Multimodal Deep Learning for Hate Speech Detection

Sai Aakash Vydana
Dept. of Computer Science
University at Buffalo
Buffalo, United States of America
saiaakas@buffalo.edu

Advised by: Dr. Shamsad Parvin
Dept. of Computer Science
University at Buffalo
Buffalo, United States of America
shamsadp@buffalo.edu

***Abstract*—With the rise of social media, people are sharing content in many different forms like text, audio, and video, at an unprecedented scale. Although this has allowed greater expression and connection, it has also led to an increase in harmful and hateful content online. Most existing research focuses on hate speech detection in text, overlooking the rich and complex signals embedded in videos. In this project, we present a multimodal deep learning approach to detect hate speech in videos by combining visual, audio, and textual features. Using the HateMM dataset, we extract features from video frames using a Vision Transformer (ViT), audio using MFCCs, and spoken content using automatic speech recognition followed by BERT embeddings. A key strength of our model is its flexibility, It is designed to operate not only with all three modalities, but also in scenarios where one or more modalities are missing. Whether only text, only audio, or partial combinations are available, the system still delivers strong performance. Our evaluation with 5-fold cross-validation shows that this adaptable, fused architecture outperforms unimodal baselines and represents a step toward robust and scalable hate speech detection across different types of media content.**

***Keywords*—Hate speech detection, Multimodal deep learning, Video classification, Audio features, Text embeddings, ViT, BERT, MFCC, Social media moderation, Content safety**

I. Introduction

Nowadays, social media platforms have evolved into powerful communication channels where users express opinions through a variety of formats through text, images, audio, and video. This explosion of content brings with it the increased risk of online hate speech dissemination. Although automated detection of hate speech in text has received substantial attention in recent years, hateful content in video, where abusive cues may lie in tone, visuals, or transcribed speech, remains an underexplored but critical challenge [15] [12].

Traditional hate speech detection systems primarily focus on textual data, which often misses important clues like the tone someone uses when speaking, or aggressive expressions in a video. This limitation makes such models ineffective for platforms like BitChute or YouTube, where hate is often masked in sarcasm, emotionally charged speech, or visual symbols [5] [12] [14]. Moreover, real-world video data often suffers from missing or noisy data, such as failed speech transcription or poor frame quality—posing additional challenges that many traditional models are not designed to handle [1] [8].

A major challenge in multimodal hate speech detection is the lack of large-scale high-quality datasets, especially those focused on video content. Although many existing datasets target textual hate speech, few include the synchronized audio and visual elements needed to understand how hate is conveyed in real-world multimedia posts [1] [4] [14]. This has limited the development and evaluation of models that go beyond simple text classification. In this context, the HateMM dataset introduced by Das et al. [2] provides a valuable foundation. It includes diverse and annotated video content collected from BitChute, enabling research on how hate manifests itself in visual, audio, and text modalities. Building on this resource, we propose a deep learning-based model that integrates features from all three modalities, visual, audio, and text, to detect hate speech more effectively.

Previous work has shown the value of incorporating multiple modalities into hate speech detection and sentiment analysis tasks [3] [4] [6], but few models explicitly support flexible modality configurations. To overcome these issues, we adopt a modular architecture that builds separate unimodal branches for text, audio, and visual inputs, along with a fusion mechanism that can gracefully handle missing modalities during inference. This design results in several key contributions:

- A flexible deep learning model that can work with any available combination of text, audio, or video, making it practical and reliable even when some parts of the input are missing or noisy.
- A fusion strategy that improves reliability across diverse content types and enhances moderation capabilities at scale.
- Strong performance gains over unimodal baselines, validated through comprehensive 5-fold cross-validation on a real-world dataset.

Together, these contributions help pave the way for more inclusive, scalable, and context-aware hate speech detection systems.

II. Related work

Hate speech detection has been a widely studied area in natural language processing (NLP), especially for platforms like Twitter and Facebook. Early methods relied on traditional machine learning techniques using manually crafted features from text data [14] [15]. More recently, deep learning models such as CNNs, RNNs, and transformer-based architectures like BERT have demonstrated strong performance on various hate speech benchmarks [5] [6] [11].

While these models perform well on text data, they fall short in handling scenarios where hate speech can also be conveyed through tone, facial expressions, or visual context. Recognizing this gap, researchers have begun to explore multimodal detection approaches. Boishakhi et al. [3] and Khera et al. [4] proposed models that combine audio and text, or visual and text cues, to detect hate speech more accurately. However, these approaches often assume all modalities are available during inference, which limits their practical application due to the noisy data in a real-world scenario.

The HateMM dataset introduced by Das et al. [2] is among the first to provide a well-annotated, real-world video dataset for hate speech detection with multimodal. Their baseline models showed improved performance using a combination of modalities, thus confirming the benefit of multimodal learning in this space.

Further studies like Malik et al. [8] and Mnassri et al. [9] benchmark deep learning architectures across multiple hate speech datasets and show how pre-trained models and ensemble strategies can further boost performance. Fonseca et al. [10] delve into hate speech dynamics on Twitter, suggesting that user interactions and conversational threads also contribute to effective detection.

While these contributions advance the field, few works directly address the issue of modality dropout or flexible input fusion, which is crucial for scalable deployment. Our model closes this gap by allowing inference with any combination of available modalities like text, audio, or video, making it highly practical for real-world scenarios.

III. Dataset

For this study, we use the HateMM dataset introduced by Das et al. [2], a multimodal benchmark specifically curated for hate speech detection in online videos. The dataset consists of approximately 1,083 videos (around 43 hours) collected from BitChute, a video-sharing platform known for hosting controversial and minimally moderated content. Each video is manually annotated as either Hate or Non-Hate. For hate-labeled videos, annotators have also marked time-stamped segments where hate speech occurs and identified the target group or community. These target groups include Blacks, Jews, Whites, Asians, LGBTQ individuals, and others.

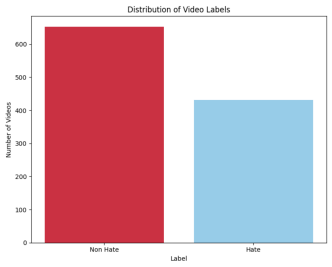

Fig. 1. Distribution of Video Labels

To better understand the composition of the dataset, Figure 1 presents the distribution of hate and non-hate videos. While there is a moderate imbalance with non-hate videos forming the majority, this distribution is beneficial for practical applications. Identifying non-hate content often requires more nuanced judgment, as it involves verifying context and intent, whereas hate speech can sometimes be easily recognized by the presence of explicit slurs or aggressive language. Therefore, having a larger proportion of non-hate examples not only reflects real-world data but also supports the development of models that are more cautious and precise in distinguishing between offensive and benign content.

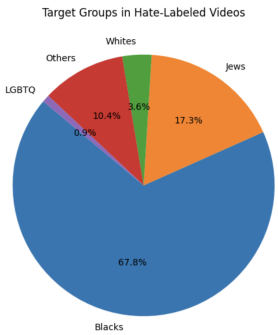

Fig. 2. Target Groups in Hate-Labeled Videos

Figure 2 provides insights into the demographic groups targeted in hate-labeled videos. The chart shows that hate speech is disproportionately directed at the Black community, which accounts for 67.8% of the hateful content. This is followed by content targeting Jews (17.3%) and others (10.4%), with smaller proportions aimed at Whites (3.6%) and LGBTQ individuals (0.9%).

IV. Methodology

Our multimodal hate speech detection model is designed to process and integrate information from three modalities: visual, audio, and textual. The architecture is built in a modular

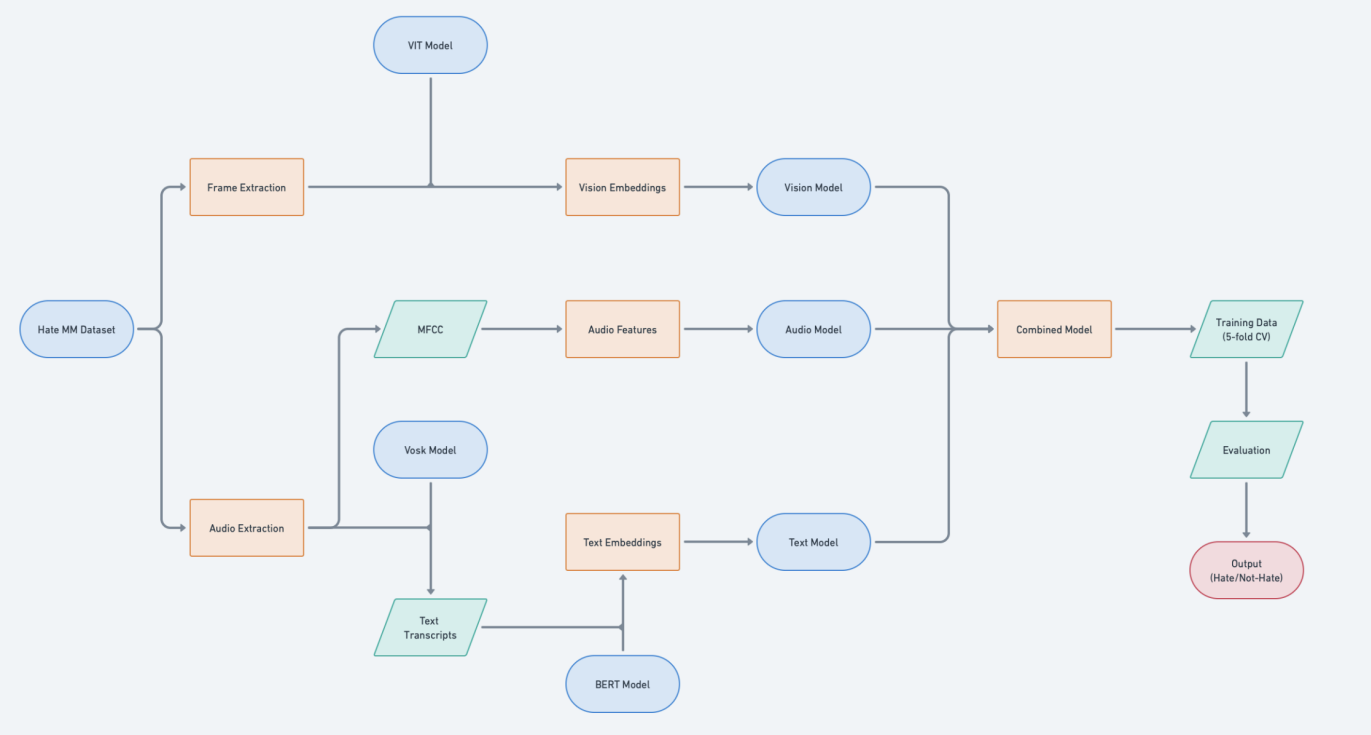

Fig. 3. The Workflow

way, where each modality is handled separately using its own feature extractor. Once the individual embeddings are generated, they are brought together through a fusion mechanism that combines them for the final classification. The model is designed to remain functional even when one or more modalities are missing or corrupted, which is often the case in real-world video data. An overview of the workflow is illustrated in Figure 3.

A. *Preprocessing*

Each video is first segmented into fixed-length clips of 10 seconds. For each clip, we perform initial preprocessing by extracting frames and audio:

- **Frame Extraction**: Using OpenCV, we extract 100 frames at a rate of 1 frame per second and store them as images.
- **Audio Extraction**: Using MoviePy, we extract the audio track from each video clip and save it as a WAV file.

These raw modality components are then processed through their respective feature extraction pipelines, as described below.

*1) **Text Embeddings**:* We convert the extracted audio into text using the Vosk automatic speech recognition (ASR) model. The transcriptions are cleaned by removing filler words, special characters, and punctuation. We tokenize the cleaned text using the BERT-base-uncased tokenizer and feed the tokens into a pre-trained BERT model.

The [CLS] token from the final layer is used as the semantic representation of the clip, resulting in a 768-dimensional text embedding. If no valid transcription is produced due to silence, poor audio quality, or ASR failure, we replace the text embedding with a zero vector of size 768, and the model receives a flag indicating the modality is missing.

*2) **Audio Embeddings**:* We use the previously extracted audio tracks and compute Mel-Frequency Cepstral Coefficients (MFCCs) using the Librosa library. These features effectively capture speech characteristics such as tone, pitch, and vocal timbre.

For each clip, we extract 20 MFCC coefficients over time and aggregate them by computing the mean and standard deviation, resulting in a (20, 2) matrix per clip and then flattened into a 40-dimensional vector. These embeddings summarize the acoustic signature of the audio and help identify emotional cues commonly associated with hate speech, such as anger or sarcasm. If the audio is missing or fails to process, we use a zero vector of size 40 in place of the audio embedding, and the model is informed that the audio modality is absent

*3) **Vision Embeddings**:* For each video clip, we extract 100 frames at 1 frame per second using OpenCV. These frames are resized to 224×224 pixels and normalized with ImageNet statistics. Each frame is then passed through a pre-trained Vision Transformer (ViT-B/16) using Hugging Face's implementation, and we extract the [CLS] token

(768-dimensional) from the final layer for each frame.

This results in a sequence of 100 frame-level embeddings, each of size 768, which captures temporal visual context including gestures, facial expressions, and scene-level features. If fewer than 100 frames are successfully extracted, the missing embeddings are padded with zero vectors to maintain a consistent input shape of (100, 768). A modality flag is also set to indicate whether the visual input is complete or partially missing.

B. *Model Architecture*

Once we obtain the final embeddings from each modality they are passed through dedicated neural network modules tailored to transform them into a uniform representation space. These transformed embeddings are then fused and used for classification. Each sub-model is designed to be lightweight yet expressive, and the overall architecture supports flexible combinations of modalities during inference.

1) ***Text Model****:* The text model is a fully connected neural network that takes the 768-dimensional embedding produced by the BERT model as input. It consists of three linear layers with Layer Normalization, Leaky ReLU activation, and Dropout for regularization. The first layer reduces the dimensionality from 768 to a hidden size, followed by another reduction, and finally maps to a fixed 64-dimensional output embedding. This 64-dimensional representation is used in the final fusion step.

2) ***Audio Model****:* The audio model shares the exact same architecture as the text model. It takes the 40-dimensional MFCC-based audio embedding as input and processes it through the same sequence of linear layers, normalization, activations, and dropout. This design choice allows the model to treat audio and text features uniformly during fusion, simplifying integration while maintaining modality-specific learning.

3) ***Vision Model****:* The vision model is implemented using a bidirectional LSTM with attention. It takes the (100, 768) sequence of frame-level embeddings obtained from the Vision Transformer (ViT). The LSTM processes this temporal sequence and outputs contextualized hidden states for each frame. An attention mechanism is applied to weigh the importance of each frame, and a context vector is generated via a weighted sum. This vector is then passed through a two-layer feedforward network with Layer Normalization, ReLU, and Dropout, resulting in a 64-dimensional visual embedding.

4) ***Final Model****:* The final classification model, Combined_model, takes the 64-dimensional embeddings from the text, audio, and vision models and combines them. It also accepts flags indicating whether each modality is present. These flags are used to assign normalized weights to the modality outputs. Each embedding is scaled accordingly, and the three embeddings are concatenated into a 192-dimensional fused vector.

This fused representation is passed through a feedforward classification head consisting of Layer Normalization, a hidden layer with ReLU activation and Dropout, and a final linear layer that outputs class logits (e.g., Hate vs. Non-Hate). This design ensures the model remains functional and accurate even when one or more modalities are missing during inference.

V. Results

In this section, we present the training setup, evaluation metrics, and comparative performance analysis of our multimodal hate speech detection system. Our experiments are designed to assess not only the individual and combined performance of different modalities but also the model's robustness across varying data splits. To ensure reliability and generalizability, we employ 5-fold cross-validation and perform final evaluation using ensemble predictions based on majority voting.

A. *Training*

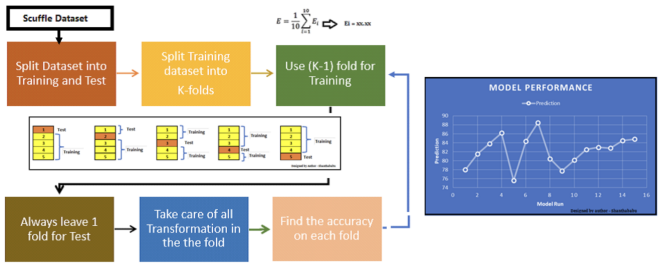

Fig. 4. 5-Fold Cross Validation techinque

To validate the performance of our model, we adopt a 5-fold cross-validation strategy using StratifiedKFold from scikit-learn [16]. As illustrated in Figure 4, the entire dataset is first merged and then split into five stratified folds. This ensures that each fold maintains the original class distribution, preserving the balance between hate and non-hate samples.

For each fold, we reserve a portion of the training data as a validation split (10% of the training fold), resulting in three subsets per fold: training, validation, and test. This procedure helps ensure that the model is evaluated under consistent and balanced conditions while making optimal use of the available data.

We train the model independently on each fold using a batch-based training loop and an Adam optimizer. For each epoch, the model is evaluated on both the validation and test

sets. The best performing model (in terms of macro F1-score on the validation set) is selected and its corresponding test set performance is saved.

After all five folds are completed, we save fold-wise predictions and metrics. These are later used for majority voting during ensemble evaluation. This cross-validation approach provides a more reliable estimate of model performance compared to a single train-test split. It helps reduce variance in evaluation, ensures better use of limited labeled data, and allows us to observe how consistent the model is across different data partitions.

B. *Evaluation and Comparison*

We evaluate our multimodal hate speech detection model using both 5-fold cross-validation and hold-out validation strategies. The model integrates features from text (BERT), audio (MFCC), and video (ViT) modalities and is assessed using commonly used performance metrics: accuracy, macro-F1 score, F1 score, precision, recall, and area under the curve (AUC).

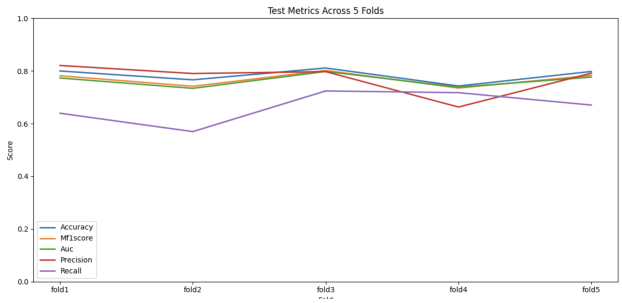

Fig. 5. Performance across 5-fold Cross Validation

Figure 5 shows the performance of our model across the five cross-validation folds. Each line in the graph represents a different evaluation metric: accuracy, macro-F1, AUC, precision, and recall. As shown, the metrics remain relatively consistent across folds, with minimal fluctuation. Accuracy and macro-F1 consistently remain above 75%, while AUC maintains stability close to 0.76. Although some variance is observed in precision and recall, especially between folds 2 and 4, the overall trend demonstrates the model's robustness across different data splits.

The recall scores are slightly lower and more variable compared to other metrics, which suggests the model is slightly conservative in flagging hate speech—favoring precision over false positives. Despite this, the high macro-F1 score indicates balanced performance across both hate and non-hate classes. Overall, these results confirm that our multimodal architecture performs reliably and generalizes well, even when evaluated across different test subsets.

*1) **Unimodal vs Multimodal Comparison**:* To further validate the strength of our multimodal fusion approach, we compare the performance of the final combined model with each of its unimodal components. Table I presents this comparison.

TABLE I
UNIMODAL VS MULTIMODAL PERFORMANCE

Model	Accuracy	Precision	Recall	F1 Score
Text (BERT)	79.13%	0.78	0.65	0.71
Audio (MFCC)	67.50%	0.73	0.65	0.62
Video (ViT)	75.60%	0.69	0.65	0.73
Multimodal (BERT + ViT + MFCC)	**83.37%**	**0.84**	**0.74**	**0.75**

As observed, the combined model clearly outperforms the individual modality models in every metric. Notably, it achieves a significantly higher precision of 0.84 and overall accuracy of 83.37%, indicating that integrating features from all three modalities enables the model to make more confident and correct predictions. This supports the conclusion that multimodal fusion leads to more context-aware and accurate hate speech detection.

*2) **Benchmark Comparison**:* To further validate the effectiveness of our approach, we compare our final model's performance against several baseline models reported in the HateMM dataset paper. These include both unimodal and multimodal configurations that use combinations of textual, audio, and visual features. Table II summarizes this comparison.

TABLE II
IMPROVEMENT OVER HATEMM BENCHMARK MODELS

Model Type	Benchmark Accuracy	Our Accuracy	Improvement
Text (fastText)	68.7%	79.13%	+10.43%
Text (BERT)	73.5%	79.13%	+5.63%
Audio (MFCC)	67.5%	67.50%	+0.00%
Vision (ViT)	74.8%	75.60%	+0.80%
Multimodal (BERT ViT MFCC)	79.8%	83.37%	+3.57%

As shown in Table II, our model consistently outperforms several key baselines reported in the HateMM benchmark. The most significant gain is observed over the text-only fastText model, where we achieve an improvement of over **10.43%** in accuracy. Our BERT-based text model also surpasses the benchmark BERT baseline by **5.63%**. For the vision modality (ViT), we improve slightly by **0.80%**, and match the audio (MFCC) performance exactly, confirming alignment in modality-specific feature extraction.

Most notably, our full multimodal system surpasses the strongest baseline (M1: BERT ViT MFCC) by **3.57% in accuracy** and **2.89 points in macro-F1 score**, underscoring the effectiveness of our fusion strategy. Unlike many benchmark models, our architecture is designed to flexibly accommodate missing modalities during inference, making it more robust and applicable to real-world scenarios where data incompleteness is common.

These improvements reflect the benefits of our modular design, which leverages specialized encoders for each modality, a bidirectional attention-based video model, and a

dynamic fusion mechanism. Together, they enable the system to integrate multimodal signals more effectively, leading to better generalization and more reliable hate speech detection in diverse multimedia content.

VI. Conclusion and Future Work

In this work, we presented a robust and flexible multimodal deep learning approach for hate speech detection in video content. By integrating features from text, audio, and vision using modality-specific encoders and a dynamic fusion strategy, our model outperforms several strong unimodal and multimodal baselines on the HateMM dataset. The system demonstrates consistent performance across 5-fold cross-validation and maintains effectiveness even under missing modality conditions, making it well-suited for real-world deployment.

Future Work: While our results are promising, there are several avenues for improvement. First, the HateMM dataset, despite being one of the few multimodal video-based hate speech benchmarks, is relatively limited in scale, with only 1,083 annotated videos. Expanding this dataset with a broader and more balanced representation of targeted communities would help improve generalizability and reduce dataset bias, especially given the current skew toward certain demographic groups.

Second, our current visual processing approach extracts uniformly sampled frames from each clip, which may not always capture the exact segments where hate speech is conveyed. A promising direction is to enhance the model to localize hateful segments more precisely within the video, essentially identifying the specific timestamps where hate occurs. This would not only improve classification accuracy but also offer significant practical value for real-time content moderation platforms such as YouTube, Instagram, or TikTok by allowing targeted review and removal of hateful content.

These extensions would further elevate the effectiveness and applicability of our system in building safer and more inclusive online platforms.

References

[1] C. J. Kennedy, G. Bacon, A. Sahn, and C. von Vacano, "Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application," arXiv preprint arXiv:2009.10277, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2009.10277

[2] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee, "HateMM: A Multi-Modal Dataset for Hate Video Classification," arXiv preprint arXiv:2305.03915, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.03915

[3] F. T. Boishakhi, P. C. Shill and M. G. R. Alam, "Multi-modal Hate Speech Detection using Machine Learning," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 4496-4499, doi: 10.1109/BigData52589.2021.9671955.

[4] Khera, S., Anuradha, Singh, A., Singh, K.K. (2025). Exploring Multimodal Hate Speech Detection Using Machine Learning and Deep Learning Models. In: Singh, A., Singh, K.K. (eds) Multimodal Generative AI. Springer, Singapore. https://doi.org/10.1007/978-981-96-2355-6_11

[5] M. Subramanian, V. E. Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," Alexandria Engineering Journal, vol. 80, pp. 110–121, 2023. DOI: 10.1016/j.aej.2023.08.038

[6] J. G, B. Saikiran, I. Reddy and M. Abhishek, "Twitter Hate Speech Detection using Machine Learning," in 2024 4th International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2024, pp. 270-278, doi: 10.1109/ICPCSN62568.2024.00051.

[7] A. Abraham, A. J. Kolanchery, A. A. Kanjookaran, B. T. Jose, and D. P. M., "Hate Speech Detection in Twitter Using Different Models," in ITM Web of Conferences, vol. 56, 2023. DOI: 10.1051/itmconf/20235604007

[8] J. S. Malik, H. Qiao, G. Pang, and A. van den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study," arXiv preprint arXiv:2202.09517v2, 2023. Available: https://arxiv.org/abs/2202.09517

[9] K. Mnassri, P. Rajapaksha, R. Farahbakhsh, and N. Crespi, "BERT-based Ensemble Approaches for Hate Speech Detection," arXiv preprint arXiv:2209.06505v2, 2022.

[10] A. Fonseca et al., "Analyzing hate speech dynamics on Twitter/X: Insights from conversational data and the impact of user interaction patterns," Heliyon, vol. 10, 2024, e32246. DOI: 10.1016/j.heliyon.2024.e32246

[11] S. Narang, S. Karki, S. Chauhan, K. Garg and S. S. Samant, "Hate Speech Analysis And Moderation On Twitter Data using BERT And Ensemble Techniques," 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10725330.

[12] Z. Mansur, N. Omar, and S. Tiun, "Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities," IEEE Access, vol. 11, pp. 14065–14094, 2023. DOI: 10.1109/ACCESS.2023.3239375

[13] S. Perera, N. Meedin, M. Caldera, I. Perera, and S. Ahangama, "A Comparative Study of the Characteristics of Hate Speech Propagators and Their Behaviours Over Twitter Social Media Platform," Heliyon, vol. 9, 2023, e19097. DOI: 10.1016/j.heliyon.2023.e19097

[14] M. A. Paz, J. Montero-Díaz, and A. Moreno-Delgado, "Hate Speech: A Systematized Review," SAGE Open, vol. 10, no. 4, pp. 1–12, 2020. DOI: 10.1177/2158244020973022

[15] M. S. Jahan and M. Oussalah, "A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing," Neurocomputing, vol. 546, 2023, Art. no. 126232. DOI: 10.1016/j.neucom.2023.126232

[16] https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/