagram, Facebook, Twitter etc. it is getting harder to differentiate between offensive, explicit and hate speeches. Since the mode of hate exhibiting by the users is through words and the words simply do not mean anything by themselves unless we try to understand them with a context.

- Detecting hate speech in social media automatically is technically difficult.
- Approach to the issues will lead to reasonable and efficient solutions.
- The solutions obtained so far still has some challenges
- Without a proper context, the system cannot detect sufficiently based on words only.

The work here focuses on hate speech text detection by exploring natural language processing and machine learning techniques for the purpose of identifying hate speech on social media platforms such as twitter. The objective is to get the insights of how different classification techniques can perform on the social media dataset to get the result of the most appropriate technique that can be used for pragmatic hate speech detection. Therefore, different pre-selected feature generation, resampling and classification algorithms are combined and applied on a social media dataset.

II. REVIEW OF LITERATURE

Sentiment analysis or polarity detection in twitter data has been one of the highly researched in the recent past. Researchers mainly focus on finding out users experience towards a any product, service or an event. The data are mainly collected from social networking sites like Twitter and analysed to obtain the intended result [1] [2] [3] [4]. Other research works those were triggered by the enormous amount of data available in the social networks are listing users with similar interest in a product or a service, [5] [6]; identifying violent contents in the social network data [7] [8]. The data collected from social networks are predominantly unstructured and the existing methodologies often find it

difficult to analyse these data. Researchers did good number of works which led to enhancement in these methodologies [9] [10]. It has often been in reports that violence between communities or groups break out and the social media posts seem to be primary reason for that. [11], [12]. These series of events encourage researchers to detect the violent contents in social networking sites.

Hate speech may be depicted as interaction among users with verbal or nonverbal contents that contain high decibel of fanaticism and hostility [13]. Hate speech can also contain abusive contents based on people's religion, race, gender, political association [14]. Regular argument with offensive content hurts the self-respect of the people and these phenomena may lead to negative impact in the society [15].

Researchers from the domain of pure natural language have been working on developing models those can classify if a sentence contains hate speech or not [16], [17]. Intuitively models of these types are computationally slower than their counterparts developed using machine learning and soft computing algorithms. One requirement that the models developed using machine learning has been those need datasets to be labelled as we need supervised learning models for conducting these tasks. Researchers have been working continuously in this area and many of the researchers developed their own datasets. One of the normal approaches in this type of studies is to extract the data from the social networking sites and then label those messages if they contain hateful messages or not. Ahmed et. el., prepared a dataset containing English and Bengali mixed corpus and followed a rigorous process to label the data binary labels. [18]. Sahi et. el., worked on Turkish language corpus and built a supervised learning model which can be used to detect hate speech against women. For this study they collected data which has mentions of clothing choices by women and used that dataset for developing machine learning model [19]. Annotators play a major role in developing the data set and Waseem studied the influence of annotators on different classification models [20]. In a separate study, Waseem et. el., built a dataset containing 16000 tweets and they also investigated to find the important features which may lead to improvement of model performance [21]. Considerable amount of research work have been carried out taking open source data sets and using those dataset for building machine learning models for detecting hate speech [22] [23] [24].

Ayo et. el., examined the research on machine learning methods for classifying hate speech and offered a metadata framework. It was discovered that the metadata architecture produced better outcomes. [25]. For the model to learn from different abstractions of the problem, Nascimento et al. examined gender bias in a different dataset and developed an ensemble learning strategy based on various feature spaces for hate speech identification. Our technique was tested on a publically accessible corpus and trained on nine distinct feature spaces, and the findings show that it performs better than state-of-the-art options. [26]. Ali et al. created a hatred lexicon in Urdu and assembled a dataset of 10,526 annotated tweets. Additionally, they used transfer learning and machine learning strategies to get improved outcomes. [27]. By merging the results of transfer and deep learning-based methods, Roy et al. proposed an ensemble model. Their model outperformed the most advanced model in terms of results. [28]. A deep multitask learning framework was presented by Kapil et al. to use relevant data from other related categorization tasks to enhance the performance of the individual task. The shared-private scheme, which allocates shared and private layers to capture shared features and task-specific information from five categorization jobs, is the foundation of the proposed multi-task model. The suggested framework achieves encouraging performance in terms of macro-F1 and weighted-F1, according to experiments1 on the five datasets. [29]

Researchers often went beyond only binary classification for hate speech data and extended this to multiclass classification. Watanabe et. el., in their study, used twitter data to develop model which can classify the tweets into three different classes. [30]. Kumar et. el used Facebook messages and developed machine learning models for classifying tweets into three classes namely Aggressive, Covertly Aggressive, and Non-aggressive [31]. Ali et. el., developed an Urdu language hate lexicon and developed an annotated dataset of 10526 texts. They applied different machine learning algorithms for hate speech detection [32].

Open-source data set from Kaggle were collected for this study. The tweets contain data from American users regarding various issues[33]. Using these data supervised machine learning models were built for detecting hate speech. There are three class labels in the data set namely hate speech, offensive language and neither. The performance measures of these models are presented in the results section of the paper.

III. Methodology

As the work in this study is a supervised classification task, this paper used the data set to train logistic regression, naïve bayes classifier, support vector machine and artificial neural networks. These are some of the most used supervised learning tasks used in the machine learning.

A. Logistic Regression:

Logistic regression is a classification algorithm that uses logistic function to calculate the probabilities of the classes to be assigned to an individual data item. For prediction a threshold value on the probability can be set and depending on the threshold, a class is assigned. For example, if the threshold value is 0.5, then in case of binary classification problems, 1 can be assigned to data having probability more than 0.5 and 0 to the data points with threshold less than or equal to 0.5. Predominantly this algorithm is used for binary classification, but it is also found working well with multi class classification problems as well.

B. Naïve Bayes Classifier

Naïve Bayes Classifier is one of the most widely used probabilistic classification algorithm which is built upon the powerful concept of Bayes Theorem. The algorithm assumes that the classes are mutually exclusive and features in the dataset are independent of each other, hence it takes the prefix naïve in front of it.

C. Support Vector Machines

Support vector machine is a supervised learning algorithm widely used for classification. In this algorithm, data are plotted in an n dimensional space and classification was done by finding a hyperplane that clearly differentiates the data as per the class labels. The support vectors are mainly the coordinates if each data points. Major advantage of using

support vector machines is, this algorithm is computationally simpler.

D. Artificial Neural Network

Artificial neural network is a computational model that comprises of number of simple processing elements, which are highly interconnected. Artificial neural networks draw their motivation from the way human brain cells work. The basic computational unit in the network is known as neuron. There are different forms of artificial neural networks based on their architecture, like feed forward networks, multilayer perceptron, convolutional neural networks etc. This paper implements a multilayer perceptron for classification of the text data.

E. Data Pre-processing

The dataset consists of 24,783 text data which were collected from twitter using API. The hatebase.org compiled some of the words and phrases into a hate speech lexicon which are recognized by internet users as hate speech. The terms in this hate speech lexicon are used to capture the tweets from Twitter database using Twitter API. It results in the below 24,783 tweets. All these tweets have been labelled by CrowdFlower workers (human annotators) into one of these classes of "hate speech", "offensive language" and "neither".

	hate_speech	offensive_language	neither	class	tweet
0	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	0	6	0	1	!!!!!!!!!!! RT @ShenikaRoberts: The shit you...

Figure 1. First five rows of the dataset

It is evident from the figure 2(a) below that the dataset is highly imbalanced. The majority class, labelled offensive has 77% data. The class labelled Hate Speech has only 5.8%. This imbalance in data needs to be taken care of otherwise the classifiers trained with this imbalanced data may become highly biased. Thus, up sampling technique were used for pulling the number of minority class up.

Figure 2(a) shows the status of the data set in term of number of tweets before up sampling was applied, whereas figure 2(b) shows the same after the up sampling.

Investigation shows that differentiating between texts containing offensive and hateful messages is one of the very challenging tasks. Both the types share similar qualities and makes the job of differentiation much difficult

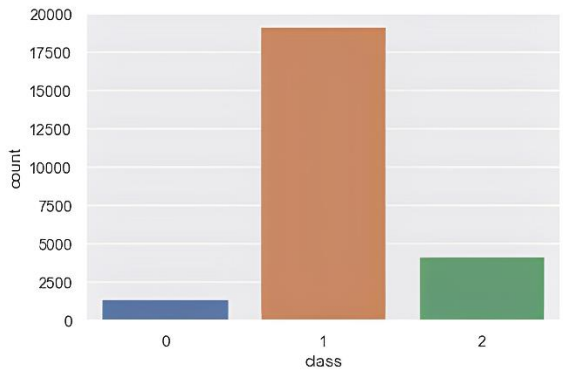

Figure 2(a) Bar diagram representing imbalanced class

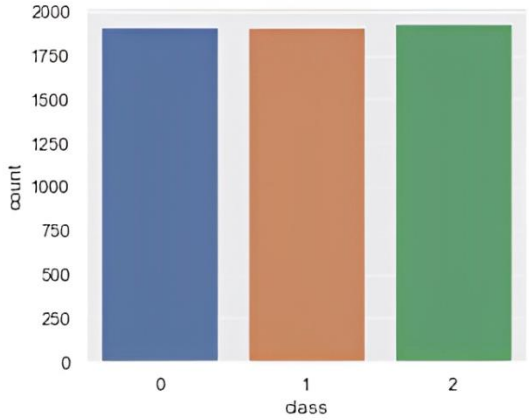

Figure 2(b) Bar diagram representing balanced class

The experiments conducted in this paper are comprised of four main steps: pre-processing, feature generation, oversampling and classification. In the pre-processing step the raw data is cleaned to remove unnecessary data to get better features for the following step. In the feature generation step, the cleaned data from previous step is transformed into feature vectors. Oversampling technique was applied to up sample the minority classes since the data set used here is highly imbalanced. The class which represents hate speech, is having small amount of data and this can cause erroneous model if not handled properly. Hence up sampling the minority class was done to bring the amount of data to similar to the majority class.

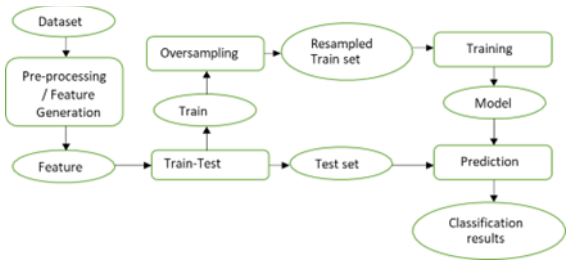

Figure 3 Experiment Pipeline

Balancing the class labels is one of the useful techniques used for improving classification results. Predominantly there are two major options for data balancing. In one of the approaches, the majority class is cut down to make the size of minority and majority class similar. But this approach results in loss of lots of data. This approach is primarily known as under sampling as one removes large amount of data in a particular class to make the count of data equivalent to the minority class. The other alternative is to pull the number of minority class up to make the amount of the data equivalent to the majority class. This approach is mainly known as oversampling as one resamples the data from the minority class to bring the amount of the data up. There are more advanced versions of data balancing tasks available in the literature. In this study synthetic minority oversampling technique was used as tool for data balancing. By far SMOTE is one of the state-of-the-art techniques used for balancing the data[34] and in this paper the same technique was used for balancing the data.

IV. Results and Analysis:

After review of the literature, four algorithms namely logistic regression, support vector machines, naïve bayes

classifier and multi layer classifier were selected for this study. During the review it was noticed that these four algorithms are mostly used ones in text classification problems. In this study, all these four algorithms were trained using the data set collected and then their performance were compared. The result section of this paper presents a detailed deliberation on the performance of these algorithms using the current data set.

Selected algorithms were implemented on the dataset and the evaluation were done. The confusion matrices were prepared first for all these four trained model for understanding the performances. A confusion matrix for a classification model presents the values which were predicted correctly by the model and also the values which were not predicted correctly by the model. The figures below shows the four confusion matrices prepared in the study.

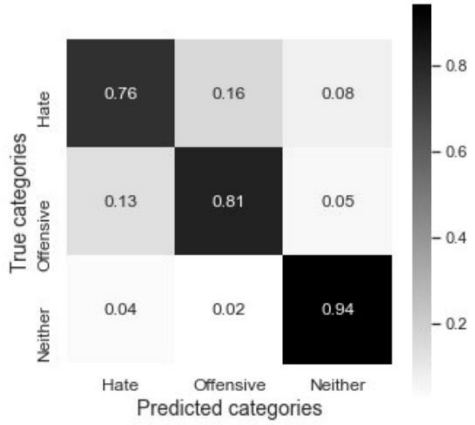

Figure 4: Confusion matrix for Logistic Regression

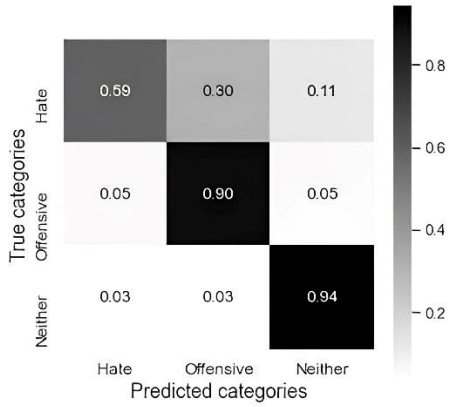

Figure 5: Confusion matrix for Support Vector Machine

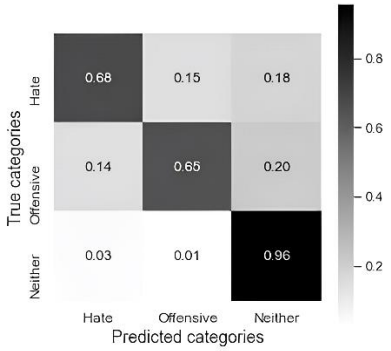

Figure 6: Confusion matrix for Naïve Bayes Classifier

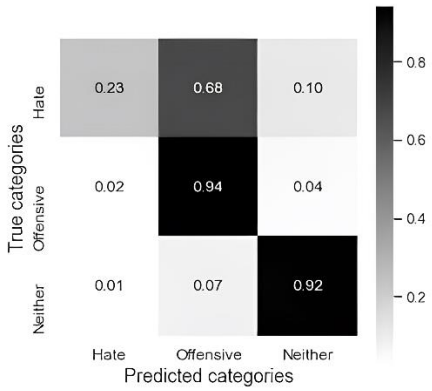

Figure 7: Confusion matrix for Neural Networks

Investigating the confusion matrices, artificial neural network is found to be a little better in comparison to the other models, but it is difficult get a clear picture of the best performer. Other performance measures viz., accuracy, precision, recall and f1 scores are also computed to find out the model which can classify the data with higher accuracy. The measures are discussed in very brief in the following section:

A. Accuracy:

One of the most used performance metrics is accuracy, which is calculated as the proportion of all correctly classified inputs to all observations. For a balanced dataset, accuracy is a good measure of performance, but in case of imbalanced data, accuracy may not give a good picture about the performance of the algorithms. From the experimental data in this study, artificial neural network is the clear lead in the group with a accuracy of 0.94. The formula for calculating accuracy is given below:

$$Accuracy = \frac{TP+NP}{TP+TN+NP+FN} \quad (1)$$

B. Precision:

Precision is measured as the ratio of correctly predicted positive values to the total number of positive values. If a model has a high value of precision, that means the model has a low false positive rate. In this study, neural network has a higher value of precision among all these algorithms. The formula for calculating precision is presented below:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

C. Recall

Recall measures the proportion of correctly predicted positive entries to all positive entries in the class. In essence, it reflects the percentage of positive observations that were correctly categorised. According to this study, again neural network is found to perform better than the other algorithms. The formula for calculating recall is presented below:

$$recall = \frac{TP}{TP+FN} \quad (3)$$

D. F1 score

The F1 score, which takes into account both false positive and false negative results, is the weighted average of accuracy and recall. F1 score becomes a stronger performance indicator than accuracy for a situation when the classes are unbalanced.

In this investigation, again ANN is found to perform better than other models in terms of F1 score also

$$f1\ score = \frac{2*recall*precission}{recall+precission} \quad (4)$$

The calculated values of the performance measures for all these algorithms are presented in the table below. A graphical representation of these measures also presented in the figure 8.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.85	0.87	0.96	0.92
Naïve Bayes	0.7	0.79	0.96	0.86
Support Vector Machine	0.84	0.87	0.96	0.91
Artificial Neural Networks	0.94	0.94	0.97	0.96

Table 1: Performance measures of the algorithms

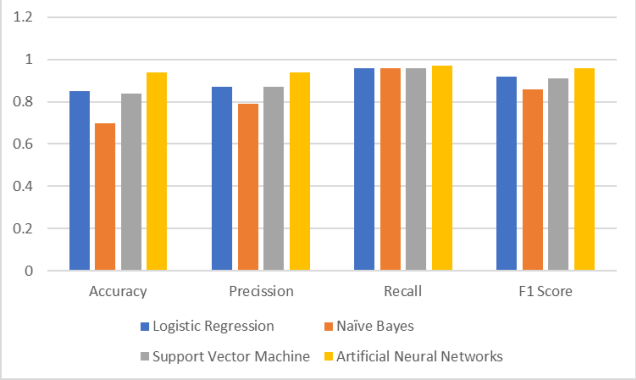

Figure 8: bar diagram for performance measures

V. Conclusion

The study finds that the performance of artificial neural networks is clearly better than the other algorithms considered. The literature hints about the fact that the classification power of ANNs to be higher than the other algorithms considered in this paper. Apart from ANN, logistic regression and support vector machine are found to be almost equal in different values of performance measures. But the algorithm, naïve bayes could not perform well.

One can extend the study and apply the same algorithms on some real world data set. The result of the experiment will be interesting to see. One can also extend the study and apply different deep learning methods for the classification tasks.

References

[1] S. Muthukumaran, P. Suresh and J. Amudhavel, "Sentimental analysis on online product reviews using LS-SVM method," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 12, pp. 1342-1352, 2017.

[2] S. A. Devi, P. Sapkota and M. Obulesh, "Sentiment analysis on products using social media," Journal of Advanced Research in Dynamical and Control Systems, pp. 137-141, 2017.

[3] M. Bhargava and D. Rao , "Sentimental analysis on social media data using R programming," International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 80-84, 2018.

[4] C. G. Krishna , D. R. Meka, V. S. Vamsi and K. M. V. S. Ravi , "A survey on twitter sentimental analysis with machine learning techniques," International Journal of Engineering and Technology(UAE), vol. 7, no. 2.32, pp. 462-465, 2018.

[5] P. Jadhav and B. V. Babu, "Detection of Community within Social Networks with Diverse Features of Network Analysis," Journal of Advanced Research in Dynamical and Control Systems , vol. 11, no. 12, pp. 366-371, 2019.

[6] L. P. Maguluri, I. Bhavitha, S. . A. v. Reddy, T. N. Reddy and A. Chowdary , "An efficient method on supervised joint topic modeling approach by analyzing sentiments," Journal of Advanced Research in Dynamical and Control Systems, vol. 9, no. 18, pp. 3219-3230, 2017.

[7] B. R. Rahin, K. K. Prem, N. Danapaquiameq, J. Arumugam and D. Saravanan, , "Blocking Abusive and Analysis of Tweets in Twitter Social Network Using NLP in Real-Time," BIOSCIENCE BIOTECHNOLOGY RESEARCH COMMUNICATIONS, vol. 11, no. 1, pp. 94-103, 2018.

[8] C. Paul, D. Sahoo and P. Bora, "Aggression In Social Media: Detection Using Machine Learning Algorithms," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH , vol. 9, no. 4, pp. 114-117, 2020.

[9] L. A. Deshpande, and M. R. Narasingarao, "ADDRESSING SOCIAL POPULARITY IN TWITTER DATA USING DRIFT DETECTION TECHNIQUE," JOURNAL OF ENGINEERING SCIENCE AND TECHNOLOGY, vol. 14, no. 2, pp. 922-934, 2019.

[10] S. P. Bhargav, G. N. Reddy, R. R. Chand, K. Pujitha and A. Mathur, "Sentiment Analysis for Hotel Rating using Machine Learning Algorithms," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 6, pp. 1225-1228, 2019.

[11] H. Correspondent, "Facebook played a role in fuelling riots, says Delhi panel," 01 September 2020. [Online]. Available: https://www.hindustantimes.com/cities/facebook-complicit-in-aggravating-n-e-delhi-riots-says-delhi-assembly-panel/story-1HkXrGw4fWSOpoLUrVuCsO.html. [Accessed 03 September 2020].

[12] K. R. Balasubramanyam, "Bengaluru Riots: Karnataka to hold talks with social media giants on filtering fiery contents," 17 August 2020. [Online]. Available: https://economictimes.indiatimes.com/news/politics-and-nation/bengaluru-riots-karnataka-to-hold-talks-with-social-media-giants-on-filtering-fiery-contents/articleshow/77582323.cms. [Accessed 03 September 2020].

[13] K. Sreelakshmi , B. Premjith and K. P. Soman , "Detection of Hate Speech Text in Hindi-English Code-mixed Data," in Procedia Computer Science, Trivandrum, 2020.

[14] J. C. Pereira-Kohatsu, L. Quijano-Sánchez , F. Liberatore and M. Camacho-Collados , "Detecting and Monitoring Hate Speech in Twitter," Sensors (Basel), pp. 1-37, 2019.

[15] A. Gaydhani, V. Doma, S. Kendre and L. Bhagwat, "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach," rXiv preprint arXiv:1809.08651., pp. 1-5, 2018.

[16] G. B. Herwanto, A. M. Ningtyas, K. E. Nugraha and P. T. I Nyoman , "Hate Speech and Abusive Language Classification using fastText," in International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Jetis, Indonesia, 2019.

[17] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the Fifth International workshop on natural language processing for social media., Valencia, Spain, 2017.

[18] S. Ahammed, M. Rahman, H. M. Niloy and S. M. H. Chowdhury, "Implementation of Machine Learning to Detect Hate Speech in Bangla Language," in International Conference on System Modeling & Advancement in Research Trends, Moradabad, India, 2019.

[19] H. Sahi, Y. Kilic and R. B. Saglam, "Automated Detection of Hate Speech Towards Women on Twitter," in 2018 International Conference on Computer Science and Engineering (UBMK), Turkey, 2018.

[20] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, , Austin, 2016.

[21] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of NAACL-HLT 2016, San Diego, California, 2016.

[22] G. Koushik, K. Rajeswari and S. K. Muthusamy, "Automated Hate Speech Detection on Twitter," in 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019.

[23] T. Davidson, D. Warmsley, M. Macy and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in ICWSM, 2017.

[24] G. K. Pitsilis, H. Ramampiaro and H. Langseth , "Effective hate-speech detection in Twitter data using recurrent neural networks," Applied Intelligence, vol. 48, no. 12, p. 4730–4742, 2018.

[25] F. E. Ayo, ,. Folorunso and F. T. Ibharalu, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," Computer Science Review, pp. 1-34, 2020.

[26] F. . R. Nascimento, G. D. Cavalcanti and M. D. Costa-Abreu , "Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensemble learning," Expert Systems With Applications, vol. 201, no. September 2022, pp. 1-14, 2022.

[27] R. Ali, U. Farooq, U. Arshad, W. Shahzad and M. O. Beg, "Hate speech detection on Twitter using transfer learning," Computer Speech & Language, vol. 74, no. 2022, pp. 1-16, 2022.

[28] K. P. Roy, S. Bhawal and C. N. Subalalitha, "Hate speech and offensive language detection in Dravidian languages using deep ensemble framework," Computer Speech & Language, vol. 75, no. 2022, pp. 1-15, 2022.

[29] P. Kapil and A. Ekbal, "A deep neural network based multi-task learning approach to hate speech

detection," Knowledge-Based Systems, vol. 210, no. 2020, pp. 1-21, 2022.

[30] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," IEEE Access, vol. 6, pp. 13825 - 13835, 2018.

[31] R. Kumar, A. K. Ojha, S. Malmasi and M. Zampieri, "Benchmarking Aggression Identification in Social Media," in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, New Mexico, USA, 2018.

[32] R. Ali, U. Farooq, U. Arshad, W. Shahzad and M. O. Beg, "Hate speech detection on Twitter using transfer learning," Computer Speech & Language, vol. 74, no. July 2022, pp. 1-16, 2022.