A Supervised Classification Approach for Detecting Hate Speech in English Tweets

N. Solomon Praveen Kumar[a], Dr.M.S.Mythili[b]

[a]Research Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620017, Tamil Nadu, India.

[b]Assistant Professor, Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli – 620017, Tamil Nadu, India.

Abstract

As social concerns about threats of hatred and harassment have grown on the internet, there has been a lot of attention paid to detecting hate speech. This research looks at how well SGD classifiers with hyper-parameter tuning perform at detecting hate speech in tweets. It describes the categorization of English tweets with stochastic gradient descent (SGD) classifiers. The categorization of text documents depends on their content, which is divided into groups based on predefined categories. The Term-Frequency (TF) and Inverse-Document Frequency (IDF) parameters are implemented in the proposed system. A Stochastic Gradient Descent method (SGD) is used to generate classifiers that learn independent features, and performance is assessed using Accuracy and F1-score.

Keywords. Hate Speech, SGD, TF-IDF, English tweets, and hyper-parameter.

*Corresponding author Email: **solomon@bhc.edu.in**

1.Introduction

The context of social networks and methods enhances our understanding of social data, which are data collected through interpersonal sharing. A blog, a comment, or another form of communication can allow people to share their views on any subject (e.g. a place or a person). There are significant challenges associated with accessing social data, including literature-based data. Due to this massive amount of social data, marketing efforts can be streamlined and campaigns can be measured in an efficient decision support system. This can be used to streamline marketing campaigns [1].

In the 21st century, social networks have had a profound impact on society. Through the use of alphanumeric, special, and hyperlinked characters as well as photos, emoticons, and other icons, Twitter users can send information across the internet. Due to the development of its functionality, it has experienced a number of adjustments. As an illustration, Twitter has raised the character restriction for each tweet from 140 to 280, allowing for greater room for engagement [2].

Despite the fact that such behaviors are prohibited by Twitter's terms of service, freedom of expression has allowed incendiary and hate speech to propagate via social media sites in many places. Hate speech does not have an international legal definition, according to the United Nations policy and plan of action on the subject. However, there is one exception, and that is incitement, which is a clear and intentional act opposed to prejudice, animosity, and violence. Hateful speech has been defined as content that employs such terms, comparable to the way hate speech has occasionally been confused with verbally abusive or derogatory expressions. Through machine learning, inflammatory remarks on Twitter have been categorized and recognized used in accordance to racial, sexist, misogynistic, religious, refugee, and immigrant situations.

The information offered is an improved subset of the Kaggle dataset. Two labels make up the Kaggle dataset. Rhetoric, Hate Speech, Sexually Objectionable, Hypothetical, Sincere, and Other Questions are the five labels of questions that make up the hate speech dataset. This fine-grained taxonomy seeks to better identify these questions while also offering appropriate responses for each kind of topic. A supervised learning strategy was suggested.

There are two possible fixes: An SGD-optimized Classifier that makes use of Tf-Idf vectorization is used at the root level. The content is further refined by using the most pertinent terms from the categories of hate speech and sexually objectionable to filter it. The model's overall accuracy is about 84%.

Recent research [3, 4] has found that keywords alone are still insufficient to classify the use of hate speech. To identify alternate applications of the phrase that lack the requisite intention or meaning for the claim to be categorized as hate speech, a semantic analysis of the text is necessary.

Some examples of how the vocabulary used in hate speech is applied are as follows: Since homophones have numerous meanings and are used in this context with a specific connotation, they do not fit the definition of hate speech.

• This category encompasses extremely inflammatory language; hence it should be considered hate speech. The show of hostility towards that group sets this sort apart.
• When the keywords are utilized to highlight the term in question or a related topic, they must not be offensive or unsuitable.
• Appropriate when used by a member of that group to achieve a different goal.

2.Related Work

Three main steps were employed by Shoven Ahammed et.al., [2019] to identify hate speech in Bangla. due to the lack of the Bangla dataset. Data from Facebook has been gathered by the author. The dataset must first be formatted. The researchers used ml algorithms to incorporate hate speech from Facebook. The Naive Bayes technique generated the best F1 score (0.73), while the SVM algorithm had the highest F1 score (0.71). [5]

Sean MacAvaney et.alstudy[2019] explores the difficulties that online automatic methods encounter when attempting to identify hate speech in text. The author provides an overview of various examples of hate speech. In place of neural techniques, the authors suggest a Multi-View SVM since it is less complex and easier to understand. [6]

The idea of analysing large-sized text was put forth by Amita Jain et al. Particle-Swarm Optimization (PSO) and Neutrosophic Set are combined to create Senti-NSetPSO, a group of binary and ternary classifiers. Large documents larger than 25KB have been used to test Senti-NSetPSO. The accuracy of the proposed study's binary and ternary classifiers was 95.3% and 81.99%, respectively. The dataset was put together by the author using the Blitzer data set. [7]

A system for predicting hate speech from Twitter has been created by Zafer Al-Makha et al. The Killer Natural Language Processing Optimization Ensemble Deep Neural Network Learning Approach (KNLPEDNN), which has the highest prediction accuracy of 98.71%, is used to analyse the data using a fusion of machine learning and natural language processing techniques. [8]

Bidirectional Encoder Representation from Transformers (BERT) and Embedding's from Language Models are two well-known machine learning techniques for text categorization that Yanling Zhore et al. used to the SemEval-2019 data sets (ELMo). The author employs a deep

learning-based fusion technique to identify hate speech. The results demonstrate a significant increase in classification accuracy and F1-Score. [9]

Oluwafemi Oriola et al. developed a corpus of English tweets using South African vocabulary to test machine learning techniques. Combining Support Vector Machine ensemble and multi-tier meta-learning models of gradient boosting, random forest, and logistic regression, character n-grams, word n-grams, negative emotion, grammatical feature, and their hybrid, hyper-parameter optimization were retrieved and assessed. With true positives of 0.887 and 0.858, respectively, and an overall accuracy score of 0.671, Support Vector Machine, Random Forest, and Gradient Boosting meta-learning models were demonstrated to be consistently dependable and balanced in classifying offensive and hate speech. [10]

Sentiment analysis using an unbalanced class label distribution has been the topic of research by R. Srinivasan et al. additionally; "Code mixing" is a priority of the author. Text that alternates between two or more languages is referred to as code mixed data. The proposed analysis made use of the F1-Score to assess the effectiveness of several machine learning methods, such as the Random-Forest Classifier, Logistic-Regression, XG-Boost Classifier, Support Vector Machine, and Naive Bayes Classifier. [11]

Automated systems are capable of spotting offensive language mixed with multilingual characters, according to a study by Omar Sharif et al. To complete the tasks at hand, the author used two machine learning algorithms (LR and SVM), three pairs of transformers, and three deep learning techniques (LSTM, LSTM + Attention) (m-BERT, Indic-BERT, and XLM-R). The weighted F1-Score for the suggested techniques was 0.76. [12]

Gretel Liz De la Pena Sarracen has been focusing on negative tweets on Twitter. The study found that by factoring in elements like user activity, user communities, and potentially shared images along with tweets, it was possible to forecast hate speech more precisely when the textual content was taken into account. The author is working on a multimodal and multilingual approach to identifying hate speech in both English and Spanish tweets. The author asserts that frameworks with graphical representations were examined using deep learning techniques. Using graph neural network (GNN) techniques, the author investigated the user networks' structure and user community interactions. To process both the text of the tweets and the relationships between the people, we are specifically developing a method based on convolutional graph neural networks. [13]

Md Rabiul Awal et al. claim that they employed a supervised method that made heavy use of unbalanced and frequently deficient annotated datasets on hate speech. This study suggests AngryBERT, a new multitasking learning-based model designed to concurrently assess sentiment,

identify targets, and detect hate speech. The AngryBERT can recognize its target, reliably detect hate speech, and ascertain the emotion being conveyed. [14]

Ishan Sanjeev Upadhyay et al. have investigated two strategies. The original method developed classifiers based on Logistic Regression, Random Forest, SVM, and LSTM using contextual embedding. The second method requires fine-tuning pre-trained transformer models (BERT, AL-BERT, RoBERTa, and IndicBERT) with an output layer in order to produce an ensemble of 11 models that can cast a majority vote. With weighted F1-Scores of 0.93, 0.75, and 0.49 for English, Tamil, and Malayalam, respectively, the second technique outperformed the previous one. [15]

The Meme integrated Open-domain Dialogue (MOD) dataset, which Zhengcong Fei et al., have suggested, focuses on appropriate responses within multimodal historical contexts with copious emotion labelling. It offers an efficient and straightforward technique for creating Internet meme-based dialogues. The capacity to manage a MOD assignment, in the author's opinion, can be used as a useful test environment for assessing the development of multimodal, open-domain discourse intelligence. [16]

3.Model of Work

The purpose of this activity is to identify hate speech on Twitter. For efficiency, we label a tweet as having hate speech material if it displays racial or gender bias. So the challenge is separating racist or sexist tweets from other tweets. Formally, your job is to predict the labels on the test dataset from a training sample of tweets and labels, where label "1" designates either racism or sexism in a tweet and label "0" indicates neither. The conceptual layout of the study is depicted in the picture.

Logistic Regression and (linear) Support Vector Machines are two examples of linear classifiers and regression that can be effectively fitted using the statistical gradient descent (SGD) method. As a stochastic spin-off of gradient descent, stochastic gradient descent addresses the shortcomings of gradient descent and outperforms it significantly on big datasets. As a result, it is commonly used as a web-based machine-learning optimization technique. In online forums, comment sections, and on social media, hate speech is a problem. Since hate speech may have a negative impact on society, it is essential to recognize it. This blog post will explore the most modern techniques for detecting hate speech using machine learning algorithms.

A. COLLECTION AND ANNOTATION OF DATA

Twitter Archiver, a publicly available Google Sheets plugin that is based on the Twitter Search API, collected 81,121 English tweets between May 5, 2019 and May 13, 2019. Non-English tweets were removed.

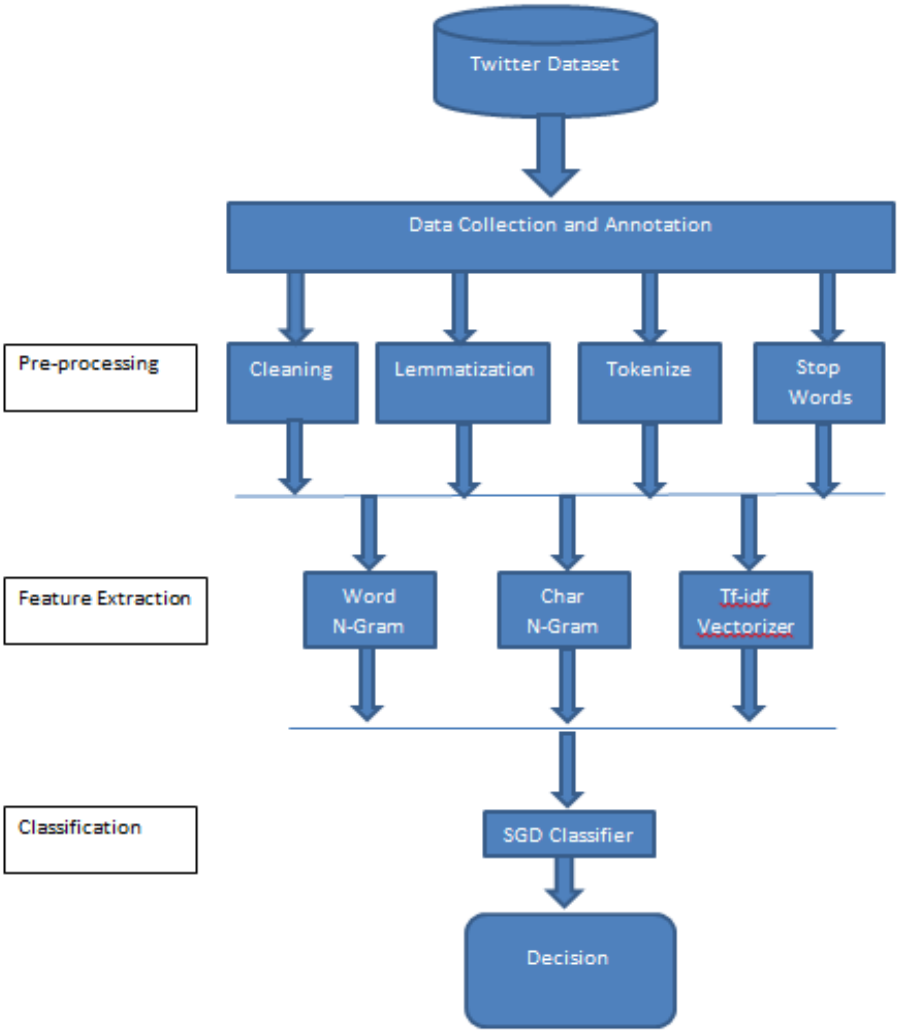


FIGURE 1. Proposed architecture for finding hateful speech in English tweets based on SGD Classifier.

Figure 1 shows a general method for locating hate speech in a sample of English tweets using the suggested model for predicting hateful speech in tweets using the SGD classifier. Results from stochastic gradient descent are more precise and work well with feature extraction for hyper-parameters.

The following rules were applied to the annotation:

When a tweet is intended toward that particular person or group of people, it is considered hate speech.

1) The wording in the tweet was usually racial or offensive.
   • It uses offensive words with the intention of hurting or inducing hurt.
   • It makes mention of and supports other racist facts, groups, and hateful tweets.
   • It makes use of insensitive or violent idioms, metaphors, colloquial terms, or other indirect communication techniques.
   • It uses violent language.
2) A tweet is deemed to be offensive speech if it:
   • Is directed at a specific individual, team, or organization.
   • The language used is not anti-Semitic.
   • By employing insulting or disparaging language, it verbally degrades the target.

3) A tweet qualifies as free speech if it:
   - Is directed at, or not directed at, a specific individual, group, or organisation.
   - The wording is neither hurtful nor inflammatory.
   - It is disseminated by a government body or officially recognised government agency with the intention of spreading awareness and encouraging widespread advocacy.

B. DATA PREPROCESSING AND TOKENIZATION

Tokenization and pre-processing were done after extracting the syntactic and sentiment indicators.

The steps that are involved in cleaning and summarizing tweets:

Because it accurately translates HTML tags, Links, retweets, report concerns, & Utf codes, NLTK employs Tweet Tokenize [17] for lemmatization.

• Stemming from the NLTK's Word-Net Lemmatize [17].
• Delete the username.
• Eliminating punctuation
• Elimination of emoji’s and other special characters and symbols.
• Elimination of hash-tag hash symbols.
  • Punctuations in English are deleted. Since stop words from other languages can be used as useful signs when they're offensive or while speaking English, they have been included. For instance, "bane" is a terrible word in English but a stop word in isiZulu and Sesotho.
• Lowercase all text wherever possible.

C. FEATURE EXTRACTION

In machine learning, the process of converting raw input data into examinable numerical values is referred to as feature extraction. In comparison to applying machine learning directly, it produces better results because it preserves the data from the original data set. By removing features, this procedure lowers the amount of duplicate data in the data collection. Reduced data makes it easier to create machine learning models, which improves generalization and learning as well as machine learning efficiency. For our investigation, we made use of all the features modified from [18].

D. TF-IDF FEATURES FOR WORD

A tweet's word count, which can range from 1 to N, is used to calculate the number of words that follow one another to form a word n-gram. The word qualities unigram (n = 1) and bigram (n = 2) were assessed in this study. The term frequency-inverse document frequency (TF-IDF) performance-improving approach [18] changes the word count in a document by the frequency of the word (term) in the corpus. The following is the TF-IDF for the word t in the document (tweet) d:

$$\text{IDF(t) TF(t, d)} = \text{IDF-TF(t, d)} \quad (1)$$

$$\text{When IDF (t)} = \log [n / (\text{DF(t)} + 1)]$$

There are n distinct documents in the document set.
Document frequency when it is DF (t).

E. Classification

The research includes 2 potential solutions. We have SGD-optimized TF-IDF vectorized classifiers at the root level. The answer is improved by picking the most important buzzwords from the categories of violent and hatful speech. SGD classifiers combine logistic and SVM classification with an accuracy and F1 score of 84%.

Performance Matrics

With the use of the true positive rate (TPR), accuracy (Acc), and macro-averaging for precision (P), recall (R), and F1-score, we looked at a classification problem and evaluated the efficacy of several solutions (F1). Through the use of macro-averaging, the average performance metrics from each of the k one-vs.-all matrices were independently determined. Since it displays the percentage of correctly predicted tweets per class, TPR, also known as sensitivity, was selected as the top algorithm for classification over accuracy. The formula for calculating the TPR is as follows: TPR = TP / (TP + FN). The number of instances of a class with an incorrect label "L" prediction and an accurate label "L" prediction are shown, respectively, by the symbols TN and TP.

4.EXPRIMENTAL RESULTS

The results exhibited here indicate the proportion of sentiments in the testing set that the SGD classifier correctly predicted. Accuracy and F1-score are employed as evaluation matrices. Higher values denote better categorization or prediction in both of these matrices, which have a range of 0 to 1. A total of 359 tweets are classified as negative tags in the test set of SGD out of 5456 total tweets.

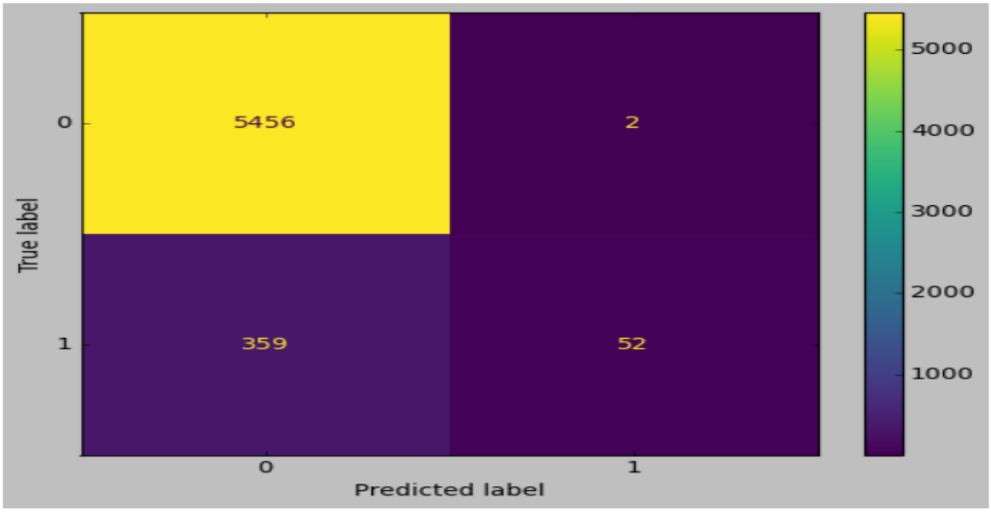


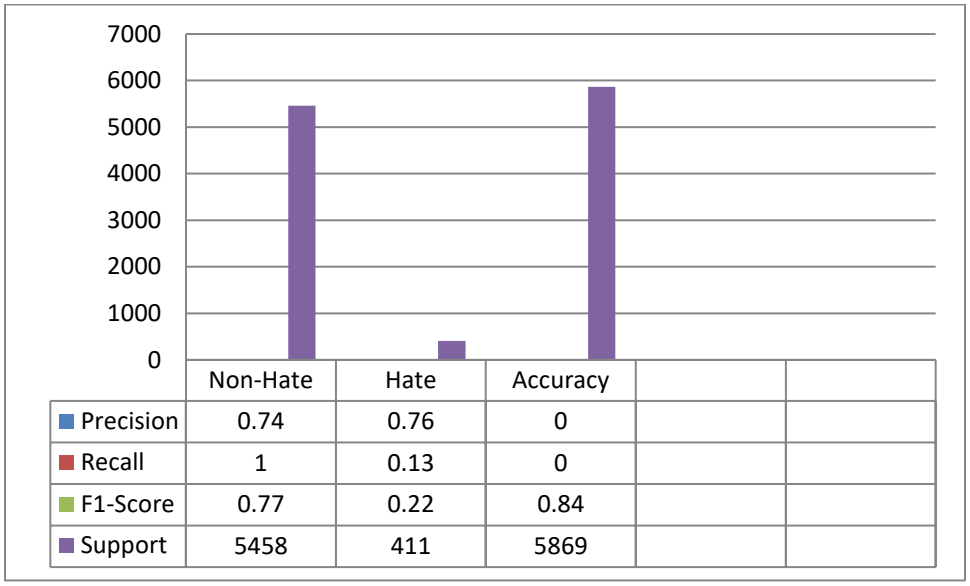
FIGURE 2. Confusion matrix for SGD classifier-based hate speech detection.

The confusion matrix for hate speech detection using an SGD classifier is displayed in Figure 2 along with the actual and predicted labels for the supported English hate speech tweets.

Index	Precision	Recall	F1-Score	Support
Non-Hate	0.74	1.00	0.77	5458
Hate	0.76	0.13	0.22	411
Accuracy	-	-	**0.84**	5869
Macro-Avg	0.85	0.56	0.60	5869
Weighted-Avg	0.84	0.84	0.82	5869

TABLE 1. The various results of SGD Classification.

SGD classification results are shown in Table 1, where accuracy and F1 scores are calculated with respect to hate and non-hate speech tweets, yielding 0.94 percent accuracy and 0.92 percent F1 score. The micro mean is 0.60 and 0.92, respectively.



	Non-Hate	Hate	Accuracy		
Precision	0.74	0.76	0		
Recall	1	0.13	0		
F1-Score	0.77	0.22	0.84		
Support	5458	411	5869		

FIGURE 3.1. Values for precision and recall based on SGD classifier.

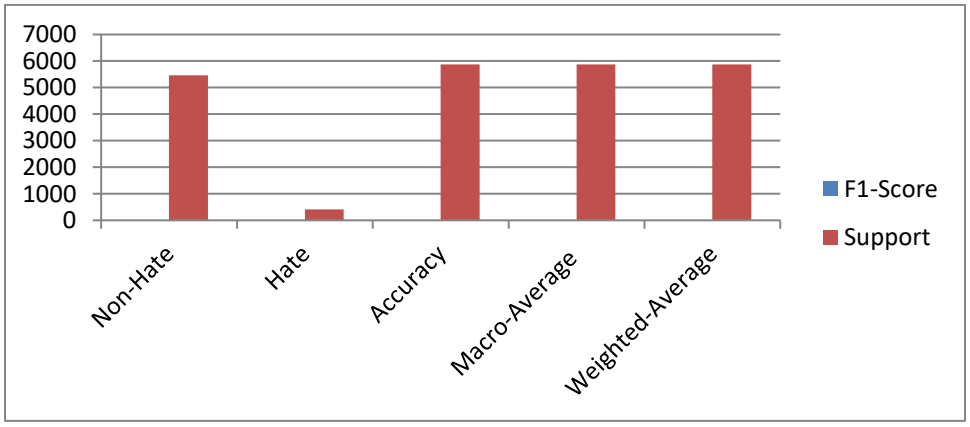


FIGURE 3.2 Values for F1-score and support based on weight and micro average.

Figure 3.1 and 3.2 The result indicates the best F1-score and accuracy by the SGD classification and hyper parameter tuning.

5.Conclusion

In order to find tweets that contained hate speech, this study used automated text classification tools. Because they serve as a benchmark for future text classification systems for automated hate speech detection, the findings of this study are practically significant. The suggested model has the drawback of being unable to effectively manage mixed datasets of hate speech. Therefore, the goal is to develop the suggested model so that it may also be used to anticipate how severe a text contains hate speech in the future.

References

[1]. Kaur, S., & Mohana, R. (2015). A roadmap of sentiment analysis and its research directions. International Journal of Knowledge and Learning, 10(3), 296-323. https://doi.org/10.1504/IJKL.2015.073485

[2]. Oriola, O., & Kotzé, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. IEEE Access, 8, 21496-21509. https://doi.org/10.1109/ACCESS.2020.2968173

[3]. Brown, A. What is Hate Speech? Part 1: The myth of hate. Law Philos. 2017, 36, 419-468. https://doi.org/10.1007/s10982-017-9297-1

[4]. Kurrek, J.; Saleem, H.M.; Ruths, D. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage.In Proceedings of the Fourth Workshop on Online Abuse and Harms, Online, 20 November (2020); Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 138-149. https://doi.org/10.18653/v1/2020.alw-1.17

[5]. Awal, M. R., Cao, R., Lee, R. K. W., & Mitrovic, S. (2021). AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection. arXiv preprint arXiv:2103.11800. https://doi.org/10.1007/978-3-030-75762-5_55

[6]. Al-Makhadmeh, Z., & Tolba, A. (2020). Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing, 102(2), 501-522. https://doi.org/10.1007/s00607-019-00745-0

[7]. Zhou, Y., Yang, Y., Liu, H., Liu, X., & Savage, N. (2020). Deep learning based fusion approach for hate speech detection. IEEE Access, 8, 128923-128929. https://doi.org/10.1109/ACCESS.2020.3009244

[8]. Oriola, O., & Kotzé, E. (2020). Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. IEEE Access, 8, 21496-21509. https://doi.org/10.1109/ACCESS.2020.2968173

[9]. MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

[10]. Upadhyay, I. S., Wadhawan, A., & Mamidi, R. (2021). Hopeful_Men@ LT-EDI-EACL2021: Hope Speech Detection Using Indic Transliteration and ssTransformers. arXiv preprint arXiv:2102.12082.

[11]. Ahammed, S., Rahman, M., Niloy, M. H., & Chowdhury, S. M. H. (2019, November). Implementation of machine learning to detect hate speech in Bangla language. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 317-320). IEEE. https://doi.org/10.1109/SMART46866.2019.9117214

[12]. De La Peña Sarracén, G. L. (2021, March). Multilingual and Multimodal Hate Speech Analysis in Twitter. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 1109-1110). https://doi.org/10.1145/3437963.3441668

[13]. Sharif, O., Hossain, E., & Hoque, M. M. (2021). NLP-CUET@ DravidianLangTech-EACL(2021): Offensive Language Detection from Multilingual Code-Mixed Text using Transformers. ArXiv preprint arXiv:2103.00455.

[14]. Srinivasan, R., & Subalalitha, C. N. (2021). Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distributed and Parallel Databases, 1-16. https://doi.org/10.1007/s10619-021-07331-4

[15]. Jain, A., Nandi, B. P., Gupta, C., & Tayal,D. K. (2020). Senti-NSetPSO: large-sized document-level sentiment analysis using Neutrosophic Set and particle swarm optimization. Soft Computing, 24(1), 3-15. https://doi.org/10.1007/s00500-019-04209-7

[16]. E. L. S. Bird, Analyzing Texts With Natural Language Toolkit: Natural Language Processing With Python, 1st ed. Newton, MA, USA: O'Reilly Media, (2009).

[17]. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar,"SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval)," in Proc. 13th Int. Workshop Semantic Eval. (SemEval),(2019), pp. 75-86. https://doi.org/10.18653/v1/S19-2010