SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Swish-Activated Deep Neural Bigru Framework With Contextual Information For Fine-Grained Multilingual Hate Speech Detection

Ms. N Zahira Jahan[1] , Dr. R Pushpalatha[2]

[1]Research Scholar, Kongu Arts and Science College, Erode, Tamilnadu, India & Assistant Professor, Department of Computer Applications, Nandha Engineering College, Erode, Tamilnadu, India Email: zahirajahann999@gmail.com
[2]Associate Professor, Department of Computer Science, Kongu Arts and Science College, Erode, Tamilnadu, India
Email: rpljour@gmail.com

Abstract
The growing prevalence of hate speech in multilingual and code-mixed social media environments necessitates robust fine-grained detection models capable of capturing semantic, contextual, and language-specific variations. In this study, a Swish Activated Deep Neural Bi-directional Gated Recurrent Unit with Contextual Information (SADNB-CI) is developed for fine-grained hate speech detection across Tamil, English, and Tanglish social media text. The proposed framework addresses the limitations of existing hate speech detection approaches by integrating Global Dense Vector Representation-based embedding, Swish-activated convolution, max pooling, CNN-based feature representation, BiGRU-based sequential modelling, and contextual information learning for multi-label classification. Separate embedding representations are generated for Tamil, English, and Tanglish inputs, while convolutional filters extract discriminative local features and the BiGRU layer captures forward and backward contextual dependencies. Contextual hate speech categories, including community-based, religion-based, gender-based, and political hate speech, are incorporated to enhance the distinction between hate speech and non-hate speech and to support fine-grained classification. The SADNB-CI model is evaluated using the AI Tamil Hate Speech Detector dataset with precision, recall, accuracy, and false positive rate as performance metrics. Experimental findings demonstrate that the proposed model outperforms k-means+textGCN and t-HateNet across Tamil, English, and Tanglish samples by achieving higher precision, recall, and accuracy while reducing the false positive rate. The results indicate that the integration of Swish-activated CNN feature extraction with BiGRU-based contextual learning improves multilingual and code-mixed hate speech detection, thereby establishing SADNB-CI as an effective approach for fine-grained social media hate speech classification.

Keywords: Hate Speech Detection, Global Dense Vector Representation, Swish Activation, Bi-direction Gated Recurrent Unit, Contextual Information.

1. Introduction

K-means+textGCN classifier with BERT was proposed in [1] with three different Turkish hate speech datasets. However, it failed to analyze false positive rate. Transfer learning method called, t-HateNet was proposed in [2] for English language. However, precision and recall were not focused. Binary classification method was proposed in [3]. Uncertainty-aware cross-modal alignment (UCA) method was proposed in [4]. In [5], recurrent neural network classifier was proposed. Deep learning methods were applied in [6]. LSTMwas presented in [7]. ML and DL investigated in [8]. Multilingual embedding model was designed in [9]. Fusion approach presented in [10].Sequential model was designed in [10]. Contributions of work are followed as, To propose fine-grained hate speech detection method called, SADNB-CI. To implement Global Dense Vector Representation-based

embedding to map speech obtained from social media and generates three embedding layers separately for Tamil, English and Tanglish. To apply consolidated network layer advantage of both CNNs and BiGRU networks by synthesizing their outputs to generate new model for generating accurate results. To prove SADNB-CIfine grained hate speech detection with minimal FPR. To simulation analysis and performance of proposed SADNB-CIagainst existing methods.

2. Related works

Systematic review of hate speech detection was proposed in [12]. ML and DL techniques were presented in [13]. Review of DL techniques was investigated in [14]. DL and hate-speech detection methods were designed in [15]. Challenges of hate speech detection were investigated in [16]. ML and DL were presented in [17]. DNN proposed in [18-19] for hate speech classification. Application of Bidirectional encoder representation designed [20]. Roman Urdu text was presented in [21]. LSTM presented in [22] ML for hate speech detection was analyzed in [23]. DL using multiple text datasets was presented in [24].K-means with feed forward neural networks were designed in [25] for clustering and classification of English. Previous existing studies did not identify false positive rate. To address these issues, Proposed SADNB-CI designed using DNN concept. DNN used for hate speech detection in [26], [27], [28], [29] with different language such as Tamil, English and Urdu etc.

3. Methodology

Objective of this work is to identify maximum number of hate speech 'HS' related classes from AI Tamil Hate Speech Detector dataset extracted from https://github.com/dreamspace-academy/ai-tamil-hate-speech-projectas by users. It is classified into two classes either Hate Speech 'HS'or Non-Hate Speech 'NHS'. This section introduces a comprehensive representation of the method architecture proposed in this work.

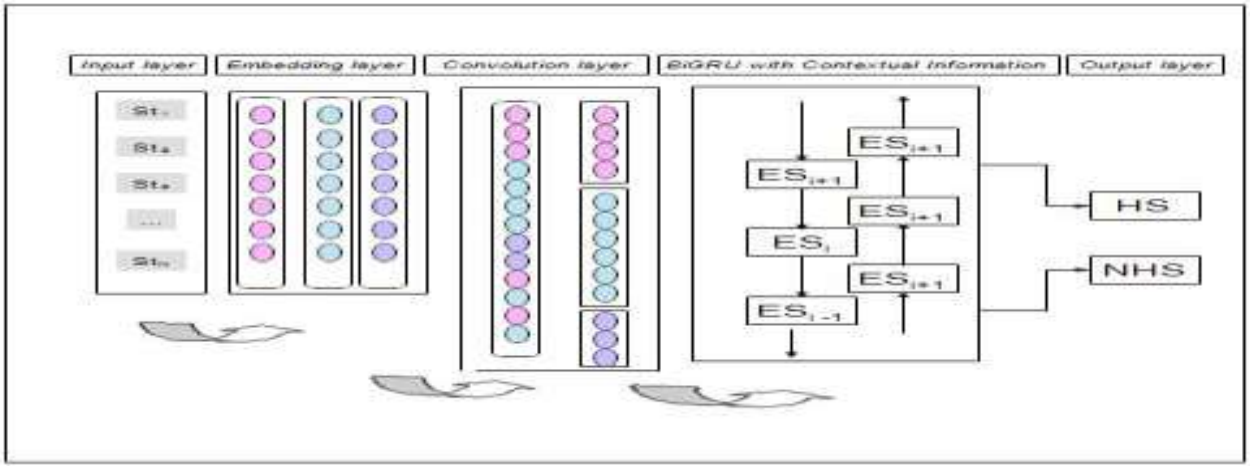

Figure 1 Structure of SADNB-CI fine grained hate speech detection

Figure 1 shows the architecture of proposed SADNB-CI fine grained hate speech detection method. From above figure, the problem statement IS represented as follows. For baseline model, conventional machine learning based classifiers were employed and made a detailed comparison and the necessitated features was extracted using tf-idf models. In contrast, proposed method with preprocessed and selected features as input using convolution process employing multiple filters present in intermediate layers performs the process of classification of hate speech in an accurate and precise manner. The proposed method consists of string embedding and convolutional layers integrated to BiGRU with Contextual Information fine-grained hate speech detection.

3.1 Input and Global Dense Vector Representation-based Embedding Layer

Global Dense Vector Representation-based (GloDenseVe) embedding layer is utilized for converting the speech obtained from social media into a numerical vector shape. Here the Global Dense Vector Representation-based embedding layer represents the generation of three distinct embedding layers. The term embedding here maps the speech obtained from social media and produces a speech matrix. With this generated speech matrix the

number of strings in each sample makes the use of padding proportionate. After successful padding each sample has a specified number of strings that say 'N *i.e.*, $ST = \{ST_1, ST_2, ..., ST_N\}$'. Now from the Global Dense Vector Representation-based embedding model for all terms 'N' the respective preprocessed samples vector extracted and concatenated in conjunction to form a sample matrix 'P' with a size of '$N*d$' where 'N' denotes the number of strings and 'd' representing the proposed word embedding dimension. Then the embedding layer has a matrix 'EM' of vocabulary size 'V_{size}' and each row of 'EM' contains the vector representation of string in the vocabulary. Then, the vector 'e_i' for each string 'ST_i' the embedding layer generates matrix multiplication results as given below.

$e_i = ST_i * EM$, *where* $EM = e_1, e_2, ..., e_N$ *of size* $N*d$ (1)

From (1) results string embedding matrix (embedding layer) 'EM' is updated during training process by reducing a loss function with reference to the parameters of the model (with respect to the string 'ST_i' and 'd' dimension of dense words to be learned). String embedding matrix of corresponding string embedding layer learns to encapsulate the semantic correlations between strings in the vocabulary. Embedding matrix 'EM' is then augmented or fed as input to following layers of DNN for facilitating processing and classification. This work has created three distinct embedding layers, one for generating vectors for Tamil strings, second for generating vectors for English string and third for generating vectors for Tanglish respectively.

3.2 Swish Activated-Convolution Layer

It is one of crucial elements of Deep Convolutional Neural Networks. It comprises of learnable filter '$ST \in R^{Nd}$', where 'N' represents the number of strings and 'd' is the dimensions. This filter is utilized in extracting local features by convolving on 'N' strings at time and performs element-wise dot product to acquire feature 'F' with acquiring more informative words or informative strings that boost the overall method performance. This convolution layer applies 128 filters of size 3 to input sequence (i.e. preprocessed samples), generating 128 feature maps as output. Output of convolution operation is then passed through Swish activation function.

$$H_{ij} = Swish(P) = \sum_{k=1}^{K} W_k U + B \quad (2)$$

$$U = \left(P * \frac{1}{1+e^{-P}} \right) \quad (3)$$

From (2) and (3), 'i','j' represent point of output feature map with 'k' denoting kernel size, '$Swish(P)$' Swish activation function for corresponding preprocessed samples. Filter weights 'W_k' are learned to acquire significant patterns from corresponding preprocessed samples and term bias 'B' is included to each output of corresponding feature map with putting in place shift in activation function. Resulting output feature map consists of set of activation values denoting presence of different patterns (Tamil, English and Tanglish) in preprocessed samples.

3.3 Max Pooling Layer

Also pooling layer is utilized to pool out indispensable features from the extracted set of features. Convolution process utilized the extracted features 'ES' from the preprocessed samples 'P' to convolve and activate them. However, it is not mandatory that all extracted features 'ES' from the preprocessed samples are predominant one. In additional with the intent of minimizing the computing overhead of the model, the selected but indispensable features required to be carried on with. By applying pooling minimizes the vector representation spatial size therefore aiding in getting along with overfitting. The most prevalent pooling operation is performed using max over time pooling operation that takes into consideration the local maximum value to encapsulate the indispensable features from convoluted feature map results.Max pooling layer with pooling window of size '2' expressed below.

$$Q_{ij} = \max\left\{ES_{i,2j}, ES_{i,2j+1}, ES_{i,2j+3}, ...\right\} \quad (4)$$

From (4), 'Q_{ij}' forms '$j-th$' output of '$i-th$' feature map post max pooling with '$ES_{i,2j}$','$ES_{i,2j+1}$' is outputs of prior convolution layer at points '$2j$' and '$2j+1$'.

3.4 CNN Output Layer

Output layer constitutes fully connected dense layer with ReLu activation function that returns the maximum of 0 and the input value that denotes that any negative values are set to 0.

$$\mathrm{Re}\, s_{CNN} = Q = f\left(W^T P + B\right) \quad (5)$$

From (5), 'Q' and 'P' is output and input layer uses weight 'W', bias 'B' by ReLU activation function 'f'.

3.5 BiGRU with Contextual Information Fine-Grained Hate Speech Detection

BiGRU, a type of RNN, is specifically utilized in sequential modeling issues formulating series from forward and backward handlings. In BiGRU, forward '$\overrightarrow{\mathrm{GRU}}$' and backward '$\overleftarrow{\mathrm{GRU}}$' GRUs combined to recover successive '(f_1 to f_{128})' and previous feature series '(f_{128} to f_1)'. Proposed method retrieves forward and backward series possessing contextual information (i.e., hate-speech related categories, community-based, religion-based, gender-based and political) by BiGRU layer on convoluted pooled layer output. Fine-grained task employed involved multi-label classification. Figure represents Contextual Information fine-grained hate speech detection employed. For each speech and its respective contextual information, first category necessitated to mark whether speech is hateful or not. According to mentioned categories, speech regarding hateful or non hateful is analyzed.

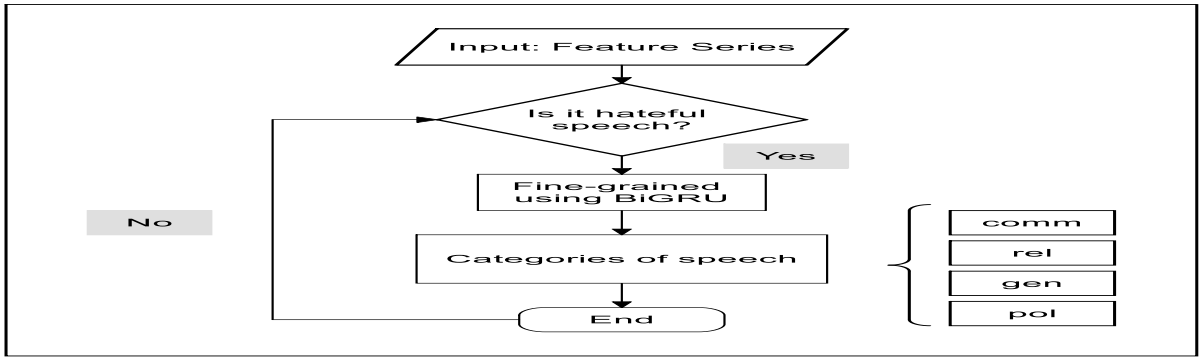

Figure 2 BiGRU with Contextual Information Fine-Grained Hate Speech Detection

From above figure, with feature series (preprocessed and selected features) provided as input are subjected to BiGRU-based sequential modeling. Fine-grained task involving multi-label classification based analysis is performed. Categories included in dataset for multi-label classification formulated.

$$comm = comm + 1 \quad (6)$$

$$rel = rel + 1 \quad (7)$$

$$gen = gen + 1 \quad (8)$$

$$pol = pol + 1 \quad (9)$$

From equations (6), (7), (8) and (9), clustered categories stored in form of vector as given below.

$$Cat = \left[comm, rel, gen, pol\right] \quad (10)$$

From (10) fine-grained hate speech detection for four different categories performed. Equations (11) and (12) given below denotes results from BiGRU with Contextual Information in forward and backward handlings respectively. Then, BiGRU with Contextual Information output is the hate-incorporating representation of preprocessed feature extracted input text by combining the contexts from the forward and backward handlings. BiGRU with Contextual Information-based representation for given convoluted pooled layer output, denotes concatenation of forward 'ES_{for}', backward 'ES_{back}' and hidden states respectively.

$$\overrightarrow{ES_{for}} = \overrightarrow{GRU}\left(Cat\left(Q_{ij}\left[P\right]\right)\right) \quad (11) \quad \overleftarrow{ES_{back}} = \overleftarrow{GRU}\left(Cat\left(Q_{ij}\left[P\right]\right)\right) \quad (12)$$

Finally equation (11) and (12) given above denotes combined Contextual Information-based representation series as a final hidden state 'H_t' that is passed to fusion layer to produce resultant predicted output results.

$$\mathrm{Re}\, s_{BiGRU} = H_t = \left[\overrightarrow{ES_{for}}, \overleftarrow{ES_{back}}\right] \quad (13)$$

From (13), multi-label output is designed that concurrent prediction call-to-action label ('HS'or'NHS') respectively.

3.6 Model fusion

Consolidated network layer in our proposed method takes advantage of both CNNs and BiGRU networks by synthesizing their outputs to generate a new model. With output of each neural network i.e., CNN and BiGRU network being a vector characterization, Consider output tensor from CNN as $\mathrm{Re}\,s_{CNN}$ and output tensor from BiGRU network as $\mathrm{Re}\,s_{BiGRU}$, then resultant consolidated network layer is formulated as given below.

$$\mathrm{Re}\,s = \mathrm{Re}\,s_{CNN} \oplus \mathrm{Re}\,s_{BiGRU} \quad (14)$$

From (14), $\oplus$ represent element-wise operation being performed by combing both tensor outputs.

3.7 Fine-Grained Training

Training process for fine-grained models with output of the model being a vector of probabilities for each output variable (four characteristics plus call-toaction), a multi-label loss function is employed that takes into consideration the probability of each class separately. 'Cat' be number of output variables (4 in our case), '$Q \in \{0,1\}^{Cat}$ the true label vector and '$Q' \in [0,1]^{Cat}$' the predicted probabilities. Loss function defined as given below.

$$L\left(\mathrm{Re}\,s, \mathrm{Re}\,s'\right) = \sum_{l=1}^{Cat} L_{bin}\left(\mathrm{Re}\,s_i, \mathrm{Re}\,s'_i\right) \quad (15)$$

$$L_{bin}\left(\mathrm{Re}\,s_i, \mathrm{Re}\,s'_i\right) = -\mathrm{Re}\,s \log\left(\mathrm{Re}\,s'\right) - \left(1-\mathrm{Re}\,s\right)\log\left(1-\mathrm{Re}\,s'\right) \quad (16)$$

Where, $\mathrm{Re}\,s$ is true label results ('0' or '1') and '$\mathrm{Re}\,s'$' is predicted results for binary detection ('HS' or 'NHS').

4. Experimental setup

Proposed SADNB-CI, existing [1], [2] evaluated using AI Tamil Hate Speech Detector dataset in Python working on MS Window platform with several metrics.

4.1 Performance analysis of Tamil, English and Tanglish

Precision, recall and accuracy involved in hate speech detection for Tamil, English and Tanglish is analyzed in [27], [30].

$$\mathrm{Pr}\,e = \frac{TP}{TP+FP} \quad (17)$$

$$\mathrm{Re}\,c = \frac{TP}{TP+FN} \quad (18)$$

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (19)$$

From (17), (18) and (19), 'TP', 'TN'is true positive and negative, 'FP', 'FN'is false positive, and negative.

Table 1 Precision, recall and accuracy analyses using SADNB-CI, k-means+textGCN [1] and t-HateNet [2] fine grained hate speech detection (English)

Samples	Precision rate (English)			Recall rate (English)			Accuracy (English)		
	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]
10000	97.53	92.15	86.35	98.59	90.35	82.35	96.5	91	87

20000	97.7	90	85	98.35	89.15	82	96.1	90	85
30000	97.28	89	84	98.43	88	81	96.11	88.35	82
40000	97.16	88.15	83.15	98.42	87.55	79.55	96.05	88	81
50000	97.03	87	82	98.29	87.35	79	96	87.15	80
60000	97.02	85.35	80	98.21	87	78.35	95.83	86	78
70000	97	85	79.35	98.16	86.35	78.15	95.71	85.35	75
80000	96.15	84.15	79	97.75	86.15	78	95.12	84	72
90000	95	83	75	97.6	86	77.55	95	83	70
100000	93.25	82.15	74.15	97.55	85.25	77	94.9	82	69

Table 1 shows tabulation of precision, recall and accuracy for English. From the above tabulation for 10000 samples, where the precision, recall and accuracy using SADNB-CI method for English was found to be 97.53%, 98.59% and 96.5% respectively whereas using the existing methods [1] and [2] it was found to be 92.15%, 86.35%, 90.35%, 82.35%, 91% and 87% respectively.

Table 2 Precision, recall and accuracy analyses using SADNB-CI, k-means+textGCN [1] and t-HateNet [2] fine grained hate speech detection (Tamil)

Samples	Precision rate (Tamil)			Recall rate (Tamil)			Accuracy (Tamil)		
	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]
10000	97.98	95	90	98.35	92.15	90.25	96	94	90
20000	97.33	93	88.35	98.05	90	88.55	95.2	92	88
30000	96.73	91	87	98	88.35	87.35	94.83	90	86
40000	96.58	90	84	97.65	85	83	94.37	89.35	85.35
50000	95.15	88	83.25	97.35	82.15	79.15	94.2	89	85
60000	95	86	81	96	81	78	92.83	87	84
70000	93.15	84.35	80	95.85	80.35	77.55	91	86.35	83
80000	92	82	79.55	94	78.15	77.25	90.75	84	81.15
90000	91	81	78	90	78	77	89.35	82	81
100000	90	80	77.35	89	77.35	76.35	89	81	80

Table 2 illustrates tabulation of precision, recall and accuracy using three methods. From Table 2 above for 10000 samples, where the precision, recall and accuracy using SADNB-CI method for Tamil speech from social media was observed to be 97.98%, 98.35% and 96% respectively whereas using the existing methods [1] and [2] it was found to be 95%, 90%, 92.15%, 90.25%, 94% and 90% respectively.

Table 3 Precision, recall and accuracy analyses using SADNB-CI, k-means+textGCN [1] and t-HateNet [2] fine grained hate speech detection (Tanglish)

Samples	Precision rate (Tanglish)			Recall rate (Tanglish)			Accuracy (Tanglish)		
	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]
10000	96.92	93.15	90	98.9	96	91	94.2	92	88
20000	95.73	92	88.45	98.77	93.15	90	92.95	90	86
30000	94.44	90	85	98.46	92	87.25	91.53	88	85
40000	93.85	90	83	97.72	90	85	90.85	85	82
50000	93.92	88.35	82.15	97.66	89	82	90.6	83	80
60000	93.73	87	80	97.45	88.45	80	90.2	81	78
70000	92.75	84	78	96.97	83.15	78	89.77	80	75
80000	91.95	82.15	77	96.75	82	75	89.57	78	73
90000	91.75	81	75.35	96.22	80	73	88.91	75	72

100000	90	80	75	95.71	78.45	71	88.45	73	71

Table 3shows tabulation of precision, recall and accuracy for three methods for Tanglish speech. From above tabulation for 10000 samples, where the precision, recall and accuracy using SADNB-CI method for Tanglishwas observed to be 96.92%, 98.9% and 94.2% respectively whereas using the existing methods [1] and [2] it was observed to be 93.15%, 90%, 96%, 91%, 92% and 88% respectively.The overallprecisionis improved using proposed SADNB-CI method by 12%, 20% (English), 9%, 14% (Tamil) and 8%, 15%(Tanglish)over to [1] and [2]. Overallrecallenhanced using proposed SADNB-CI method by 12%, 24% (English), 15%, 17% (Tamil) and 12%, 21% (Tanglish)over [1] and [2]. Overallaccuracyimproved using proposed SADNB-CI method by 11%, 24% (English), 6%, 10% (Tamil) and 10%, 15% (Tanglish)over [1] and [2].

4.2 Performance Analysis of False Positive Rate 'FPR'

It evaluated as percentage ratio between number of negative events wrongly categorized as positive.

$$FPR = \frac{FP}{FP + TN} \qquad (20)$$

Table 4 False Positive Rate Analysis

Samples	**False Positive Rate (%) – English**			**False Positive Rate (%) – Tamil**			**False Positive Rate (%) – Tanglish**		
	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]	SADNB-CI	k-means+textGCN [1]	t-HateNet [2]
10000	2.45	2.95	3.15	1.35	1.75	2.25	2	2.55	2.85
20000	2.85	3.25	3.35	1.75	2.25	2.55	2.15	2.75	3.15
30000	3	3.55	3.85	2.15	2.45	2.85	2.55	3	3.35
40000	3.15	3.85	4	2.35	2.95	3.15	2.85	3.35	3.70
50000	3.35	4	4.35	2.55	3	3.35	3	3.55	4
60000	4	4.35	4.85	2.85	3.15	3.55	3.15	3.95	4.35
70000	4.15	4.55	5.15	3	3.35	3.85	3.35	4	4.85
80000	4.25	4.85	5.85	3.15	3.85	4	3.55	4.25	5
90000	4.55	5	6.25	3.35	4	4.35	3.85	4.45	5.25
100000	5	5.35	7	3.55	4.35	4.55	4	5	5.85

Table 4 shows tabulation of False Positive Rate using three methods. However from the above simulation for 10000 samples, the false positive rate using SADNB-CI method reduced for English, Tamil and Tanglishwas found to be 2.45%, 1.35% and 2% respectively. Also, the false positive rate minimized for English, Tamil and

Tanglish2.95% and 3.15%, 1.75% and 2.25%, 2.55% and 2.85% using [1] and [2] respectively.Overall false positive rate of SADNB-CI method improved by 12%, 22% (for English speech), 17%, 25% (for Tamil speech), and 18%, 28% (for Tanglish speech) respectively.

5. Conclusion

Proposed SADNB-CI fine grained hate speech detection has been presented withGlobal Dense Vector Representation-based embedding, Swish Activated convolution and BiGRU with Contextual Information fine-grained hate speech detection in three different languages, Tamil, English and Tanglish. Proposed SADNB-CI provides better precision, recall, accuracy and FPR as compared to the existing similar methods. Results of SADNB-CI showreducedfalse positive rate by 17%, 21% and 23% for English, Tamil and Tanglish hate speech detection.

References

[1] Ayse Berna Altinel, Gozde Karatas Baydogmus, Sema Sahin, Mustafa Zahid Gurbuz, "So-haTRed: A Novel Hybrid System for Turkish Hate Speech Detection in Social Media with Ensemble Deep Learning Improved by BERT and Clustered-Graph Networks", IEEE Access, Jun 2024,Volume 12, Pages 86252 – 86270. DOI: 10.1109/ACCESS.2024.3415350

[2] Lanqin Yuan, Tianyu Wang, Gabriela Ferraro, Hanna Suominen, Marian-Andrei Rizoiu, "Transfer learning for hate speech detection in social media", Journal of Computational Social Science, Springer, Oct 2023, Volume 6, Pages 1081-1101.https://doi.org/10.1007/s42001-023-00224-9

[3] Olumide Ebenezer Ojo, Thang-Hoang Ta, Alexander Gelbukh, Hiram Calvo, Grigori Sidorov, Olaronke Oluwayemisi Adebanji, "Automatic Hate Speech Detection Using Deep Neural Networks and Word Embedding", Computación y Sistemas, Volume 26, Issue 2, Oct 2022, Pages 1-7. https://doi.org/10.13053/cys-26-2-4107

[4] Chuanpeng Yang, Fuqing Zhu, Yaxin Liu, Jizhong Han, Songlin Hu, "Uncertainty-Aware Cross-Modal Alignment for Hate Speech Detection", 4 ELRA Language Resource Association, Mar 2024, Pages 16973–16983.https://aclanthology.org/2024.lrec-main.1475/

[5] Georgios K. Pitsilis, Heri Ramampiaro, Helge Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks", Applied Intelligence, Springer, Jul 2018, Volume 48, Pages 4730-4742.https://doi.org/10.1007/s10489-018-1242-y

[6] Kevin Usmayadhy Wijaya, Erwin Budi Setiawan, "Hate Speech Detection Using Convolutional Neural Network and Gated Recurrent Unit with FastText Feature Expansion on Twitter", Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, Sep 2023, Volume 9, Issue 3, Pages 619-631. DOI: 10.26555/jiteki.v9i3.26532

[7] Chandan Senapati, Utpal Roy, "Bengali Hate Speech Detection Using Deep Learning Technique", CEUR Workshop Proceedings, Oct 2021, Pages 1-10. https://ceur-ws.org/Vol-3681/T6-21.pdf

[8]Alaa Marshan, Farah Nasreen Mohamed Nizar, Athina Ioannou, Konstantina Spanaki, "Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online", Information Systems Frontiers, Springer, Nov 2023, Volume 27, Pages 487-505. DOI:10.1007/s10796-023-10446-x

[9] Ayme Arango Monnar, Jorge Perez Rojas, Barbara Polete Labra, "Cross-lingual hate speech detection using domain-specific word embeddings", PLOS ONE, Jul 2024, Volume 20, Issue 4, Pages 1-10.https://doi.org/10.1371/journal.pone.0306521

[10] Waqas Sharif, Saima Abdullah, Saman Iftikhar, Daniah Al-Madani, SHAHZAD MUMTAZ, "Enhancing Hate Speech Detection in the Digital Age: A Novel Model Fusion Approach Leveraging a Comprehensive Dataset", IEEE Access, Feb 2024, Volume 12, Pages 27225 – 27236.DOI: 10.1109/ACCESS.2024.3367281

[11] Ashraf Ullah, Khair Ullah Khan, Aurangzeb Khan, Sheikh Tahir Bakhsh, Atta Ur Rahma, Sajida Akbar, BibiSaqia, "Threatening language detection from Urdu data with deep sequential model", PLOS ONE, Jun 2024, Volume 19, Issue 6, Pages 1-14. doi: 10.1371/journal.pone.0290915

[12] Md Saroar Jahan, Mourad Oussalah, "A systematic review of hate speech automatic detection using natural language processing", Neurocomputing, Elsevier, May 2023, Volume 546, Issue 6, Pages 1-30. https://doi.org/10.1016/j.neucom.2023.126232

[13] Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models", Alexandria Engineering Journal, Elsevier, Aug 2023, Volume 80, Pages 110-121.https://doi.org/10.1016/j.aej.2023.08.038

[14] Ara Zozan Miran, Adnan Mohsin Abdulazeez, "Detection of Hate-Speech Tweets Based on Deep Learning: A Review", (Jurnal Informatika dan Sains, Dec 2023, Volume 6, Issue 2, Pages 1-15. DOI: https://doi.org/10.31326/jisa.v6i2.1813
[15] Jitendra Singh Malik, HezheQiao, Guansong Pang, Anton van den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study", arxiv, Dec 2023, Volume 1, Pages 1-18. DOI:10.48550/arXiv.2202.09517
[16] György Kovács, Pedro Alonso, Rajkumar Saini, "Challenges of Hate Speech Detection in Social Media", Computer Science, Springer, Feb 2021, Volume 2, Issue 95, Pages 1-15. DOI:10.1007/s42979-021-00457-3
[17] Zainab Mansur, Nazlia Omar, Sabrina Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities", IEEE Access, Feb 2023, Volume 11, Pages 16226 – 16249.
DOI: 10.1109/ACCESS.2023.3239375
[18] Shankar Biradar, Sunil Saumya, Arun chauhan, "Fighting hate speech from bilingual hinglish speaker's perspective, a transformer and translation based approach", Social Network Analysis and Mining, Springer, Nov 2022, Volume 12, Issue 87.
doi: 10.1007/s13278-022-00920-w.
[19] Prashant Kapil,Asif Ekbal, "A deep neural network based multi-task learning approach to hate speech detection", Knowledge-Based Systems, Elsevier, Oct 2020, Volume 210, Pages 106458.https://doi.org/10.1016/j.knosys.2020.106458
[20] Ashwin Geet d'Sa, Irina Illina, Dominique Fohr, "Classification of Hate Speech Using Deep Neural Networks", Data and Information Processing to Knowledge Organization: Architectures, Models and Systems, Jan 2021, Volume 25, Issue 1, Pages 1-12. https://hal.science/hal-03101938v1/document
[21] Faiza Mehmood, Hina Ghafoor, Muhammad Nabeel Asim, Muhammad Usman Ghani, Waqar Mahmood, Andreas Dengel, "Passion-Net: a robust precise and explainable predictor for hate speech detection in Roman Urdu text", Neural Computing and Applications, Springer, Nov 2023, Volume 36, Pages 3077-3100. https://doi.org/10.1007/s00521-023-09169-6
[22] Sanjiban Sekhar Roy, Akash Roy, Pijush Samui, Mostafa Gandomi, and Amir H. Gandomi , Senior Member, IEEE, "Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach", IEEE Transactions on Computational Social Systems, Oct 2023, Volume 11, Issue 4, Pages 5028-5037.DOI: 10.1109/TCSS.2023.3260217
[23] Ali Alhazmi, Rohana Mahmud, Norisma Idris, Mohamed Elhag Mohamed Abo, Christopher Ifeanyi Eke, "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models", PLOS ONE, Jul 2024, Volume 19, Issue 7,Pages 1-24.
https://doi.org/10.1371/journal.pone.0305657
[24]Md Abul Bashar, Richi Nayak, Khanh Luong, Thirunavukarasu Balasubramaniam, "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, Springer, Jul 2021, Volume 11, Issue 69, Pages 1-18. doi: 10.1007/s13278-021-00780-w
[25] Ramanpreet Kaur and Amandeep Kaur, "Text Document Clustering and Classification using K-Means Algorithm and Neural Networks", Indian Journal of Science and Technology, Oct 2016, Volume 19, Issue 40, Pages 1-5. DOI: 10.17485/ijst/2016/v9i40/97722
[26] Purbani Kar, "Towards Secure Social Platforms: Hate Speech Detection and Classification in Indian Languages Using Hybrid Soft Computing Techniques", Springer, July 2025, Pages 1-53. DOI:10.21203/rs.3.rs-7085357/v1
[27] Asif Hasan, Tripti Sharma , Azizuddin Khan ,and Mohammed Hasan Ali Al-Abyadh, "Analysing Hate Speech against Migrants and Women through Tweets Using Ensembled Deep Learning Model", Computational Intelligence and Neuroscience, Hindawi, April 2022, Volume 2022, Pages 1-9.doi: 10.1155/2022/8153791.
[28] Jayasree Ravi, "Handwritten alphabet classification in Tamil language using convolution neural network", International Journal of Cognitive Computing in Engineering, Elsevier, June 2024, Volume 5, Pages 132-139. https://doi.org/10.1016/j.ijcce.2024.03.001
[29] Muhammad Ahmad, Iqra Ameer, Wareesa Sharif, Sardar Usman, Muhammad Muzamil, Ameer Hamza, Muhammad Jalal, Ildar Batyrshin & Grigori Sidorov, "Multilingual hope speech detection from tweets using transfer learning models", Scientific Reports, Springer, March 2025, Volume 15, Issue 9005, Pages 1-17. DOI:10.1038/s41598-025-88687-w