International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V74I2P116

Volume 74 Issue 2, 223-234, February 2026

Original Article

Censored Regressive Canonical Optimized Convolutional Deep Belief Classifier For Hate Speech Detection in Online Social Network

I. Imthiyas Banu[1], Velumani Thiyagarajan[2], Vijay Arputharaj J[3], P. Thenmozhi[4]

[1,2]*Department of Computer Science, Rathinam College of Arts and Science, Coimbatore, India.*
[3]*Department of Computer Science, Christ University, Bengaluru, India.*
[4] *Department Of Computer Science, Nandha Arts And Science College (Autonomous), Erode, India.*

[2]*Corresponding Author : Velumani46@gmail.com*

Received: 01 August 2025 Revised: 14 January 2026 Accepted: 20 January 2026 Published: 14 February 2026

Abstract - *Social networking uses internet-based platforms to facilitate users to make connections with others and share various forms of content, including text, images, videos, and links. Social networking services are mainly used for non-social interpersonal communication. Many approaches have been developed for hate speech detection, but they still face significant challenges, particularly in classifying text into multiple labels accurately and in a timely manner. For accurate hate speech detection in social networks, a Censored Regressive Canonical Optimized Convolutional Deep Belief Classifier (CRCOCDBC) model is developed. The objective of the developed CRCOCDBC is to detect multi-class hate speech with minimal time and error rate. Comparative analysis shows improved performance in terms of minimum error and higher authentication accuracy and precision than other well-known methods.*

Keywords - *Hate Speech Detection, Deep Belief Networks and Convolutional Neural Networks, Canonical Correlation, Krill Herd Algorithm.*

1. Introduction

Online social media is essential for every individual's life in the community for providing enhanced communication. While positive communication within diverse communities significantly enhances confidence, negative comments harm people's reputations and well-being. Therefore, detecting the rapid spread of hate speech is a crucial task in creating safer and more inclusive digital spaces. In the beginning, fundamental communication selects a particular range that imparts an individual with the privilege to convey their perspectives and notions.

The platform for sharing user information within a network is online social media. These platforms facilitate communication and interaction among friends, family, colleagues, and even businesses with their customers. In spite of virtual communication through social media programs, hate speech detection is tremendously advantageous and has grown into an unavoidable component.

The hate detection of social media is noticed in the form of offensive content referred to as hate speech, and fake news that has a crucial uneasiness in society. Such objectionable content can influence an individual's mental health, which cannot always be recovered. So, ascertaining and arbitrating such content is a dominant need of the moment. Conv-BiRNN-BiLSTM framework [1] was introduced to detect hate speech on online social media. However, it does not utilize an efficient meta-heuristic algorithm for hyperparameter optimization, enhancing accuracy. The FAST-RNN technique [2] was developed with a DNN using regularization methods. The time required for hate speech detection was not minimized. Two transformer-based models were designed in [3].

The Social Hater BERT approach is introduced in [4]. In [5], hate speech detection on social networks is provided by a curated dataset.DL based method was designed in [6]. In [7], a two-channel DL model was designed. Detection of hate speech using contextual information is described in [8]. Ensemble DL Model was intended in [9].

In [10], G-BERT was presented for identifying hate speech in Bengali social media. The goal of hate speech detection research is to organize healthy, fair, and scalable systems in real-time content moderation. The existing method that fails to understand evolving language or is biased against specific demographics can cause more harm, leading to unfair

censorship or to the proliferation of harmful content. To overcome this issue, the research method explores hybrid models that integrate DBNs' strengths in unsupervised feature learning with the superior contextual understanding of state-of-the-art models like Transformers, while focusing on improving generalizability and interpretability.

1.1. Contributions of Proposed Work

- To improve multi-class hate speech detection accuracy with minimum time in an online social network, the CRCOCDBC model is designed.
- To identify stop words from the input texts, preprocessing is performed by applying the Gestalt pattern recognition method. The Gestalt pattern recognition method is used for accurate stop word removal by matching the length of the word before and after processing. This process minimizes the time complexity of hate speech detection.
- To select the relevant features (i.e., keywords) for accurate classification, a censored regression function is applied to the preprocessed work.
- To perform in the Max pooling layer using correlation estimation, the Canonical correlation is measured between extracted keywords to classify multiple classes of hate speech. Correlation results are used to detect hate speech with higher accuracy

1.2. Paper Organization

The rest of the paper is organized into different sections as follows: Section 2 provides a brief elaboration on the related works. Section 3 describes the different processes of the CRCOCDBC model, with a neat diagram. Section 4 explains the simulation settings and dataset description, followed by the implementation procedure. Sections 5 present the performance evaluation and discussion. Finally, Section 6 describes the discussion of the proposed model, and the conclusion is provided in Section 7.

2. Related Works

Online Multilingual Hate Speech detection was described in [11]. In [12], the DL predictor is named as Passion-Net. The feature combination model is introduced in [13]. In [14], a transfer learning approach was designed. An interpretable two-dimensional visualization tool was used to develop transfer learning [15].

In [16], Neutrosophic NN are developed with Whale Optimization Algorithm and PSO. Yet another novel conception of unsupervised progressive domain adaptation on DL using multiple text datasets was presented in [17]. Hate speech classification using DNN was proposed in [18]. The current plethora of status and recommendations was designed in [19]. However, another hate speech classification method employing SVM and Naïve Bayes was proposed in [20].NLP employing the DL technique was proposed in [21]. An adaptive ensemble classifier was proposed in [22]. An FS model employing the Ruzicka similarity function and applying a regression function was applied in [20]. Hate speech detection employing LSTM utilizing TF-IDF was presented in [23]. DNN-based multi-task learning was proposed in [24]. In [25], design the ViTHSD - a targeted hate speech detection dataset for Vietnamese Social Media Texts.

The multilingual hate speech detection model is introduced in [26] to classify content in both Arabic and English. In [27], a stack ensemble classification system that classifies tweets into three groups: hate speech, abusive language, or neutral. The enhancements of automatic hate speech detection in Albanian social media using advanced Deep Neural Techniques were performed in [28].

A multilingual dataset in English and Urdu, and applied a translation-based approach, is designed in [29] to handle multilingual challenges and utilizes several state-of-the-art machine learning, deep learning, and transfer learning methods.

3. Proposed Methodology

Various research have been performing for hate speech detection over past few years in social media. The presence of various aspects causes issues of hate speech detection and results in enhanced issues not only to society, but also to policy-makers and researchers.

Thus, hate speech is fundamentally the exploitation of offensive language on social media. However, the misclassification of hate speech using existing work was higher, which reduces the accuracy of hate speech detection.

Besides, the amount of time required for hate speech analysis is not minimized. To address the issues mentioned above, the CRCOCDBC model has been developed.

The overall flow process of the proposed CRCOCDBC model is demonstrated in Figure 1 to attain better detection of hate speech. The CRCOCDBC model improves accuracy in detecting hate speech in online social networks with minimum time consumption.

3.1. Deep Belief Convolutional Neural Network

A Deep Belief Convolutional Neural Network is a specific type of Deep Neural Network that utilizes a mathematical function called convolution.

A proposed Deep Neural Network learning technique analyzes the text sample data for improving the accuracy of hate speech detection.

The main advantage of the Deep Convolutional Neural learning is the ability to achieve high accuracy rates while handling a large volume of social media text data with minimum error.

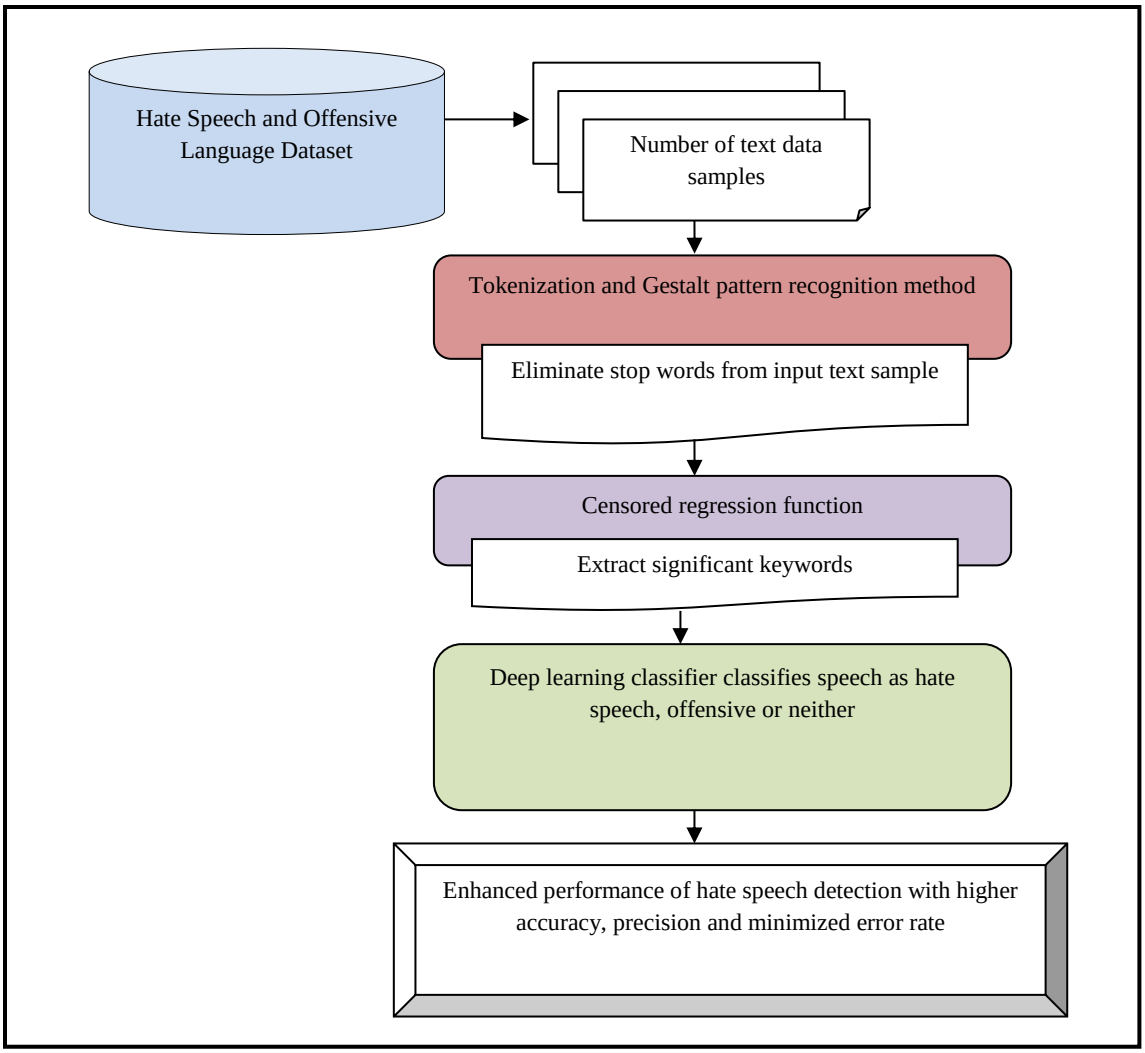

Fig. 1 Architecture diagram of CRCOCDBC model

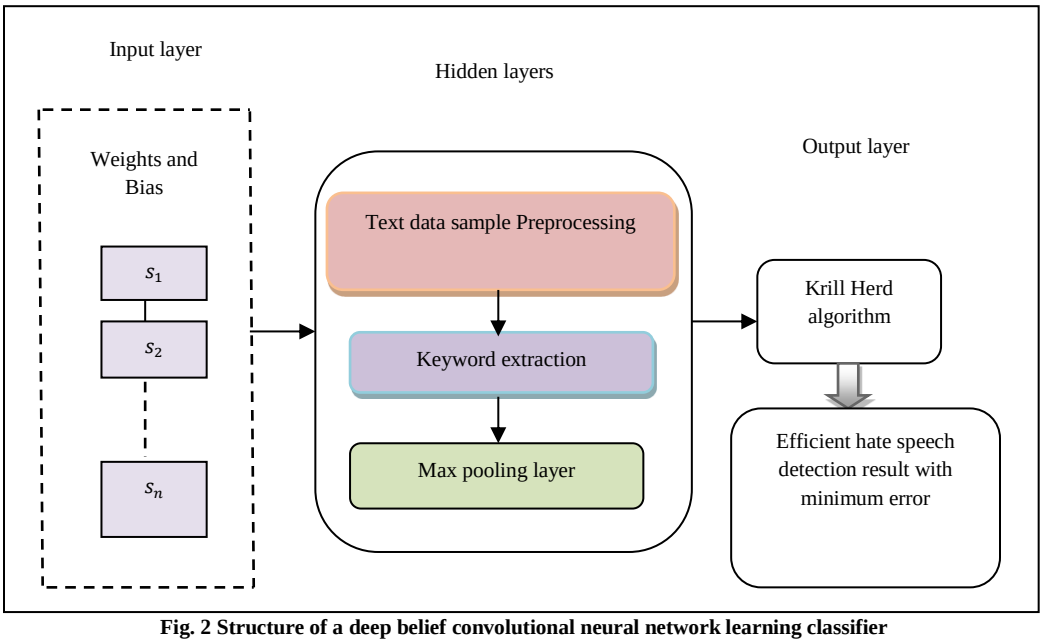

Fig. 2 Structure of a deep belief convolutional neural network learning classifier

Figure 2 illustrates the structure of a deep belief Convolutional Neural Network classifier comprising different types of layers, such as one input layer, multiple hidden layers, and one output layer. The input layer receives the input and data to be processed. A random number of hidden layers is positioned between the input and output layers. Within each layer, there are small individual units known as artificial neurons or nodes that handle the input text data samples and transmit them to neurons in the subsequent layer. The types of hidden layers carry preprocessing, keyword extraction, and classification with max-pooling layers. The classification results are ultimately obtained at the output layer. Let us consider the number of data samples $s = \{s_1, s_2, \dots, s_n\}$and '$m$' number of features '$f = \{f_1, f_2, \dots, f_m\}$' that collected from hate detection dataset 'DS'. Then, the activity of the neuron is formulated at the input layer and expressed as:

$$X(t) = [\sum_{i=1}^{n} s_i(t) * W_i] + b \quad (1)$$

From equation (1), the activity of neurons at the input layer '$X(t)$For each text, a data sample is determined. Here, 's_i' denotes sum of input data samples, 'W_i' indicates a weight and 'b' denotes a bias that stores the value of '1'. After that, the input data was transformed into the hidden layers to enhance hate speech detection in social media.

3.1.1. Data Sample Preprocessing

The CRCOCDBC model initially performs data sample preprocessing in the first hidden layer. Preprocessing is the process of preparing the text data through tokenization and stop word removal from the text data. In preprocessing, stop words are removed from the input text data. First, the tokenization process is carried out to separate text into a number of words for identifying stop words. Followed by, the Gestalt pattern recognition method is applied to identify stop words from the input texts. This process helps to minimize the complexity of detecting hate speech in online networks.

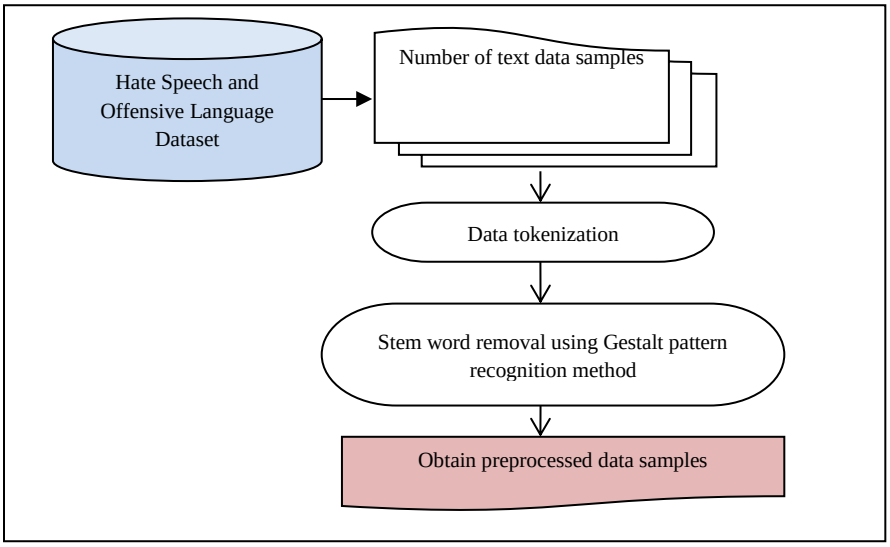

Fig. 3 Process of the preprocessing phase

The processing diagram of data sample preprocessing is illustrated in Figure 3 above to obtain preprocessed data samples. At first, input data samples are considered and separated into a number of words using a tokenization process. This tokenization splits the sentences from online media into words by means of punctuation and spaces in square brackets. Separated words are stored in a string for matching words from the user. Words extracted from the input are formulated as,

$$s = w_1, w_2, w_3, \dots . w_m \quad (2)$$

From (2), the data samples 's' are partitioned into several words. $w_1, w_2, w_3, \dots . w_m$ ' using tokenization. After the tokenization result, the stop words are identified and removed using the Gestalt pattern recognition method. The pattern recognition is applied for finding the stop words. Gestalt pattern recognition is a statistical method that is used to find the relationship between a dependent variable and one or more independent variables for identifying the stop words from the customer comments in social media.

Here, Gestalt Pattern Matching is estimated based on the similarity between the length of the output word string and the length of the input word string. Thus, data preprocessing removes stop words, thereby minimizing the detection time of hate speech. The mathematical formulation to measure Gestalt Pattern Matching is expressed as follows.

$$G_{PM} = \frac{2 * n_{mw}}{n_w} \quad (3)$$

From equation (3), Gestalt Pattern Matching. G_{PM} ' is estimated based on the number of matching words in the output string. 'n_{mw}' and the number of words in the input

string. 'n_w'. It provides the output values that range between zero and one.

$$G_{PM} = \begin{cases} 1\ ; & stopwords \\ 0\ ; & notstopwords \end{cases} \quad (4)$$

From the pattern matching results, the accurate stop word removal process is performed. If the matching result returns '1', it denotes that words are identified as stop words. If the matching result provides 0, it indicates words are identified as not stop words.

The identified stop words are removed, and the other words are used for performing the keyword extraction and hate speech detection process. As a result, the time complexity of detection is said to be minimized.

3.1.2. Censored Regression Function-Based Keyword Extraction

After the text data sample preprocessing, relevant keyword extraction is performed in hidden layer 2 of the deep learning classifier. In that layer, a censored regression function is used to select the relevant features from the preprocessed data. The Censored Regression is a Machine Learning Technique used to find statistically highly correlated results by defining the threshold value. Based on a specific threshold value, the frequent occurrence score of words is analyzed to select relevant keywords. The words with higher correlation score values are selected as relevant keywords. Censored regression function by comparing with the threshold is expressed as.

$$R = \begin{cases} C_T > \beta\ ; & Selected \\ C_T < \beta; & Removed \end{cases} \quad (5)$$

From equation (5), the output of the regression function 'R' is estimated with consideration of the threshold value 'β' and the correlation output.C_TThe regression function result is determined. If the correlation results are greater than the threshold, the word is chosen as a relevant keyword for accurate classification. Or else, correlation results are removed to minimize the dimension of the input data.

3.1.3. Max-Pooling Layer

After extracting the relevant keyword, data sample classification is performed at the Max pooling layer using CCA. Canonical Correlation Analysis measures the relationship between extracted keywords for classifying words into multiple classes. Correlation analysis involves analyzing training and testing variables to determine relationships between them for better detection.

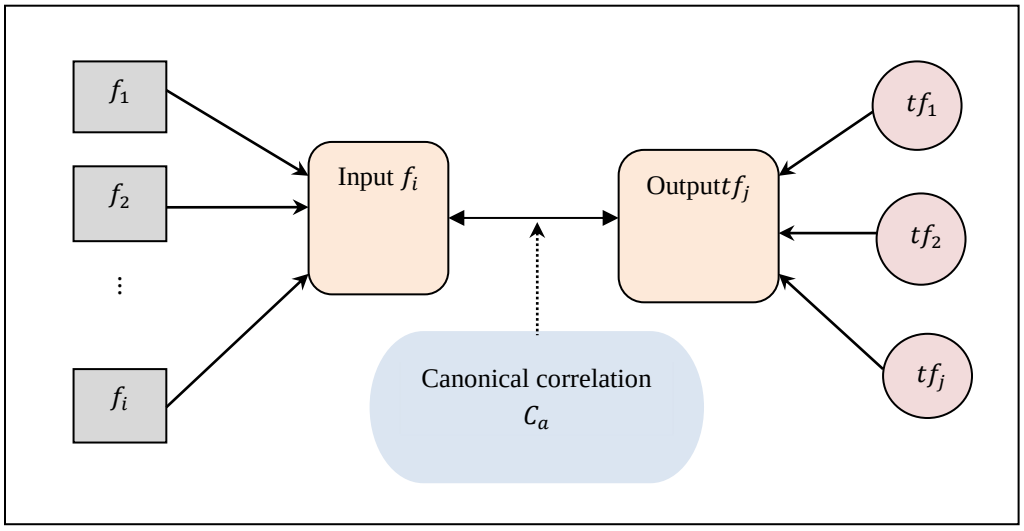

Fig. 4 Canonical correlation function

In the proposed Deep Neural Learning Network Model, canonical correlation is performed at the last hidden layer to identify the relationship between training features and testing keyword features of the data sample.

Canonical correlation is expressed as:

$$C_a = \frac{\sum(\varepsilon f_i - \overline{\varepsilon f_i})(tf_i - \overline{tf_i})}{n} \quad (6)$$

From equation (6), the Canonical Correlation. C_a ' is estimated. In the above expression, 'εf_i' denotes an extracted keyword feature value of data, '$\overline{\varepsilon f_i}$' refers to a mean value of the extracted feature, 'tf_i' refers to a testing feature value, '$\overline{tf_i}$' refers to a testing feature mean value, and 'n' point out the total number of the text data sample features extracted. The results of correlation are given to the output layer for detecting multiple classes of hate speech from the dataset. The result of the output layer for classifying given data to identify multiple levels of hate speech is expressed below.

$$Y(t) = \begin{cases} C_a = 0 & hatespeech \\ C_a = 1 & offenisve \\ C_a = 2 & neither \end{cases} \quad (7)$$

The classification result in the output layer '$Y(t)$' is represented in equation (7). 'C_a' denotes a correlation coefficient result. Based on the correlation value, multiple classes of hate speech are effectively detected, namely hate speech, offensive, and neither. The obtained classifier results include errors. Here, the weight in the network is updated, and the error is determined. Lastly, Krill Herd Optimization is employed to determine the minimum error results of classification.

Krill Herd Optimization is a nature-inspired metaheuristic optimization algorithm based on the collective behavior of krill swarms. This optimization simulates the social interactions and movement patterns of krill to solve multi-objective optimization problems.

The main advantages of Krill Herd Optimization over other optimization techniques are providing global optimization capabilities, diversity maintenance, efficiency, robustness, and user-friendly implementation. During the fine-tuning process, the proposed deep learning classifier optimizes hyperparameters, including weights, to minimize classification errors.

For each data sample classification result, the error is computed and mathematically expressed as follows.

$$\delta = \sum_{i=1}^{n} \left| {C_{\alpha i}}^{iden} - {C_{\alpha i}}^{act} \right| \quad (8)$$

From (8), classification error 'δ' is measured based on.${C_{\alpha i}}^{iden}$, identified classification results and. '${C_{\alpha i}}^{act}$refers to actual classification results. With the aid of error estimation, the classification result with minimum error is provided at the output layer as follows.

$$Y(t) = arg\ min[\delta] \quad (9)$$

Final classification result with minimum error '$Y(t)$' is obtained using equation (9). '$arg\ min$' denotes the argument of the minimum function, and 'δ' symbolizes error.

Multi-class classification results for hate speech detection achieve higher precision at the output layer.

// Algorithm 1:Censored Regressive Canonical Optimized Convolutional Deep Belief Classifier Model

Input: Hate Speech and Offensive Language Dataset, number of data samples $s = \{s_1, s_2, \dots, s_n\}$, features '$f = \{f_1, f_2, \dots, f_m\}$'

Output: Enhanced hate speech detection

Begin

1. Collect the number of data samples '$s_1, s_2, \dots, s_n$' from dataset ----[input layer]
2. For each data sample'ss_n'
3. Perform preprocessing ---[hidden layer 1]
4. For each word in the string
5. Remove stop words using Gestalt Pattern Matching
6. $if\ \ (G_{PM} = 1)$ then
7. Stop words are identified and removed
8. Else
9. Words s are identified as not stop words
10. End if
11. End for
12. End for
13. Perform keyword extraction ---[hidden layer 2]
14. Compute censored regression 'R'
15. Select relevant keywords
16. Perform classification process ---hidden layer 3
17. For each selected keyword.f_i'
18. Measure the canonical correlation 'δ'
19. It identifies hate speech, offensive, and neither
20. End for
21. Krill Herd Optimization
22. return classification result with minimum error at the output layer

End

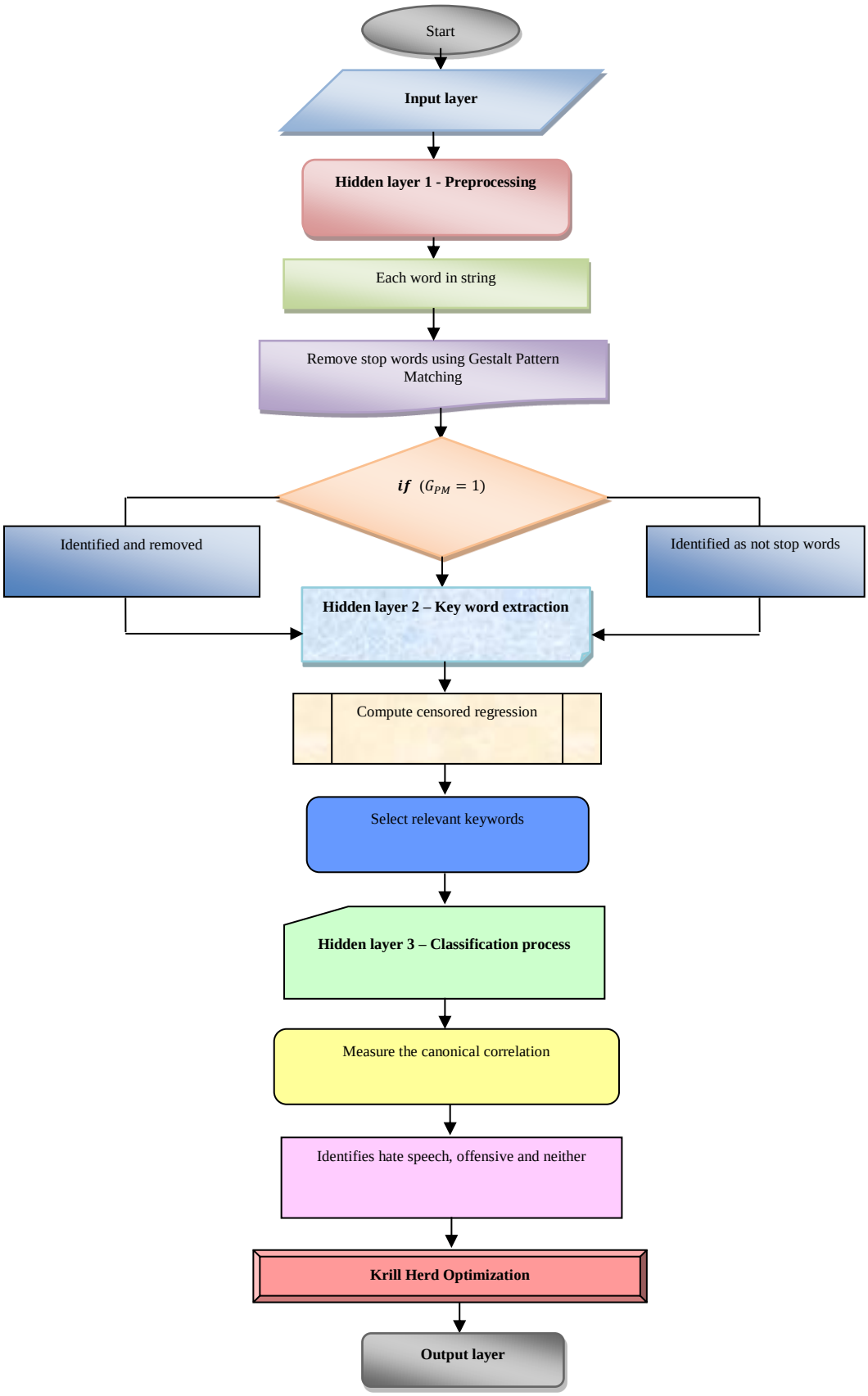

Fig. 5 Workflow diagram of the proposed CRCOCDBC model

4. Experimental Settings

Experimental assessment of the CRCOCDBC model and existing Conv-BiRNN-BiLSTM framework [1], FAST-RNN technique [2], and Advanced Deep Neural Techniques [28] are implemented using Python language with R Statistical Programming Tool. Hate Speech and Offensive Language Dataset is taken from https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset. This dataset is used to predict hate speech in online social networks based on the various processes, namely, preprocessing, keyword extraction, and classification. The dataset consists of 7 different attributes or features along with 25296 data samples. The attributes are index, count, hate_speech, offensive_language, neither, class, and tweet. The dataset is split into a 70% training dataset, 10% validation, and 20% test sets.

5. Performance Results Analysis

The performance of the CRCOCDBC model and existing [1, 2, 28] models is determined in terms of accuracy, precision, recall, F-measure, error rate, and classification time with various numbers of data samples.

5.1. Impact of Accuracy

It is measured as a ratio of the number of text data samples from social media that are correctly detected as hate speech. It is measured in percentage (%). Accuracy is formulated as given below,

$$Accuracy = \left[\frac{p_t+n_t}{p_t+n_t+p_f+n_f}\right] * 100 \quad (10)$$

Table 1. Values of accuracy

Number of data samples	Accuracy (%)			
	Conv-BiRNN-BiLSTM framework	FAST-RNN technique	CRCOCDBC model	Advanced deep neural techniques
2500	86	89	95	92
5000	85.6	87.5	95	93
7500	87.6	89.2	94.7	91.8
10000	85.98	87.3	94.3	92.4
12500	87.8	88	93	91
15000	86.3	89.5	93.15	91.5
17500	88.4	89.1	92	90
20000	84.85	90.5	92.8	92
22500	86.2	90.85	94.5	93
25000	88.7	91	95.2	94.6

Table 1 Presents Experimental Results of Accuracy. The average of comparison results indicates that the accuracy of CRCOCDBC is increased by 8%, 5% and 2% to [1, 2, 28].

5.2. Impact of Precision

Precision is measured as the ratio of the true positives to the false positives in the data samples. The formula for calculating precision is given below.

$$Precision = \left[\frac{p_t}{p_t+p_f}\right] * 100 \quad (11)$$

Figure 6 Depicts the Performance Results of Precision Using Three Methods. Overall Comparison Results Indicate Precision is Improved By 6.9%, 4.8% and 2% over [1, 2, 28].

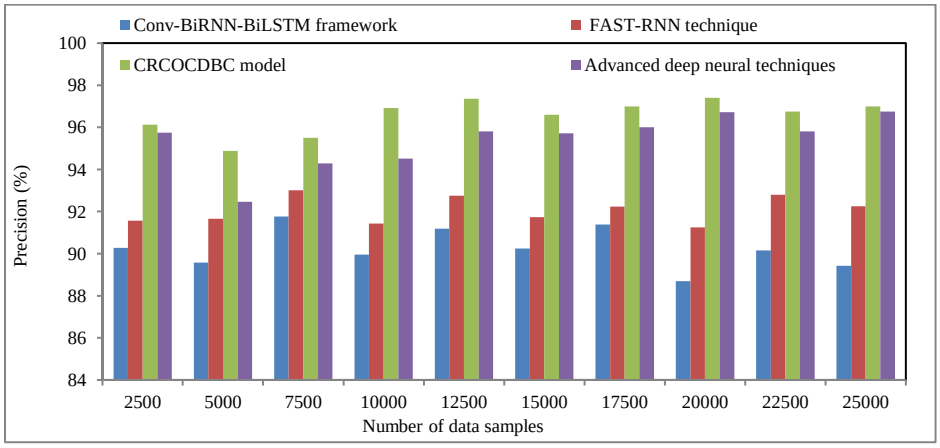

Fig. 6 Evaluation results of precision

5.3. Impact of Recall

Recall is measured as the ratio of the number of true positives to the number of false negatives in the data samples. Recall is computed as given below,

$$Recall = \left[\frac{p_t}{p_t+n_f}\right] * 100 \quad (12)$$

Table 2. Values of recall

Number of data samples	Recall (%)			
	Conv-BiRNN-BiLSTM framework	FAST-RNN technique	CRCOCDBC model	Advanced deep neural techniques
2500	93.3	96	98	97
5000	92.4	94.55	96.21	95.47
7500	93.11	94.53	95.97	95
10000	93.51	94.55	95.82	95.27
12500	93.07	93.88	95.32	95.10
15000	94.06	94.8	96.12	95.87
17500	93.86	94.67	95.72	94.95
20000	94.778	95.08	96.05	95.76
22500	93.039	93.98	95.02	94.66
25000	94.031	94.8	95.706	95.22

The recall performance analysis is shown in Table 2. Overall performance of recall using CRCOCDBC is improved by 2.6%, 1.3% and 1% in [1, 2, 28].

5.4. Impact of F-Measure

F-measure is computed based on the result of precision and recall. It is formulated as,

$$f - measure = 2 * \left[\frac{precision*recall}{precision+recall}\right] * 100 \quad (13)$$

Figure 7 illustrates the F-measure. The overall comparison results indicate F-measure of CRCOCDBC is improved by 5%, 3.5% and 2% to [1, 2, 28].

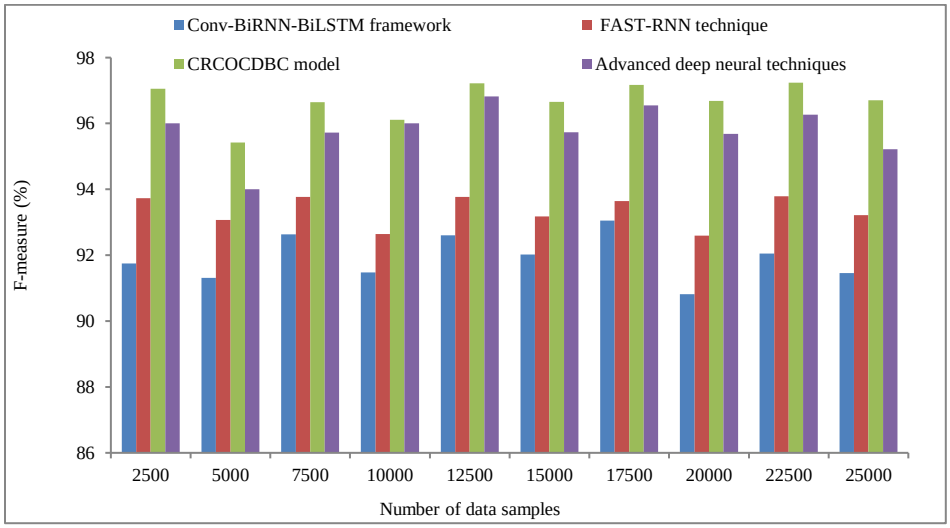

Fig. 7 Evaluation results of F-measure

5.5. Impact of Error Rate

Error rate is evaluated as a ratio between the number of data samples that are wrongly categorized as hate speech and the total number of input data samples. It is stated as given below.

$$ER = \sum_{i=1}^{n} \frac{S_{wronglydet}}{S_i} * 100 \quad (14)$$

Error rate 'ER' is measured based on the input data samples. S_i' and wrongly detected data samples '$S_{wronglydet}$'. It is measured as a percentage (%).

Table 3 Details Overall Performance Results of the Error Rate Versus Number of Data Samples.

Overall Performance of Error Rate is Minimized Using CRCOCDBC by 54%, 43% and 26% to the [1, 2, 28].

Table 3. Values of error rate

Number of data samples	Error rate (%)			
	Conv-BiRNN-BiLSTM framework	FAST-RNN technique	CRCOCDBC model	Advanced deep neural techniques
2500	14	11	5	8
5000	14.4	12.5	5	9
7500	12.4	10.8	5.3	7
10000	14.02	12.7	5.7	7.5
12500	12.2	12	7	10
15000	13.7	10.5	6.85	8.75
17500	11.6	10.9	8	9
20000	15.15	9.5	7.2	8.3
22500	13.8	9.15	5.5	7.3
25000	11.3	9	4.8	6.8

5.6. Impact of Classification Time

Classification time is measured as the amount of time consumed for detecting hate speech in English in an online social network with a Deep Learning Neural Network. Time is calculated as,

$$C_{Time} = \sum_{i=1}^{n} s_i * time\ (classifysingledata) \quad (15)$$

Figure 8 Provides a Performance Analysis of Data Classification Time Using Different Methods. Overall Results Indicate CRCOCDBC Minimizes Time by 26%,17% and 10% to the [1, 2, 28].

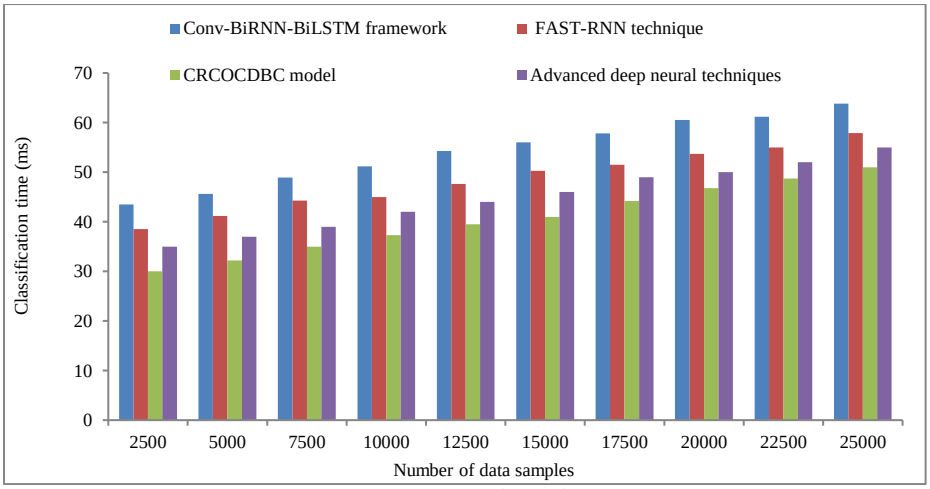

Fig. 8 Evaluation results of classification time

6. Discussion

This study compares the proposed CRCOCDBC technique with the existing Conv-BiRNN-BiLSTM framework [1], the FAST-RNN technique [2], using the Hate Speech and Offensive Language Dataset based on different metrics, namely, accuracy, precision, recall, F-measure, error rate, and classification time with various numbers of data samples. In this approach, a number of text data samples from an online dataset are taken as input for hate speech detection.

Initially, the unwanted words, which are stop words, are identified and removed through the text data preprocessing. Here, tokenization and the Gestalt pattern recognition method are applied to identify stop words from input text data. This helps to minimize the time consumption of the detection process. With preprocessed words, the significant keywords are extracted based on the censored regression function along with a word-specific threshold. After the keyword selection, the classification is performed using Canonical Correlation Analysis into a max-pooling Deep Neural Network.

The correlation measures the relationships between keywords to provide better data classification results. Finally, the multi-class classification results for hate speech detection are obtained at the output layer with higher accuracy and minimized error rate. The results confirm the proposed CRCOCDBC method.

7. Conclusion

The novel approach, named the CRCOCDBC model, is suggested for detecting multiple classes of hate speech in social media platforms. The quantitative analysis confirms

that the CRCOCDBC model has achieved higher accuracy of hate speech detection with lesser time consumption as well as fewer errors when compared to other methods. The limitations of hate speech detection in online social networks include the subtlety and context-dependency of language, which can lead to errors, the biases inherent in training data, which can reinforce prejudice, and the difficulty of building universal models due to cultural and linguistic diversity. The subjective nature of hate speech, the challenge of creating high-quality datasets, and the opaque decision-making of automated systems also present significant hurdles. In the future, hate speech detection in online social networks will be driven by the advancement of Large Language Models (LLMs) like GPT-4 and transformer-based architectures, focusing on multilingualism, context awareness, and handling nuanced language like code-mixing. Key areas include improving dataset quality and creating standardized frameworks for consistent and unbiased detection across different domains and languages, especially low-resource ones.

Future research will also focus on developing models that can assist human moderators without introducing bias and exploring real-time detection for proactive moderation.

Author contributions

The corresponding author claims a significant contribution to the paper, including formulation, analysis, and editing. The co-author provides guidance to verify the analysis result and manuscript editing.

Compliance with ethical standards

This article is an entirely original work of its authors; it has not been published before and will not be sent to other publications until the journal's editorial board decides not to accept it for publication.

References

[1] Hareem Kibriya et al., "Towards Safer Online Communities: Deep Learning and Explainable AI for Hate Speech Detection and Classification," *Computers and Electrical Engineering*, vol. 116, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ehtesham Hashmi, and Sule Yildirim Yayilgan, "Multi-Class Hate Speech Detection in the Norwegian Language using FAST-RNN and Multilingual Fine-Tuned Transformers," *Complex and Intelligent Systems*, vol. 10, no. 3, pp. 4535-4556, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Gretel Liz De la Peña Sarracén, and Paolo Rosso, "Systematic Keyword and Bias Analyses in Hate Speech Detection," *Information Processing and Management*, vol. 60, no. 5, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Gloria del Valle-Cano et al., "SocialHaterBERT: A Dichotomous Approach for Automatically Detecting Hate Speech on Twitter Through Textual Analysis and user Profiles," *Expert Systems with Applications*, vol. 216, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Devansh Mody et al., "A Curated Dataset for Hate Speech Detection on Social Media Text," *Data in Brief*, vol. 46, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Djamila Romaissa Beddiar, Md. Saroar Jahan, and Mourad Oussalah, "Data Expansion using Back Translation and Paraphrasing for Hate Speech Detection," *Online Social Networks and Media*, vol. 24, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Krishanu Maity et al., "A Deep Learning Framework for the Detection of Malay Hate Speech," *IEEE Access*, vol. 11, pp. 79542-79552, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Juan Manuel Pérez et al., "Assessing the Impact of Contextual Information in Hate Speech Detection," *IEEE Access*, vol. 11, pp. 30575-30590, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] "Retracted: Analysing Hate Speech Against Migrants and Women through Tweets using Ensembled Deep Learning Model," *Computational Intelligence and Neuroscience*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ashfia Jannat Keya et al., "G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media," *IEEE Access*, vol. 11, pp. 79697-79709, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Neeraj Vashistha, and Arkaitz Zubiaga, "Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media," *Information*, vol. 12, no. 1, pp. 1-16, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Faiza Mehmood et al., "Passion-Net: A Robust Precise and Explainable Predictor for Hate Speech Detection in Roman Urdu Text," *Neural Computing and Applications*, vol. 36, no. 6, pp. 3077-3100, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] José Antonio García-Díaz et al., "Evaluating Feature Combination Strategies for Hate-Speech Detection in Spanish using Linguistic Features and Transformers," *Complex and Intelligent Systems*, vol. 9, no. 3, pp. 2893-2914, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ishaani Priyadarshini, Sandipan Sahu, and Raghvendra Kumar, "A Transfer Learning Approach for Detecting Offensive and Hate Speech on Social Media Platforms," *Multimedia Tools and Applications*, vol. 82. no. 18, pp. 27473-27499, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Lanqin Yuan et al., "Transfer Learning for Hate Speech Detection in Social Media," *Journal of Computational Social Science*, vol. 6, no. 2, pp. 1081-1101, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yasmine M. Ibrahim, Reem Essameldin, and Saad M. Darwish, "An Adaptive Hate Speech Detection Approach using Neutrosophic Neural Networks for Social Media Forensics," *Computers, Materials and Continua*, vol. 79, no. 1, pp. 233-262, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[17] Md Abul Bashar et al., "Progressive Domain Adaptation for Detecting Hate Speech on Social Media with Small Training Set and its Application to COVID-19 Concerned Posts," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1-18, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Prashant Kapil, and Asif Ekbal, "A Deep Neural Network based Multi-Task Learning Approach to Hate Speech Detection," *Knowledge-Based Systems*, vol. 210, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[19] Vaishali U. Gongane, Mousami V. Munot, and Alwin D. Anuse, "Detection and Moderation of Detrimental Content on Social Media Platforms: Current Status and Future Directions," *Social Network Analysis and Mining*, vol. 12, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] D.C. Asogwa et al., "Hate Speech Classification using SVM and Naive BAYES," *arXiv Preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Gyorgy Kovacs, Pedro Alonso, and Rajkumar Saini, "Challenges of Hate Speech Detection in Social Media," *SN Computer Science*, vol. 2, no. 2, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Shivang Agarwal, and C. Ravindranath Chowdary, "Combating Hate Speech using an Adaptive Ensemble Learning Model with a Case Study on COVID-19," *Expert Systems with Applications*, vol. 185, pp. 1-9, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Sanjiban Sekhar Roy et al., "Hateful Sentiment Detection in Real-Time Tweets: An LSTM-based Comparative Approach," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5028-5037, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Shankar Biradar, Sunil Saumya, and Arun chauhan, "Fighting Hate Speech from Bilingual Hinglish Speaker's Perspective, A Transformer and Translation based Approach," *Social Network Analysis and Mining*, vol. 12, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Cuong Nhat Vo et al., "ViTHSD: Exploiting Hatred by Targets for Hate Speech Detection on Vietnamese Social Media Texts," *Journal of Computational Social Science*, vol. 8, no. 2, pp. 1-34, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Hassan AL-Sukhani et al., "Multilingual Hate Speech Detection: Innovations in Optimized Deep Learning for English and Arabic Hate Speech Detection," *SN Computer Science*, vol. 6, no. 3, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Muhammad Mubeen et al., "Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms using Stack Ensemble Classification Method," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Endrit Fetahi et al., "Enhancing Social Media Hate Speech Detection in Low-Resource Languages using Transformers and Explainable AI," *Social Network Analysis and Mining*, vol. 15, no. 1, pp. 1-30, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Muhammad Ahmad et al., "Multilingual Hope Speech Detection from Tweets using Transfer Learning Models," *Scientific Reports*, vol. 15, no. 1, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]