International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 273-278.
www.theijire.com ***ISSN No: 2582-8746***

Artificial Intelligence Techniques for Hate Speech Detection

Mohini Chakarverti
Assistant Professor, Bennett University, Uttar Pradesh, India.

How to cite this paper:
Mohini Chakarverti, "Artificial Intelligence Techniques for Hate Speech Detection", IJIRE-V4I02-273-278.

Abstract: *Social media and other large online communication venues allow users to express themselves freely and occasionally anonymously. While having the right to express oneself freely is a value-laden human right, inciting and disseminating hatred against another group is a misuse of this freedom. The HS detection pipeline starts with the dataset gathering and preparation stage. Social media sites like Face book, YouTube, Twitter, and others are often used to collect data. Pre-processing is done in accordance with the quality and structure of the dataset. The artificial intelligence plays important role for the hate speech detection. The artificial intelligence based techniques is designed in the last years for the hate speech detection. In this paper various techniques for hate speech detection is reviewed and analysed in terms of various parameters.*

Key Word: *Artificial Intelligence, Hate Speech Detection, Data Labelling.*

I.INTRODUCTION

Unfortunately, hate crimes are nothing new in our culture. Social media and other online communication tools have started to play a bigger part in hate crimes, though. For instance, suspects in a number of recent terror incidents that were motivated by hatred had a long history of posts on social media that supported their views, which suggests that social media plays a role in their radicalization. Social media and other large online communication venues allow users to express themselves freely and occasionally anonymously[1]. While having the right to express oneself freely is a value-laden human right, inciting and disseminating hatred against another group is a misuse of this freedom. As a result, a lot of online communities, including Face book, YouTube, and Twitter, have regulations to get rid of hate speech. There is a tremendous incentive to research automatic detection of hate speech given the public concern and how pervasive it is becoming online. The dissemination of hateful content can be stopped by automating its identification. Due to the voluminous amount of information that is shared online, there is a tremendous incentive to research the automatic detection of hate speech. To reduce crime and safeguard people's views, hate speech must be identified. This study is particularly significant in light of continuous conflicts that distort truth and dehumanise the victimised Ukrainian people [2].

1.1 Automatic Hate Speech Detection Based on AI

Natural language processing (NLP) technology advancements in recent years have made it possible to complete a number of studies on the automatic detection of hate speech in text. In the modern era, artificial intelligence has permeated several industries. Artificial intelligence has applications in a wide range of fields, including research, education, finance, and business. One use of AI is in machine learning. Computers can learn the connection between input and output without being explicitly programmed thanks to machine learning. As a result, in machine learning, as opposed to traditional programming, where writing algorithms is required, it is necessary to uncover the algorithm that extracts patterns [3] from a given dataset and creates a predictive model from which the computer can learn the patterns between input and output.The machine learning algorithm can now make predictions based on fresh, unexplored data. Under the umbrella of AI, deep learning is also a form of machine learning.

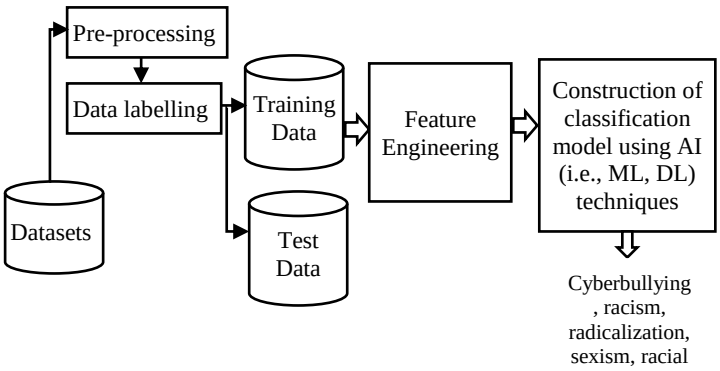

Figure 1: Typical automatic HS detection system pipeline

The whole procedure for the text classification system-based HS detection task is shown in Figure 1. The HS detection pipeline starts with the dataset gathering and preparation stage. Social media sites like Face book, YouTube, Twitter, and

others are often used to collect data. Pre-processing is done in accordance with the quality and structure of the dataset [4]. Generally, this entails filtering and normalising textual inputs, which may include, among other things, tokenization, noise reduction, stop word removal, misspelling correction, and lemmatization. Another possibility is that the dataset will be delivered to clients right away, eliminating the need for collecting. The training and testing portions of the dataset should be separated in order to prepare it for the next machine learning stage. The required traits are later extracted from the textual inputs in the analysis step after feature engineering, turning the unstructured text sequences into structured features. The TF-IDF, semantic, lexical, topic modelling, sentiment, BOW, and word embedding (FastText, GloVe, and Word2Vec) are examples of common feature extraction methods. Dimension reduction is sometimes used to simplify time and memory.

A few examples of dimension reduction methods include principal component analysis (PCA), linear discriminant analysis (LDA), non-negative matrix factorization (NMF), random projection, autoencoders, and t-distributed stochastic neighbour embedding (t-SNE) [5]. One of the most crucial processes in the pipeline for text categorization is the training of a machine learning or deep learning model on the training dataset. A variety of classifiers, including as RF, NB, LR, CNN, RNN, BERT, etc., can be changed depending on the requirements of the task. In a neural network model, word embedding is frequently combined with another embedding layer to enhance deep learning performance. The machine learning/deep learning model can distinguish between several varieties of hate speech and non-hate speech, or it can produce a multi-class output (for instance, hate speech vs non-hate speech) [6]. This final stage of the text categorization pipeline estimates the performance of the machine learning/deep learning model. Some of the evaluation metrics used for this include accuracy, F1 score, precision, Matthews Correlation Coefficient (MCC), receiver operating characteristics (ROC), and area under the ROC curve (AUC).

1.2AI Models for Hate Speech Detection

Below is a brief summary of some of the most popular AI-based methods for detecting hate speech:

i. Support Vector Machine: With the Support Vector Machine (SVM) machine learning technology, data is analysed and patterns are found using a mapping model. Regression analysis and classification are two of its primary uses. In order to determine which of two categories to assign incoming data to, an SVM algorithm builds a non-probabilistic binary linear classification model from a collection of data that falls into one of them [7]. In the space of the mapped data, it represents the categorization model as a border. An SVM algorithm creates the border that has the widest range. It is advantageous for both linear and nonlinear classification. In order to execute non-linear classification, it is crucial to map the provided data onto a high-dimensional feature space.

Bayesian Network: A Bayesian network graphically depicts the probability relationships between significant variables. A Bayesian model has various advantages over other data analysis techniques when applied with statistics. First, because it illustrates how all variables are interdependent, the model may quickly resolve a scenario where there are gaps in data entry. Second, a Bayesian network can be used to find causal connections, comprehend a problem area better, and predict the results of intervention [8]. Third, a Bayesian model may relate facts to prior knowledge because it includes semantics that are both causal and probabilistic (since the latter frequently takes shape as causal). Using Bayesian networks along with statistical methods is the ultimate solution to prevent data over fitting.

i. Convolutional Neural Networks: In the deep learning field, CNNs are among the most popular and successful architectures, particularly for computer vision problems. Convolutional layers, which convolve a kernel (or filter) of weights to extract features; nonlinear layers, which apply an activation function to feature maps (typically element wise) to allow the network to model non-linear functions; and pooling layers, which provide substitute of a small neighbourhood of a feature map with statistical data (mean, max, etc.) [9] about the neighborhood and reduce spatial resolution. Each unit in a layer receives weighted inputs from a small area of units in the layer below it, known as the receptive field, because layers are locally coupled. Higher-level layers are able to learn features from progressively wider receptive fields by stacking layers to create multi-resolution pyramids. The key computational benefit of CNNs is that they have a lot fewer parameters than fully-connected neural networks because all the receptive fields in a layer share weights.

ii. RNN architecture: RNN is an evolving innovative deep architecture for sequence data. RNNs are potentially useful to get long-run dependence for sequence data. The RNN architecture is generally implemented using Long Short Term-Memory (LSTM) for CNER. LSTM is the best-known application of the RNN framework. A basic LSTM unit consists of three multiplier gates, with an input gate to regulate the share of input information delivered to a memory cell; a forget gate to regulate the proportion of historic information to the past position; and an output gate to regulate the proportion of output information to move on to the further stage. Many regular deep learning techniques such as character embedding and dropout have been also implemented. For the input layer, word embeddings and character embedding are integrated into one input vector. The final classification layer of the RNN uses the CRF loss function [10].

Bidirectional LSTM: Standard RNNs only process information coming from one direction and focus on processing future data. Applying the bidirectional topology of LSTM solves this issue. By taking into account both the past and the future, the bidirectional LSTM [11] captures all available time information at time t. The standard RNN's hidden neurons are split using this technique into a forward state and a backward state. Both the forward and backward states of the brain's neurons are not interconnected with one another. the two-way LSTM expansion's three-time steps' fundamental organization. This structure is identical to a typical one-way RNN without a reverse state. This topology eliminates the requirement for additional delays present in conventional RNNs.

GRU: GRU (gated recurrent unit), which is comparable to LSTM, can be regarded as a variation of LSTM. In normal RNNs, the vanishing gradient problem is mostly addressed by GRU, which enhances the learning of long-term relationships in the network [12]. The tanh and sigmoid functions are also used by the GRU block to compute the required values. The GRU block does not, however, have a separate storage unit, in contrast to the LSTM block. The input/update gate is in charge of regulating the information flow because there isn't a separate forget gate for this kind of block. It has lesser number of parameters and a simpler architecture because of the distinction between these two structures, which eventually makes it more computationally efficient and simpler to train. The GRU block also contains a reset gate in addition to the update gate. Four values—update gate, reset gate, candidate activation, and output activation—are computed in a GRU block. The current block input and the prior activation value are utilized as inputs to calculate the weight and bias for each gate and candidate activation [13]. The gate value is determined in the first stage using the sigmoid function.

II.LITERATURE REVIEW

M. Z. Ali, et.al (2021) suggested a method and focused on preparing a comprehensive data set in which Urdu Tweets were comprised to detect the hate speech on the basis of analyzing the sentiments [14]. Dynamic SWF (stop words filtering) was adopted to handle the issue of sparsity; VGFSS (Variable Global Feature Selection Scheme) for tackling dimensionality; and SMOTE (Synthetic Minority Optimization Technique) model to deal with class imbalance, so that efficiency of the suggested method was enhanced. The experiments were conducted using 2 ML (machine learning) methods namely SVM (Support Vector Machine) and MNB (Multinomial Naïve Bayes). The outcomes indicated that when the class skew was tackled and the high dimensionality issue was tackled, the performance to detect the hate speech was enhanced.

A. Alhothali, et.al (2023) recommended DSWE (domain-specific word embedding) with a BiLSTM (Bidirectional Long Short-Term Memory)-based algorithm as a classification algorithm for detecting hate speech in automatic way [15]. This algorithm ensured that the word was able to assign its negative meaning itself and it was effective for detecting the coded words. Additionally, TL (transfer learning) method called BERT (Bidirectional Encoder Representation from Transformers) was adopted for detecting the hate speech problem as a binary classification task due to its capacity of generating superior results to accomplish NLP (Natural Language Processing) tasks. The experimental results exhibited that the recommended approach offered f1-score of 93%, and BERT attained f1-score on 96% on a combined balanced dataset while detecting hate speech.

S. Alsafari, et.al (2020) constructed an ensemble of CNN (Convolution Neural Network) and BiLSTM (Bidirectional Long Short-Term Memory) models relied on 2-class, 3-class, and 6-class Arabic-Twitter datasets for training with non-contextual and contextual (called Multilingual Bert and AraBert) word-embedding algorithms [16]. These models were capable of classifying the hate and offensive speech. The constructed approach was evaluated in a series of experiments on testing datasets. The results depicted the supremacy of constructed approach and attained F-scores of 91%for two-class, 84%for two-class, three-class, and 80% for six-class prediction tasks.

M. Alowaidi, et.al (2023) introduced MPCA+ECNN algorithm in which MPCA (Modified Principal Component Analysis) was combined with ECNN (Enhanced Convolutional Neural Network) for detecting the hate speech in the text [17]. NLP (Natural Language Processing) was utilized to assess syntax and meaning. This approach was useful to pre-process the data, extract the features and classify the data. The extra spaces, punctuation, and stop words were eliminated to clean the text. The processed attributes were employed in MPCA so that the features were extracted. Thereafter, the introduced algorithm helped to detect the examples of hate speech. The latter algorithm performed more effectively to recognize the hateful content online on enormous datasets. The results demonstrated that the introduced algorithm was useful for maximizing F-measure values, as well as accuracy, precision, and recall.

Y. Kim, et.al (2022) presented CL (contrastive learning) approach in order to fine-tune the implicit hate speech detector for maximizing the generalization capability [18]. In general, the shared implication was implemented as a positive sample for its corresponding hateful posts, and an ImpCon (implication-based contrastive learning)technique was established to detect the hate speech. Three datasets were executed to simulate the established technique while detecting the hate speech. The experimental outcomes validated that the established technique led to maximize the efficiency up to 9.10% on BERT, and 8.71% on HateBERT.

G. Koushik, et.al (2019) developed a framework with the purpose of detecting the hate content on Twitter in automatic way [19]. This framework was designed on the basis of BoW (bag of words) and TFIDF (term frequency-inverse document frequency) attributes. The ML (machine learning) algorithms were trained using these features. Moreover, LR (logistic regression) algorithm was presented and implemented to classify the tweets into hate and normal. The twitter dataset executed to quantify the developed framework. The experimental results reported that the developed framework attained an accuracy of 0.9411 with BoW attributes and 0.9462 using TFIDF feature to detect tweet as hateful or normal.

T. Turki, et.al (2022) projected an ML (machine-learning) model for detecting the hate speech from Twitter data [20]. Initially, a count vectorizer method adopted for generating the feature vectors, and integrated them with its corresponding labels. Bagging, AdaBoost, and RF (Random Forest) utilized these vectors as input. Subsequently, the hateful tweets were examined through a word-cloud visualization and the considerable textual data was represented. This model aimed to report the statistical information for discovering the tweets of highest frequency so that the data was comprehended more efficiently. The results of experimentation revealed the superiority of RF over other methods as it generated effective outcomes in terms of accuracy, F1 score, and AUC (area under curve). Moreover, the projected model with RF became a promising tool to detect the hate speech within Twitter.

H. M. S. T. Sandaruwan, et.al (2019) designed lexicon based and ML (machine learning) based technique for detecting Sinhala hate and offensive speeches inautomatic way after their sharing on Social Media [21]. The first task was to

initiate thelexicon based technique with the lexicon generating procedure and the accuracy of corpus based lexicon was found 76.3% to detect the speech as hate, offensive and neutral. Thereafter, ML method was implemented for which a corpus of 3000 comments was created and these comments were related to the hate, offensive and neutral speeches. This corpus was assisted in recognizing the fitting feature groups and models to detect the hate speech. The experimental results confirmed that the character trigram with MNB (Multinomial Naïve Bayes) yielded a recall of 84% and accuracy of 92.33%.

A. Chopra, et.al (2022) suggested a mechanism in which a Tablet classification method was implemented and its training was done on features extracted via MuRIL from transliterated code-mixed textual data [22]. The accuracy based grid search on hyper parameters such as embedding was considered to optimize the structure of the suggested mechanism. Furthermore, several algorithms were compared to the task of detecting the code mixed hate speech. The suggested mechanism performed well as compared to the traditional methods and provided 90% accuracy in code mixed with a test. Moreover, it was able to maintain superior accuracy of 90% on other 2 datasets.

P. K. Roy, et.al (2020) emphasized on tackling the problems of hate speech on Twitter [23]. It was complex to filter any information in manual way from a large incoming traffic. To overcome such issue, an automated system called the DCNN (Deep Convolutional Neural Network) was devised. This algorithm made the implementation of tweet text with GloVe embedding vector for capturing the semantics of tweets. For this, the convolution operation was executed. According to the findings, the devised algorithm offered a precision of 97%, recall of 88% and F1-score of 92% in contrast to the state-of-art methods.

F.Y. Al Anezi, et.al (2022) investigated DRNNs (deep recurrent neural networks) in order to classify and detect the hate speech [24]. This approach was known asDRNN-2 in which ten layers were included with thirty-two batch sizes and fifty iterations to classify the hate classification task. Moreover, a model recognized as DRNN-1 was also adopted in which 5 hidden layers comprised to carry out the binary classification. The implemented algorithms offered accuracy of 99.73% to perform binary classification, 95.38% for 3 classes of Arabic comments, and 84.14% for 7 classes. The investigated algorithm attained superior accuracy to classify the complex language, such as Arabic as compared to the other methods.

2.1. Comparison Table

Author	Year	Technique Used	Results	Limitations
M. Z. Ali, et.al	2021	SWF (stop words filtering), VGFSS (Variable Global Feature Selection Scheme) and SMOTE	The outcomes indicated that when the class skew was tackled and the high dimensionality issue was tackled, the performance to detect the hate speech was enhanced.	This approach was ineffective of tackling class skew issue without using lexical scores of the terms in the features set. Moreover, the issue related to class imbalance was occurred.
A. Alhothali, et.al	2023	DSWE (domain-specific word embedding) as features and a BiLSTM (bidirectional Long Short-Term Memory)-based algorithm	The experimental results exhibited that the recommended approach offered f1-score of 93%, and BERT attained f1-score on 96% on a combined balanced dataset while detecting hate speech.	This approach was not detected multi-class hate speech.
S. Alsafari, et.al	2020	Ensemble of CNN (Convolution Neural Network) and BiLSTM (Bidirectional Long Short-Term Memory)	The results depicted the supremacy of constructed approach and attained F-scores of 91%for two-class, 84%for two-class, three-class, and 80% for six-class prediction tasks.	The data was classified in this work using huge amount of labelled data for learning and it was its major limitation.
M. Alowaidi, et.al	2023	MPCA+ECNN algorithm	The results demonstrated that the introduced algorithm was useful for maximizing F-measure values, as well as accuracy, precision, and recall.	The datasets of huge level were not handled in this work.
Y. Kim, et.al	2022	ImpCon (implication-based contrastive learning)technique	The experimental outcomes on cross-dataset validated that the	The generalized potential of this technique was

			established technique led to maximize the efficiency up to 9.10% on BERT, and 8.71% on HateBERT.	restricted due to hate speech on unseen target.
G. Koushik, et.al	2019	BoW (bag of words) and TFIDF (term frequency-inverse document frequency) method	The experimental results reported that the developed framework attained an accuracy of 0.9411 with BoW attributes and 0.9462 using TFIDF feature to detect tweet as hateful or normal.	The accuracy of this framework was mitigated in case of deployment of linguistic attributes.
T. Turki, et.al	2022	ML (machine-learning) model	The results of experimentation revealed the superiority of RF over other methods as it generated effective outcomes in terms of accuracy, F1 score, and AUC.	This technique was not applicable to deal with all the tasks of NLP (natural language processing) in healthcare domains.
H. M. S. T. Sandaruwan, et.al	2019	lexicon based and ML (machine learning) based techniques	The experimental results confirmed that the character trigram with MNB (Multinomial Naïve Bayes) yielded a recall of 84% and accuracy of 92.33%.	This work lackedenough comments for considering Singlish (Sinhala words written in English) hate speeches. Thus, it was complex task to detect the hate speech.
A. Chopra, et.al	2022	Tablet classification method	The suggested mechanism performed well as compared to the traditional methods and provided 90% accuracy in code mixed with a test. Moreover, it was able to maintain superior accuracy of 90% on other 2 datasets.	The suggested mechanism had inaccuracies in the section of analyzing errors on a large dataset in which infrequent and common hate words were comprised.
P. K. Roy, et.al	2020	DCNN (Deep Convolutional Neural Network)	According to the findings, the devised algorithm offered a precision of 97%, recall of 88% and F1-score of 92% in contrast to the state-of-art methods.	This algorithm was effective to detect the hat speech using only textual data. Moreover, it was not applicable to detect hat speech from the other source data.
F.Y. Al Anezi, et.al	2022	DRNN-2	The implemented algorithms offered accuracy of 99.73% to perform binary classification, 95.38% for 3 classes of Arabic comments, and 84.14% for 7 classes. The investigated algorithm attained superior accuracy to classify the complex language, such as Arabic as compared to the other methods.	This algorithm was not able to test and process the data in real time on social media platforms.

References

[1] N. Shawkat, J. Simpson and J. Saquer, "Evaluation of Different ML and Text Processing Techniques for Hate Speech Detection," 2022 4th International Conference on Data Intelligence and Security (ICDIS), Shenzhen, China, 2022, pp. 213-219

[2] N. S. Mullah and W. M. N. W. Zainon, "Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review," in IEEE Access, vol. 9, pp. 88364-88376, 2021

[3] B. R. Amrutha and K. R. Bindu, "Detecting Hate Speech in Tweets Using Different Deep Neural Network Architectures," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 923-926

[4] G. H. Panchala, V. V. S Sasank, D. R. HarshithaAdidela, P. Yellamma, K. Ashesh and C. Prasad, "Hate Speech & Offensive Language Detection Using ML &NLP," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 1262-1268,

[5] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 204951-204962, 2020

[6] Rahul, V. Gupta, V. Sehra and Y. R. Vardhan, "Ensemble Based Hinglish Hate Speech Detection," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1800-1806

[7] M. U. S. Khan, A. Abbas, A. Rehman and R. Nawaz, "HateClassify: A Service Framework for Hate Speech Identification on Social Media," in IEEE Internet Computing, vol. 25, no. 1, pp. 40-49, 1 Jan.-Feb. 2021

[8] Rahul, V. Gupta, V. Sehra and Y. R. Vardhan, "Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1112-1118

[9] U. A. N. Rohmawati, S. W. Sihwi and D. E. Cahyani, "SEMAR: An Interface for Indonesian Hate Speech Detection Using Machine Learning," 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2018, pp. 646-651

[10] O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," in IEEE Access, vol. 8, pp. 21496-21509, 2020

[11] S. A. Kokatnoor and B. Krishnan, "Twitter Hate Speech Detection using Stacked Weighted Ensemble (SWE) Model," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Bangalore, India, 2020, pp. 87-92

[12] H. Rathpisey and T. B. Adji, "Handling Imbalance Issue in Hate Speech Classification using Sampling-based Methods," 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 2019, pp. 193-198

[13] K. Mnassri, P. Rajapaksha, R. Farahbakhsh and N. Crespi, "BERT-based Ensemble Approaches for Hate Speech Detection," GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 4649-4654

[14] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed and S. Hussain, "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis," in IEEE Access, vol. 9, no. 6, pp. 84296-84305, 2021

[15] A. Alhothali and K. Moria, "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model", Applied Artificial Intelligence, vol. 37, no. 1, pp. 1-6, 2023

[16] S. Alsafari, S. Sadaoui and M. Mouhoub, "Deep Learning Ensembles for Hate Speech Detection," 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA, 2020, pp. 526-531

[17] M. Alowaidi, "Hate Speech Detection Using Modified Principal Component Analysis and Enhanced Convolution Neural Network on Twitter Dataset", vol. 23, no. 1, pp. 363-374, January 2023

[18] Y. Kim, S. Park and Y.-S. Han, "Generalizable Implicit Hate Speech Detection using Contrastive Learning", 29th International Conference on Computational Linguistics, October 12–17, 2022, pp. 6667–6679

[19] G. Koushik, K. Rajeswari and S. K. Muthusamy, "Automated Hate Speech Detection on Twitter," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-4

[20] T. Turki and S. S. Roy, "Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer", Applied Sciences, vol. 10, pp. 145-151, 2022

[21] H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, pp. 1-8

[22] A. Chopra, D. K. Sharma, A. Jha And U. Ghosh, "A Framework for Online Hate Speech Detection on Code Mixed Hindi-English Text and Hindi Text in Devanagari",ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 9, no. 2, pp. 53-62, 20 October 2022

[23] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 204951-204962, 2020

[24] F.Y. Al Anezi, "Arabic Hate Speech Detection Using Deep Recurrent Neural Networks", Applied Sciences, vol. 34, no. 7, pp. 4335-4344, 2022