International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 279-285.
www.theijire.com
ISSN No: 2582-8746

Deep Learning Models for the Hate Speech Detection: A Survey

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

How to cite this paper:
Mohini Chakarverti, "Deep Learning Models for the Hate Speech Detection: A Survey", IJIRE-V4I02-279-285.

Abstract: *Hate can be directed at somebody based on their gender, colour, religion, ethnicity, etc. All expression that spreads, incites, supports, or justifies racism, xenophobia, antisemitism, or other types of intolerance, including intolerance expressed by hostile nationalism and ethnocentrism, discrimination against minorities, immigrants, and people of immigrant origin, may be considered hate speech. The deep learning models is the unsupervised learning models which can learn from the patterns. The convolution layer of the CNN model is used for the pattern detection. The last layer called dense layer will classify data into certain classes. In this paper various deep learning model technique for the hater speech detection is reviewed and analysed in terms of certain parameters.*

Key Word: *Hate Speech, Deep Learning, CNN*

I.INTRODUCTION

Over the past ten years, social media has grown tremendously in both scope and significance as a communication tool. Due to the open nature of social media, anyone can post whatever they want, advocating any viewpoint, whether it be instructive, disgusting, or somewhere in between. The number of persons who can see such posts varies depending on the forum [1]. The definition of improper content varies among forums, as do the methods for recognising it, however given the size of the medium, automated methods play a significant role in this effort. A key component of this offensive information is hate speech. Speaking hatefully is an antisocial behaviour. Hate can be directed at somebody based on their gender, colour, religion, ethnicity, etc. All expression that spreads, incites, supports, or justifies racism, xenophobia, antisemitism, or other types of intolerance, including intolerance expressed by hostile nationalism and ethnocentrism, discrimination against minorities, immigrants, and people of immigrant origin, may be considered hate speech. Yet, the word "hate speech" is ambiguous and has many definitions. Regardless of how the phrase or issue is defined, it is evident that there are instances where automated systems for detecting hate speech are required. It is crucial that the techniques used in these situations are precise, effective, and efficient [2].

1.1 Automatic Hate Speech Detection

Recent developments in natural language processing (NLP) technology have allowed for the completion of a number of research on the automatic detection of hate speech in text. Several competitions have been held by well-known contests including SemEval-2019 and 2020 and GermEval-2018 in an effort to advance automated hate speech identification. To enable field study, scientists have created large databases with data from a number of sources. The issue of hate speech in numerous non-English languages and online groups has also been examined in many of these research [3]. Researchers are compelled to examine and contrast alternative processing pipelines, including feature set selection, Machine Learning (ML) methods, classification algorithms, and more as a result. Examples of these include Naive Bayes, Linear Regression, Convolution Neural Network (CNN), LSTM, and BERT deep learning architectures.

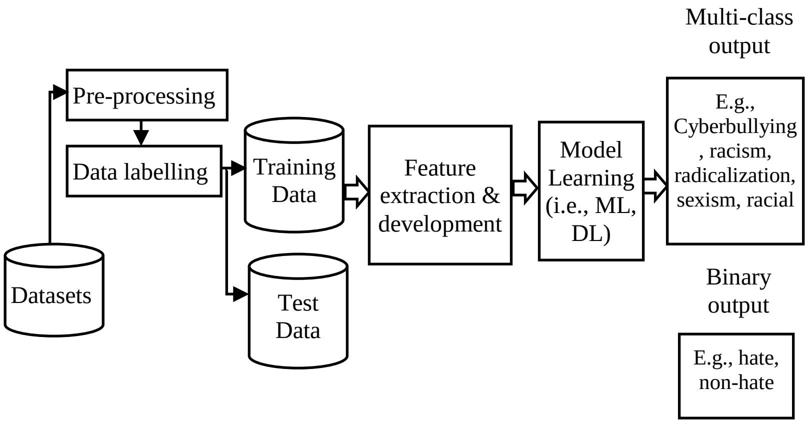

Figure 1: Typical automatic HS detection system pipeline

Figure 1 depicts the overall workflow for the HS identification task, which is based on a text classification system. The dataset collection and preparation phase is where the HS detection pipeline begins. Datasets are commonly gathered via social media platforms like Face book, YouTube, Twitter, and others. Pre-processing is carried out in accordance with the dataset's quality and structure [4]. Typically, this involves filtering and normalising textual inputs, which may include lemmatization, stop word removal, misspelling correction, noise reduction, and tokenization, among other things.It is also seen that it's possible that the dataset will be sent to customers immediately, obviating the necessity for collection. When preparing the dataset for the subsequent machine learning stage, the training and testing sections of the dataset should be divided. In the analysis step after feature engineering, the necessary characteristics are subsequently extracted from the textual inputs, turning the unstructured text sequences into structured features. Popular feature extraction techniques include the TF-IDF, semantic, lexical, topic modelling, sentiment [5], BOW, and word embedding (FastText, GloVe, and Word2Vec). Sometimes dimension reduction is used to reduce the complexity of 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) [6]. 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) [7]. 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).

1.2 Deep Learning for Hate Speech Detection

Hate speech detectors based on deep neural networks are referred to as deep learning methods. Any feature encoding technique, including established ones like TF-IDF and more recent ones like word embedding or pre-training techniques, may be used [8] to encode the input data for these neural networks. The latter strategy, which helps avoid conventional feature engineering or feature construction procedures, is typically more effective than the previous way. Instead, it picks up feature representations from the texts that are being read. Convolutional neural networks (CNN), long short-term memory (LSTM), and bi-directional LSTM (Bi-LSTM) are a few common deep neural network architectures.LSTM models are used to learn the words that have a long-range dependency of the characters, while CNN models are used for learning compositional aspects of words or characters in hate speech detection. A brief discussion of all these models is provided below:

i. Long Short-Term Memories (LSTM): These particular neural network types were created with the goal of performing effectively while dealing with sequential data sets and long-term dependencies [9]. When one wants a network to retain information for a longer period of time, these networks can be helpful. This characteristic qualifies LSTM for handling textual data. An LSTM's typical architecture is depicted in Fig. 1. An LSTM is a group of cells that are identical to one another, as shown in the diagram, and each cell processes the input in a certain way. Each cell also takes input from the cell that came before it in the chain, in addition to input from external sources. This cell design makes it easier for the LSTM to retain earlier information for longer periods of time.

ii. Bi-Directional Long Short-Term Memories (Bi-LSTM): The standard version of an LSTM can recall or make references to the data it has already processed. Nevertheless, it lacks any supporting documentation for the information that was provided after the point was reached. This turns into a significant disadvantage when working with sequence data, particularly text [10]. Another LSTM variant that may retain data from both ways is called bidirectional LSTM. Backpropagation is possible in two ways with a bidirectional LSTM. From the front and the back, respectively. Bi-LSTTM is a potent tool for textual data analysis thanks to this procedure.

iii. Recurrent Neural Network: It has been demonstrated that recurrent models deliver highly solid outcomes for language modelling. Recurrent neural networks (RNNs) are so-called feedback neural networks in which directed cycles can arise in the connections between the neurons.

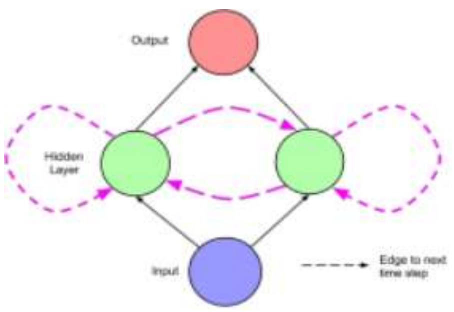

Figure 2: A simple recurrent network

The semantics of every preceding piece of text are stored in a hidden layer when this network examines a text word by word. A recurrent neural network is a type of neural network that works with the variable length sequence x $= (x_1, \dots . x_T)$ and has a hidden state h and an optional output y. The RNN's hidden state, $h_{(t)}$, is changed at each time step t by [11]:

$$h_{(t)} = f(h_{(t-1)}, x_t).$$

Where f is an activation function that is not linear. The long short-term memory (LSTM) unit is an example of a very sophisticated function yet it can also be extremely basic. One of the most well-liked and effective techniques for minimising the consequences of vanishing and exploding gradients is LSTM. So, it has the capacity to learn enduring dependencies. RNN has its advantage of capturing the contextual information in a better method, which could be helpful when dealing with large material [12]. However, because this model is biased, using it across the entire document could lessen its usefulness.

iv. Convolutional Neural Network: Convolutional neural networks, often known as CNN or ConvNet, have been developed to address the bias issue that arises with recurrent neural networks. Figure 3 depicts how a convolutional neural network works.

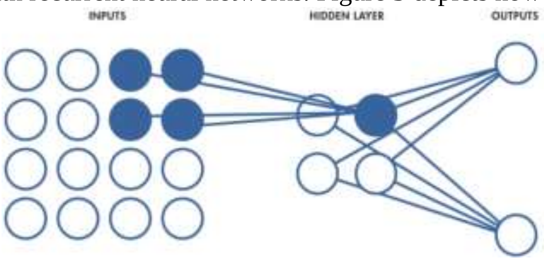

Figure 3: CNN Architecture

Only a small portion of the input layer neurons in this architecture are connected to the neurons in the hidden layer. A CNN is a deep, feed-forward artificial neural network with numerous hidden layers, an input, and an output [13]. A CNN is different because it has a convolutional multi-layer that recognizes characteristics in the input space automatically. The networks are typically employed for applications like picture and sound recognition, but more recent uses include text classification. When classifying hate speech, CNN instinctively collects word or character combinations (such as phrases or n-grams) from Tweets, whereas RNN learns word or character dependencies (orderly information).

II.LITERATURE REVIEW

D. R. Beddiar, et.al (2021) suggested an innovative DL (deep learning)-based technique for fusing a BT (Back Translation) technique, and a Paraphrasing method in order to augment the data [14]. Diverse WEs (word-embedding) based methods were put forward to classify the hate speech. The BT method was planned on the basis of ED (encoder–decoder) model, pre-trained on an enormous corpus and often worked effectively to translate the machine. Furthermore, the transformer model and the mixture of experts deployed in the paraphrasing for producing different paraphrases. Eventually, a comparative analysis was conducted on LSTM (Long Short-Term Memory) and CNN (Convolution Neural Network) for attaining improved outcomes. Five datasets: AskFm corpus, Formspring, Warner and Waseem, Olid, and Wikipedia toxic comments dataset, were executed to quantify the suggested technique. Some related outcomes indicated that this technique was efficient and reliable. The suggested technique yielded a recall of 0.997 and a precision of 0.996 on the expanded Warner and Waseem dataset, and best accuracy and F1 score up to 0.994 on expanded Wikipedia toxic comments dataset.

C. -C. Wang, et.al (2022) introduced a novel method for creating a political hate speech lexicon and training AI (artificial intelligence) classification algorithms for detecting the hate speech [15]. The major emphasize was on gathering the Chinese hate speech dataset for which a Chinese hate speech lexicon was generated and a DL (deep learning)-based and a lexicon-based method was developed for detecting the Chinese hate speech. The fundamental objective was to detect the hate speech, however, the introduced method and the developed method assisted in detecting the hate speech from other languages. Additionally, the extended hate speech dataset was applied for detecting the hate speech on the basis of BERT (Bidirectional Encoder Representations from Transformers)-DL model. The experimental results depicted that the introduced method offered a precision of 55.4% using second method and 69.7% with BERT model.

S. Khan, et.al (2022) developed a new DNN (deep neural network) algorithm known as HCovBi-Caps, in which convolutional, BiGRU (Bidirectional-Gated Recurrent Unit), and CapsNet (capsule network) layers were combined to detect the hate speech [16]. This algorithm employed CapsNet to exploit the contextual information at diverse orientations. Two datasets namely DS1 (balanced) and DS2 (unbalanced) utilized for computing the developed algorithm while classifying the hate speech from general text. The results exhibited that the precision of developed algorithm was 90%, recall of 80% and F-Score of 84% on latter dataset. Moreover, the supremacy of the developed algorithm was proved over the existing techniques. The effect of several hyperparameters of neural and CapsNet was considered for analyzing the efficiency of this algorithm.

Y. Zhou, et.al (2020) presented a DLF (Deep Learning Based Fusion) technique in which distinct ML (machine learning) techniques namely ELMo (Embeddings from Language Models), BERT (Bidirectional Encoder Representation from Transformers) and CNN (Convolutional Neural Network) were implemented to classify the text [17]. SemEval 2019 Task 5 executed to simulate these techniques for detecting the hate speech. The work focused on enhancing the performance for which the results of the implemented techniques were fused and the results of 3 CNN algorithms were fused with diverse metrics. The outcomes revealed that the presented method assisted in enhancing the accuracy and F1-score for classifying the hate speech. Additionally, the practicality of the presented method was proved at lower cost.

S. Khan, et.al (2022) recommended BiCHAT in which a new BiLSTM (Bidirectional Long Short-Term Memory) algorithm, deep CNN (convolutional neural network) and HADL (Hierarchical ATtention-based deep learning) algorithm were deployed jointly to detect the hate speech when the tweet representation was learned [18]. This approach utilized tweets for input and BERT (Bidirectional Encoder Representations from Transformers) layer employed them. Thereafter, convolutional encoded representation inserted in Attention Aware-Bi-LSTM algorithm. The next task was to label the tweet as hateful or normal in the a softmax layer. Twitter dataset helped to train and compute the recommended algorithm. The recommended algorithm led to enhance the baseline techniques concerning precision up to 8%, recall by 7% andf-score up to 8%. Moreover, this algorithm improved the accuracy around 0.5 to train the data and 0.9 to validate the data.

Z. Mossie, et.al (2019) projected a HSD (hate speech detection) method for detecting the hate speech against vulnerable minority groups on social media [19]. First of all, SDP (Spark distributed processing) model employed to gather and pre-process the posts, and the word n-grams and word embedding methods namely Word2Vec exploited for extracting the attributes. After that, DL (deep learning) called GRU (Gated Recurrent Unit), which were a kind of RNN (Recurrent Neural Network) put forward for detecting the hate speech when it was classified. In the end, the utilized method assisted in clustering the hate words for predicting the potential target group for hatred. Amharic language in Ethiopia was utilized as instance for carrying out the experiments. The experimental results confirmed that the projected method outperformed the traditional models to detect the hate speech, and effective for recognizing Tigre ethnic group as susceptible community concerning hatred. G. d. Valle-Cano, et.al (2022) established an expert system to recognize and monitor the evolution of hate speech on Twitter based on an LTSM+MLP (Long Short-Term Memory and MultilayerPerceptron) algorithm [20]. Thereafter, a HaterBERT model was generated on the basis of BERT (Bidirectional Encoder Representations from Transformers) and HaterNet's dataset applied to test it. A method was further utilized for creating a user database as a relational network for inferring the textual and centrality features. This resulted in testing the Social Graph against diverse algorithms. Finally, a SocialHaterBERT algorithm was formulated in which both the earlier methods were integrated subsequent to analyze the attributes. According to the experimental results, the formulated algorithm had generated the optimal results for detecting the hate speech.

I. Z. Muhammad, et.al (2020) developed a system for detecting the hate speech in the form of tweets on twitter [21]. This system was planned on the basis of DBN (Deep Belief Network) technique for which the Global Vector feature was weighted for maximizing the accuracy prior to classify the hate speech. This technique was capable of discovering and detecting the hate speech from the text in advance. The analysis results validated that the accuracy of the developed technique was calculated 86%, precision was 82%, recall was 89.13% and Fl-Score was found 85.42%. Thereafter, a computer was further utilized for discovering and classifying the presence of hate speech in the text.

Rahul, et.al (2021) focused on designing an independent and self-sufficing framework for classifying Hingish texts as hate speech, abusive or non-offensive [22]. This approach aimed to deploy the CLEs (character level embeddings) for Hinglish Language so that the context was extracted from Hinglish sentences according to the level of variation in syntax and semantics of the code-mixed language. The next goal was to train different DL (deep learning) methods. Afterward, GRU (Gated Recurrent Unit) was integrated with Attention Model and attained higher efficiency. CLE, GRU, and attention layer were deployed for detecting the hate speech in Hinglish Code-Mixed Language.The experimental results indicated the robustness and applicability of the designed framework to learn complex dependencies while detecting the hate speech in comparison with the other methods.

J. Melton, et.al (2020) suggested a new model that had 3 objectives [23]. At first, an ensemble of DL (deep learning) algorithms was projected in which the potentials of existing methods were integrated. At second, a tuning factor was deployed in this model for leveraging TL (transfer learning) so that the hate speech was classified automatically on unlabeled datasets called Gab. At third, a WSL (weak supervised learning) technique was designed for training the system on unlabeled data. The projected approach offered a recall of 83% on HON dataset in contrast to the conventional methods. In addition, when the designed method was trained with classifier, the efficiency of the suggested model was maximized on unlabeled data from Gab and yielded the recall of 67%.

2.1 Comparison Table

Author	Year	Technique Used	Results	Limitations
D. R. Beddiar, et.al	2021	an innovative DL (deep learning)-based technique	The suggested technique yielded a recall of 0.997 and a precision of 0.996 on the expanded Warner and Waseem dataset, and best accuracy and F1 score up to 0.994 on expanded Wikipedia toxic comments dataset.	This technique was incapable of tuning the parameters of the back-translation automatically and capturing the grammar and semantic meaning.
C. -C. Wang, et.al	2022	DL (deep learning)-based and a lexicon-based method	The experimental results depicted that the introduced method offered a precision of 55.4% using second	The dataset employed in this work was small and only153 terms were comprised in the

			method and 69.7% with BERT model.	lexicon.
S. Khan, et.al	2022	HCovBi-Caps	The results exhibited that the precision of developed algorithm was 90%, recall of 80% and F-Score of 84% on latter dataset. Moreover, the supremacy of the developed algorithm was proved over the existing techniques.	This algorithm had not considered diverse contextual semantic to detect the hate content and unable to utilizeprofile-related attributes of user.
Y. Zhou, et.al	2020	DLF (Deep Learning Based Fusion) method	The outcomes revealed that the presented method assisted in enhancing the accuracy and F1-score for classifying the hate speech. Additionally, the practicality of the presented method was proved at lower cost.	The degree of integration was not deep because it was done when the data was classified.
S. Khan, et.al	2022	BiCHAT	The recommended algorithm led to enhance the baseline techniques concerning precision up to 8%, recall by 7% and f-score up to 8%. Moreover, this algorithm improved the accuracy around 0.5 to train the data and 0.9 to validate the data.	This algorithm was ineffective of classifying the multi-lingual and code-mixed hate content.
Z. Mossie, et.al	2019	HSD (hate speech detection) method	The experimental results confirmed that the projected method outperformed the traditional models to detect the hate speech, and effective for recognizing Tigre ethnic group as susceptible community concerning hatred.	This method was not suitable for handling negation and utilizing the information throughout the posts and comments.
G. d. Valle-Cano, et.al	2022	Social Hater BERT	According to the experimental results, the formulated algorithm had generated the optimal results for detecting the hate speech.	This approach did not take history and evolution of review of hate, trends, public and anonymous users who were affected due to it.
I. Z. Muhammad, et.al	2020	DBN (Deep Belief Network) technique	The analysis results validated that the accuracy of the developed technique was calculated 86%, precision was 82%, recall was 89.13% and Fl-Score was found	The major limitation of this technique was that it had not detecting the hate speech on the basis of wide topics including religion, race, etc.

			85.42%.	
Rahul, et.al	2021	An independent and self-sufficing framework	The experimental results indicated the robustness and applicability of the designed framework to learn complex dependencies while detecting the hate speech in comparison with the other methods.	This work made the employment of small corpus of tweets in Hinglish.
J. Melton, et.al	2020	Ensemble of DL (deep learning) algorithms	The projected approach offered a recall of 83% on HON dataset in contrast to the conventional methods. In addition, when the designed method was trained with classifier, the efficiency of the suggested model was maximized on unlabeled data from Gab and yielded the recall of 67%.	The additional experimentation was not possible in this work for optimizing the metrics.

III.CONCLUSION

Hate speech detectors based on deep neural networks are referred to as deep learning methods. Any feature encoding technique, including established ones like TF-IDF and more recent ones like word embedding or pre-training techniques, may be used to encode the input data for these neural networks. The latter strategy, which helps avoid conventional feature engineering or feature construction procedures, is typically more effective than the previous way. The deep learning models can be improved in future which leads to increase in accuracy for the hate speech detection.

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] D. R. Beddiar, M. S. Jahan and M. Oussalah, "Data expansion using back translation and paraphrasing for hate speech detection", Online Social Networks and Media, vol. 10, pp. 133-141, 23 June 2021

[15] C. -C. Wang, M. -Y. Day and C. -L. Wu, "Political Hate Speech Detection and Lexicon Building: A Study in Taiwan," in IEEE Access, vol. 10, pp. 44337-44346, 2022
[16] S. Khan et al., "HCovBi-Caps: Hate Speech Detection Using Convolutional and Bi-Directional Gated Recurrent Unit With Capsule Network," in IEEE Access, vol. 10, pp. 7881-7894, 2022
[17] Y. Zhou, Y. Yang, H. Liu, X. Liu and N. Savage, "Deep Learning Based Fusion Approach for Hate Speech Detection," in IEEE Access, vol. 8, pp. 128923-128929, 2020
[18] S. Khan, M. Fazil and A. R. Baig, "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection", Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 7, pp. 4335-4344, 21 May 2022
[19] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media", Information Processing & Management, vol. 7, no. 4, pp. 170003-170011, 23 July 2019
[20] G. d. Valle-Cano, L. Quijano-Sánchez and J. Gómez, "SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles", Expert Systems with Applications, vol. 207, no. 12, pp. 769-778, 20 December 2022
[21] I. Z. Muhammad, M. Nasrun and C. Setianingsih, "Hate Speech Detection using Global Vector and Deep Belief Network Algorithm," 2020 1st International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, 2020, pp. 1-6
[22] 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
[23] J. Melton, A. Bagavathi and S. Krishnan, "DeL-haTE: A Deep Learning Tunable Ensemble for Hate Speech Detection," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2020, pp. 1015-1022