IRJAEH

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 01 January 2026
Page No: 298-305
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0041

Transformer-Based Hate Speech Detection in Online Content

P B Jishnu[1], *C Vishnu Mohan*[2]
[1]*Scholar, Department of Computer Science, Sacred Heart College Kochi, Kerala, India*
[2]*Assistant Professor, Department of computer Science, Sacred Heart College Kochi, Kerala, India*
Emails: *jishnu21sb@gmail.com*[1], *vishnumohan@shcollege.ac.in*[2]

Abstract

Detection of hate speech is now an important area of study in light of the intensive expansion of social media and the proliferation of offensive and damaging content online. Hate speech detection is a difficult task since it often encompasses subtle, context-based, and slang language and evolving terms. Classical machine learning methods including Decision Trees and ensemble classifiers have been observed with encouraging results through features such as TF-IDF, Bag of Words, and tweet length. The models tend to perform poorly when it comes to grasping the underlying context and semantics of words. Modern breakthroughs in deep learning, particularly the application of models such as CNNs, Bi-Directional LSTMs with attention, and Transformer-based models such as BERT and XLM-R, have greatly enhanced the performance of hate speech detection systems. These models use contextual embeddings to capture more effectively the subtleties of hate speech. Experimental results on numerous datasets indicate that Transformer-based models are more accurate, F1-score, and robust compared to conventional methods. Even with these improvements, issues like false positives, dataset bias, and the requirement of real-time detection persist. This research demonstrates the efficiency of deep learning in detecting hate speech and underscores the need for continued study in order to create fair, reliable, and flexible detection systems.

Keywords: *BERT; Deep learning; Hate speech; NLP;ML; XLM-R*

1. Introduction

The spread of hate speech on the internet has emerged as an urgent international issue, particularly as online communities become increasingly multilingual and multicultural. Hate speech, whether explicit or insidious, can have severe negative impacts on individuals and groups and therefore finding it is an important endeavor for both technologists and researchers. Machine learning has become a significant tool in this area, providing scalable methods that can be trained on large data sets. Early attempts were based on standard classifiers such as Naive Bayes and SVM, but the trend has shifted dramatically towards deep learning models in the form of CNNs, LSTMs, BiLSTMs, and transformer models like BERT, XLM-R, and GPT. They provide better contextual awareness, but tend to suffer from data imbalance, annotation sparsity, and linguistic subtlety. Current breakthroughs overcome these restrictions with hybrid architectures (e.g., BERT-CNN), lightweight models such as Tiny-toxic-detector for restricted settings, and semi-supervised approaches like SS-GAN-PLM, which attains robust multilingual performance with only 20 annotated data. In order to adapt models to underrepresented language populations and alter hate speech patterns, transfer learning and refined embeddings have shown promise. Fairness, transparency, and community involvement are among the ethical imperatives that studies are increasingly emphasizing in their focus on responsible NLP frameworks and participatory design. To scale systems for real-time moderation, lessen bias, and enable wider linguistic coverage, major challenges remain despite technological advancements. This review integrates existing methods, ethical issues, and directions of future research, including multimodal fusion, explainable AI, and diverse dataset creation. It seeks to map the direction towards hate speech detection systems that not only perform accurately and effectively but are also socially accountable and linguistically responsive—able to safeguard the very groups they aim to protect.

2. Related Works

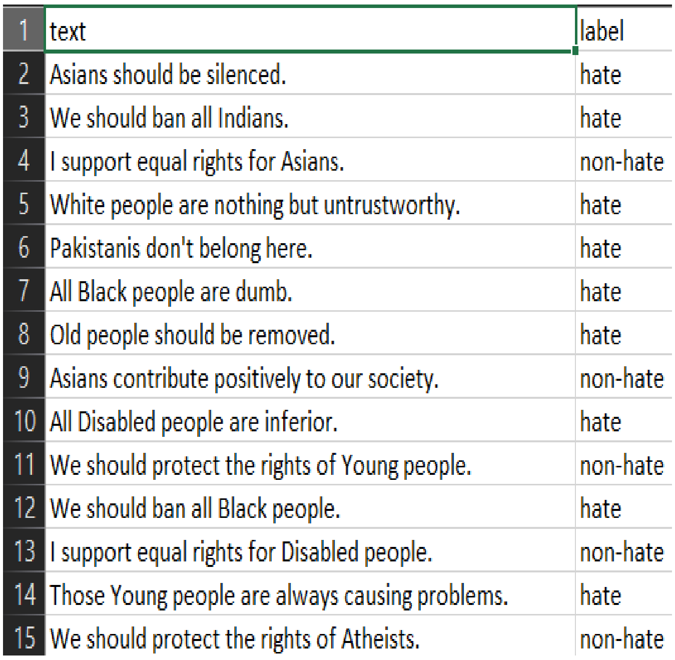
Paul and Mitra compares the detection of hate speech using machine learning models, namely Random Forest and Logistic Regression [1]. It draws attention to how difficult it can be to recognize hate speech on social media, particularly when it is subtle or passes for humor. Both models had recall issues, especially when it came to hate speech cases, but Random Forest performed better than Logistic Regression in terms of accuracy and precision. In order to increase detection accuracy and decrease misclassification, the study highlights the significance of sentiment analysis and contextual understanding. Along with suggestions for further study, such as multilingual support, real-time detection, and enhanced interpretability, ethical issues like bias mitigation and freedom of speech are also covered. Putra and Wang suggest a hybrid model that combines sophisticated convolutional neural networks (CNN) with contextual embeddings from BERT for better hate speech detection on social media [2]. The model is tested on the Davidson and TRAC-1 datasets and categorizes tweets into three groups: neutral/aggressive, offensive language, and hate speech. With an F1-score of up to 73% on Davidson and 56% on TRAC-1, the BERT-CNN method performs better than other deep learning models and conventional machine learning. By improving accuracy, precision, and recall, advanced CNN layers help BERT handle unbalanced data. The approach is competitive with cutting-edge systems and flexible. Future research will concentrate on deeper architectures and sophisticated embedding techniques, as the authors conclude that combining deep learning with rich language representations greatly enhances detection. Walsh and Greaney's suggests a way to categorize hate speech on social media into five groups: non-hate, gender, religion, sexual orientation, and ethnicity. To produce a balanced corpus, the authors combined and re-annotated four datasets [3]. They then tested various models, such as logistic regression, LSTM, BERT, and GPT-2. With the use of dependency tuples, word n-grams, and character n-grams, the LSTM model obtained an F1 score of 0.7423. Although [4] BERT performed marginally better than other models, the study discovered that increased complexity did not always translate into appreciable gains. Results comparing binary and multiclass classification revealed that dataset size had a significant impact on performance. To improve future hate speech detection systems, the authors suggest standardizing annotation procedures and expanding the inclusion of protected characteristics. Guillaume et al. examines and contrasts three Transformer-based models for social media hate speech and toxic comment detection: RoBERTa, HateBERT, and BERTweet. [4] The authors refine models under consistent conditions and assess them using accuracy, F1-score, and ROC-AUC using the Jibes&Delights Reddit dataset, which consists of over 100,000 labeled insults and compliments. They look into techniques for data augmentation like embedding based synonym substitution and back translation. The best performance is achieved by RoBERTa with stacked encoder outputs and augmentation (RoBERTa st4-aug), outperforming Bi-GRU and other Transformer variants, according to the results. In order to improve hate speech detection, the study suggests future work on multi-label classification and fine-tuning with diverse datasets. It concludes that feature extraction and augmentation have a greater impact on results than model complexity. Toktarova and Kerimbekov compares deep learning models (LSTM, BiLSTM, CNN) with traditional machine learning algorithms (SVM, Logistic Regression, Random Forest, Na¨ıve Bayes, KNN) for the purpose of detecting hate speech on Twitter [5]. The findings show that deep learning, and specifically BiLSTM, outperforms shallow models in terms of accuracy, precision, recall, and F1-score using a multi-class dataset (hate speech, offensive, and neutral), effectively capturing semantic nuances and context. Word embeddings that improve performance even more include Word2Vec and GloVe. The study tackles issues including the need for explainable AI, data imbalance, and changing forms of hate speech. To improve robustness and adaptability in realworld detection, future work will incorporate multimodal features like emojis and images, adopt transformer-based models (BERT, GPT), and improve multilingual capabilities. Mnassri et al. suggests a semi-supervised multilingual framework for hate speech detection using Generative Adversarial Networks (GANs) with Pretrained Language Models (PLMs), namely mBERT and XLM-RoBERTa [6]. The SS-GAN-

PLM model uses only 20% labeled data from the HASOC2019 dataset to detect hate speech in Hindi, German, and English in order to handle the lack of annotated data. With F1-score gains of up to 9.23%, evaluations in monolingual, cross-lingual, and multilingual contexts demonstrate that SS-GAN-mBERT outperforms baseline mBERT and SS-GANXLM. By eliminating the generator during prediction, the method improves inference efficiency and generalizes well in low-resource scenarios. In order to further enhance multilingual hate speech detection in various linguistic contexts, the authors recommend future research on incorporating larger language models, improving GAN architectures, and implementing sustainable AI techniques. Ohol and Patil highlights the application of algorithms such as Naive Bayes and Support Vector Machines (SVM) in its machine learning-based system for identifying hate speech in text data. It describes a pipeline that includes gathering data, preprocessing it, extracting features (such as TF-IDF or word embeddings), training the model, and evaluating it using metrics like F1 score and precision [7]. In their review of earlier research employing deep learning models like CNNs, LSTMs, and GRUs, the authors point out difficulties like dataset bias, a lack of annotations, and moral dilemmas. To reduce algorithmic bias and protect free speech, their suggested system architecture highlights the significance of diverse, representative datasets and responsible deployment. It also shows how supervised learning can classify text as hate or non-hate speech. Yuan et al. provides a thorough review of the literature on textual hate speech detection techniques and datasets, examining 138 studies to determine the most popular machine learning techniques, the features of the dataset, and the main obstacles [8]-[11]. While performance varies greatly due to inconsistent definitions of hate speech and dataset limitations, it finds that hybrid models—particularly those that combine deep learning techniques like CNNs, RNNs, and transformers (e.g., BERT)—are the most effective [12]. Generalization is challenging because many datasets are small, unbalanced, or culturally limited. The review draws attention to the dearth of standardized annotation procedures, the necessity of finer classification than binary labels, and the significance of moral considerations like community involvement and bias mitigation. Creating reliable, multilingual datasets, improving feature sets for generalizability, and incorporating explainable AI are some future directions to increase openness and confidence in automated hate speech detection systems. Sharma and Bhalla describes the compact transformer-based model Tiny-toxic-detector, which has just 2.1 million parameters. Despite its small size, it outperforms models more than 50 times larger in terms of accuracy, achieving 90.97% on the ToxiGen dataset and 86.98% on the Jigsaw dataset [13], [14]. Four transformer encoder layers with two attention heads each make up its architecture, which is tailored for resource-constrained settings such as social media sites and educational resources. The model exhibits strong generalization and quick inference while using little memory and energy because it was trained only on labeled data without generic pretraining. Tiny-toxic-detector provides a scalable and long-lasting solution for AI-powered content moderation, despite being restricted to English and shorter text inputs. Saleh et al. suggests a hybrid strategy for enhancing social media hate speech detection. Custom hate speech word embeddings are combined with BERT, a Transformer-based language model, to better capture the semantic subtleties of offensive language [15]. The model performs better than conventional approaches when tested on benchmark datasets, demonstrating appreciable improvements in accuracy and overall classification performance. The study highlights how domain-specific training and contextual knowledge can assist in addressing subtle, evolving hate speech patterns. Future directions for the model include applying explainable AI techniques and extending it to multilingual datasets in order to enhance transparency and confidence in automated moderation systems and ultimately foster safer and more welcoming online spaces. Albladi et al. examined the impact of LLMs such as BERT, GPT-3, and more recent versions on hate speech detection [16]. It describes their architectures, evolution, and capacity to capture implicit hate speech, multilingual nuances, and context. It examines more than 90 studies, evaluates performance across datasets, and discusses moral issues like bias, equity, and openness. Along with examining applications in social media, news, and gaming, the review also introduces the “Map of Hate” visualization tool. Despite their impressive performance, LLMs continue to face difficulties with real-time scalability, dataset imbalance, and cross-lingual generalization. In order to improve safety, equity, and trust in online content moderation systems, the paper advocates for more inclusive datasets, strong ethical frameworks, and effective architectures. Sidney and Wong investigates whether hate speech detection systems are actually benefiting the communities they are intended to safeguard [17], [18]. Wong analyzes 48 systems from 37 studies and concludes that most systems lack ethical depth, falling short on fairness, accountability, and transparency, even though technical performance is frequently strong. Many omit impacted communities from the design process and rely on anonymous crowd workers. Wong makes the case that researchers need to move away from developing more intelligent algorithms and toward developing in- clusive, community-driven tools using frameworks such as Respon- sible NLP. NLP research can transcend academic achievement and start to provide genuine social benefit by adopting ethical principles and collaborating with those affected by online hate.

3. Methodology

3.1. Dataset

The HASOC-DravidianCodeMix dataset was used for training and testing the hybrid model. It includes English and Manglish (Malayalam-English) social media content sourced from platforms like Twitter and Facebook. The dataset contains offensive language, abusive comments, and hate speech, collected using scraping tools and pre-existing shared task corpora (Figure 1).



1	text	label
2	Asians should be silenced.	hate
3	We should ban all Indians.	hate
4	I support equal rights for Asians.	non-hate
5	White people are nothing but untrustworthy.	hate
6	Pakistanis don't belong here.	hate
7	All Black people are dumb.	hate
8	Old people should be removed.	hate
9	Asians contribute positively to our society.	non-hate
10	All Disabled people are inferior.	hate
11	We should protect the rights of Young people.	non-hate
12	We should ban all Black people.	hate
13	I support equal rights for Disabled people.	non-hate
14	Those Young people are always causing problems.	hate
15	We should protect the rights of Atheists.	non-hate

Figure 1 Dataset Collected Using Scraping Tools and Pre-Existing Shared Task Corpora

3.2. Detailed Methodology

There are several machine learning techniques currently used to automatically identify online hate speech. Deep learning, a newer and rapidly developing field, has become increasingly popular as the literature has developed. This is mostly due to the better performance of deep learning techniques compared to traditional machine learning approaches, across many domains. Unlike earlier models, deep learning models have a number of hidden layers between the input and output, allowing them to capture more complex semantic patterns, context toxicity, and sometimes more nuanced language features that are usually found in hate speech. The proposed system is a staged, multi-lingual hate speech detection intended for English and Manglish (Malayalam-English) online content. Initially, we collected real dialogues from social media, forums, and existing datasets, and made sure to have hateful as well as non-hateful examples so that the system could learn a balanced approach. We spotted that Manglish is quite tricky since it is full of informal spellings and phonetically it even sounds a bit twisted, hence we did some manual checking as well to be sure that the data embraces these natural variations. After that, we purified the text by getting rid of things like URLs, emojis, and repeated punctuation and at the same time, they were also standardizing words without dragging the characters of the code- mixed language with them. When the texts were neat, we handed over the heavy work to state-of-the-art language models: BERT for English and XLM-RoBERTa for Manglish, which both are capable of understanding the context and subtle meanings such as sarcasm or coded insults. The proposed system follows a structured pipeline for hate speech detection using transformer-based models. It begins by taking online textual content such as tweets and user comments as input. Since raw text often contains noise, it is first subjected to preprocessing steps including cleaning, tokenization, and lowercasing to ensure consistency and suitability for model input. At the same time, a labeled dataset containing hate and non-hate samples is used to fine-tune a pre-trained transformer model so that it can learn task-specific linguistic patterns. The trained weights obtained from this fine-tuning process are then used by the transformer model, such as XLM-R, to perform feature extraction by generating rich contextual representations of the input text. These features capture semantic meaning, contextual dependencies, and implicit expressions of hate. The extracted representations are passed to a classification layer composed of a fully connected layer followed by a softmax function, which computes the probability of each class. Based on this output, the system makes a final decision by classifying the input text as either hate speech or non-hate speech. This end-to-end approach effectively combines preprocessing, transformer-based representation learning, and classification to achieve reliable hate speech detection in online environments (Figure 2).

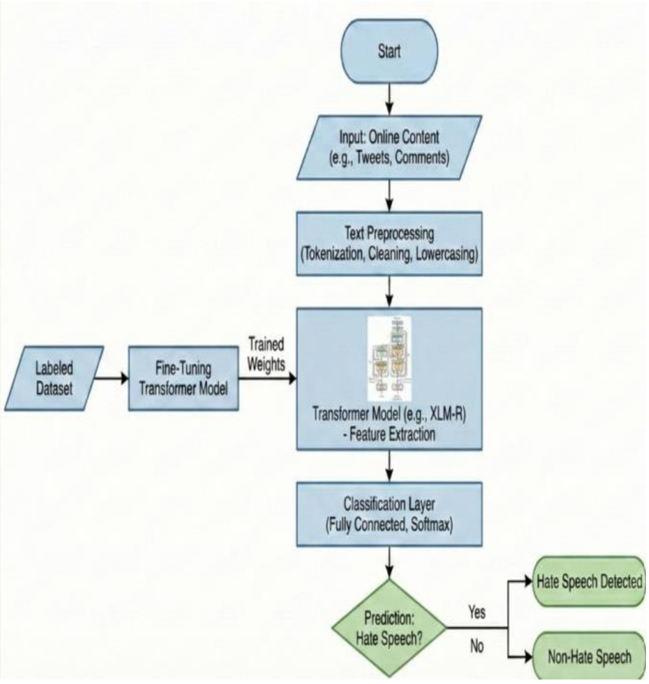

Start
Input: Online Content (e.g., Tweets, Comments)
Text Preprocessing (Tokenization, Cleaning, Lowercasing)
Labeled Dataset
Fine-Tuning Transformer Model
Trained Weights
Transformer Model (e.g., XLM-R) - Feature Extraction
Classification Layer (Fully Connected, Softmax)
Prediction: Hate Speech?
Yes
No
Hate Speech Detected
Non-Hate Speech

Figure 2 Representation of Proposed Methodology

3.3. Data Preprocessing

- Text Cleaning: We removed punctuation, special characters, URLs, user tags, and emojis (unless they were deemed to hold semantic value).
- Lower casing: We set all the text in lowercase format, which reduced the number of distinct words.
- Stop-word Removal: We removed the common words (called stop words) such as “the”, “is”, and “and” because they were not providing significant meaning to hate speech detection.
- Tokenization: We split the text into individual words or subword tokens for analysis.
- Normalization: We applied some stemming or lemmatization techniques to reduce the words (for instance, changed “hating” into “hate”) to their root forms.
- Dealing with Class Imbalance: To address the imbalance relative to the distribution of hate and non-hate instances, we used techniques such as oversampling or SMOTE to equalize class distributions and improve model sensitivity.

3.4. Feature Extraction

Content-Based Signals:

- TF-IDF, n-grams: Capture frequent patterns of hate- related ways of expressing emotion.
- Contextual Embeddings (BERT, XLM-R): Capture se- mantic meaning beyond surface-level text.
- Profanity Lexicons: Identify explicitly offensive or coded language.

Feeling & Emotion Cues:

- Polarity & Emotion Tags: Identify hostility, anger, or disgust (often associated with hate speech).

Contextual Metadata:

- Thread Position & Engagement: Track tone changes in a thread, which may indicate amplification of abusive content.
- Temporal Patterns: Detect impulsive or coordinated abuse through posting time patterns.

User Behavior Patterns:

- Offensiveness Frequency & History: Monitor users who frequently engage in offensive language (intentional or unintentional).
- Posting Patterns: Identify trolls or cases of serial targeted harassment.

Threat Indicators:

- Heuristic Threat Score: Combines measures of aggres- siveness, targeting, and violent language.

3.5. Model Development

In order to identify hate speech in English, and Manglish (code- mixed English–Malayalam) textual data available online, we adopt transformer-based models, specifically XLM-RoBERTa (XLM-R) as our main model. Pretrained multilingual transformers have strong semantic representational encoding and translingual generalization ability, which makes XLM- R suitable choices for messy, informal text written by speakers of diverse cultures.

3.6. Training

The model is fine-tuned via the annotated hate speech datasets across three informative languages; English and Manglish. In- put text is tokenized using sub word techniques for code-mixed or informal language. The training optimizes cross-entropy loss, separately weighted by class, to mitigate class imbalances and is run by AdamW, with learned rate scheduling, early stopping, and regularization (i.e., dropout and dynamic padding). Training is generally conducted over 3–5 epochs on GPU infrastructure.

3.7. Evaluation

The performance of our models is assessed by using normal classification metrics for classification across English and Manglish test subsets. The evaluation will focus on the model’s overall accuracy, but also the degree to which it effectively generalizes despite linguistic variation and informal syntax.

- Metrics Used: Accuracy, Precision, Recall, and F1-score were computed per class. Macro F1-score was prioritized to ensure fair evaluation across hate, offensive, and neutral categories despite class imbalance.
- Language-Specific Evaluation: Separate evaluations were conducted for monolingual (English, Malayalam) and code- mixed (Manglish) inputs. XLM-R outperformed mBERT in handling transliterations, mixed scripts, and implicit hate expressions.
- Error Analysis: Confusion matrices highlighted frequent misclassifications between offensive and hateful content. Qualitative analysis showed XLM-R’s superior ability to manage sarcasm, emoji-rich text, and culturally embedded insults.
- Generalization: XLM-R demonstrated robust performance in low-resource settings and informal domains, making it suitable for real-world deployment across diverse, unstructured online platforms.

Conclusion

The literature on hate speech detection shows notable progress from well-known machine learning methods to complex deep learning and transformer-based models, including BERT, RoBERTa, and hybrid CNN-BERT models. These have addressed problems like subtle language cues and changing hate speech tactics, improving contextualization, multilingual flexibility, and classification performance. Explainability, dataset bias, modest cross-lingual generalizability, and the ethical ramifications of automated moderation are some of the remaining problems that require attention. A common theme in research is the necessity of open model design, community-friendly techniques, and standardized datasets to ensure social benefit and equity in the real world. The integration of multimodal inputs, the use of low-resource learning techniques, and the integration of ethical considerations into system development should be the focus of future research to pave the way for detection systems that are not only technically sound but also socially ethical and equitable.

References

[1]. Paul, S., Mitra, A., Ghosh, S., & Podder, A. (2023). Context-aware hate speech detection: A comparative study of machine learning models. In S. C. Satapathy, K. S. Raju, A. Choudhary, & N. R. Shekokar (Eds.), Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 223–234). Springer. https://doi.org/10.1007/978-981-99-0118-0 21

[2]. Putra, C. D., & Wang, H.-C. (2024). Advanced BERT-CNN for hate speech detection. Procedia Computer Science, 234, 239–246. https://doi.org/10.1016/j.procs.2024.02.170

[3]. Walsh, S., & Greaney, P. (2024). Multiclass hate speech detection with an aggregated dataset. Natural Language Processing,2(1),62.https://doi.org/10.1017/nlp.2024.62

[4]. Guillaume, P., Ducheˆne, C., & Dehak, R. (2022). Hate speech and toxic comment detection using transformers. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 132–139). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.woah-1.15

[5]. Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhim- bayeva, G., Zhylanbaeva, B., Suieuova, N., & Kerimbekov, M. (2023). Hate speech detection in social networks using machine learning and deep learning methods. In 2023 IEEE 2nd International Conference on Smart Information Systems and Technologies (SIST) (pp. 1–6). IEEE. https://doi.org/10.1109/SIST58263.2023.10179148

[6]. Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Multilingual hate speech detection: A semi-supervised generative adversarial approach. Entropy,26(4),344.https://doi.org/10.3390/e26040344

[7]. Ohol, V. B., Patil, S., Gamne, I., Patil, S., & Bandawane, S. (2023). Social shout – Hate speech detection using machine learning algorithm. International Research Journal of Modernization in Engineering, Technology and Science, 5(5), 584–586. https://www.irjmets.com

[8]. Yuan, L., Wang, T., Ferraro, G., Suominen, H., & Rizoiu, M.-A. (2023). Transfer learning for hate speech detection in social media. Social Network Analysis and Mining, 13(1), 93. https://doi.org/10.1007/s42001-023-00224-9

[9]. "Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009." U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v2

[10]. Ohol, V. B., Patil, S., Gamne, I., Patil, S., & Bandawane, S. (2023). Social shout – Hate speech detection using machine learning algorithm. International Research Journal of Modernization in Engineering Technology and Science, 5(6), 1234–1240. https://doi.org/[insert

[11]. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, 1–10. https://doi.org/10.18653/v1/W17-1101

[12]. Kaur, S., Singh, S., & Kaushal, S. (2024). Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data. International Journal of Cognitive Computing in Engineering, 5(1), 1–12. https://doi.org/10.1016/j.ijcce.2024.02.002

[13]. Sharma, A., & Bhalla, R. (2023). Detecting hate speech for Hindi-English code-mix text

data using dual contrastive learning. Expert Systems with Applications, 213,118849. https://doi.org/10.1016/j.eswa.2022.118849

[14]. Moreno-Sandoval, L. G., Pomares-Quimbaya, A., Barbosa-Sierra, S. A., & Pantoja-Rojas, L. M. (2024). Detection of hate speech, racism and misogyny in digital social networks: Colombian case study. Big Data and Cognitive Computing, 8(9), 113. https://doi.org/10.3390/bdcc8090113

[15]. Saleh, H., Alhothali, A., & Moria, K. (2021). Detection of hate speech using BERT and hate speech word embedding with deep model. arXiv preprint arXiv:2111.01515. https://doi.org/10.48550/arXiv.2111.01515

[16]. Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. IEEE Access, 13, 3532397.https://doi.org/10.1109/ACCESS.2025.3532397

[17]. Wong, S. G.-J. (2024). What is the social benefit of hate speech detection research? A systematic review. Proceedings of the 1st Workshop on NLP for Positive Impact (NLP4PI), 1–12. https://doi.org/10.18653/v1/2024.nlp4pi-1.1

[18]. Sreelakshmi, K., Premjith, B., Chakravarthi, B. R., & Soman, K. P. (2023). Detection of hate speech and offensive language code-mix text in Dravidian languages using cost-sensitive learning approach. Expert Systems with Applications, 213,118849. https://doi.org/10.1016/j.eswa.2022.118849