ISSN 2007-9737

MultiLate Classifier: A Novel Ensemble of CNN-BiLSTM with ResNet-based Multimodal Classifier for AI-generated Hate Speech Detection

Advaitha Vetagiri[1], Prateek Mogha[2], Partha Pakray[1,*]

[1] Dept of CSE, National Institute of Technology Silchar, Assam, India

[2] Dept of EE, National Institute of Technology Silchar, Assam, India

advaitha21_rs@cse.nits.ac.in, prateek21_ug@ee.nits.ac.in, partha@cse.nits.ac.in

Abstract. The rise of multimodal hate speech, which combines text and visual elements, poses significant challenges for online content moderation. Traditional detection models often focus on single modalities and struggle with AI-generated content that is contextually nuanced and semantically complex. These limitations lead to suboptimal performance, as existing frameworks are not robust enough to handle the evolving nature of hate speech across diverse contexts and datasets. An integrated approach that captures the interplay between text and images is needed for more accurate identification. This paper introduces a novel MultiLate classifier designed to synergistically integrate text and image modalities for robust hate speech detection to address these challenges. The textual component employs a CNN-BiLSTM architecture, augmented by a feature fusion pipeline incorporating Three W's Question Answering and sentiment analysis. For the image modality, the classifier utilizes a pre-trained ResNet50 architecture alongside Diffusion Attention Attribution Maps to generate pixel-level heatmaps, highlighting salient regions corresponding to contextually significant words. These heatmaps are selectively processed to enhance both classification accuracy and computational efficiency. The extracted features from both modalities are then fused to perform comprehensive multimodal classification. Extensive evaluations of the MULTILATE and MultiOFF datasets demonstrate the efficacy of the proposed approach. Comparative analysis against state-of-the-art models underscores the robustness and generalization capability of the MultiLate classifier. The proposed framework enhances detection accuracy and optimizes computational resource utilization, significantly advancing multimodal hate speech classification.

Keywords. Hate speech detection, multimodal classification, CNN-BiLSTM, ResNet50, diffusion attention attribution maps.

1 Introduction

With the ever-evolving growth of websites, the opportunity to communicate and interact with people worldwide has increased dramatically. Nevertheless, this interconnection feature in the digital society has enabled the proliferation of hate speech [21] and, consequently, caused numerous social and personal adverse effects. Hate speech is a set of words and actions that are abusive, prejudicial, threaten or depict other individuals or a group as inferior based on characteristics such as race, religion, nationality, gender, sexual preference or disability. It is, therefore, imperative to establish efficient approaches for hate speech identification and prevention, which have spurred investigations in the domains of natural language processing (NLP) [61] and computer vision (CV) [9].

Identifying hate speech is challenging as the accurate interpretation of language and words depending on the context is complex to detect, and in addition, the hate speech on social media exists in a multimodal [12] format where text and images are making them complex towards processing and classification. Earlier research in the context of hate speech detection has predominantly used text

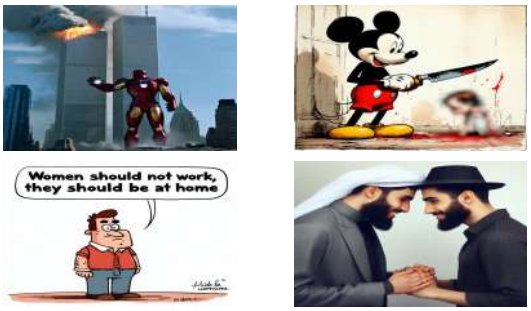


Fig. 1. AI-generated hateful images based on themes of historical events, sexism, racism, and biases.

Fig. 2. An AI-generated image from MULTILATE dataset [52] from corresponding prompt text "*Native Americans - a primitive people who want to live the way they did hundreds of years ago, -* ***Hate***".

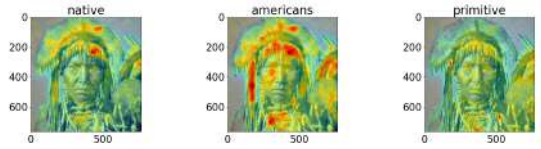


Fig. 3. Heat maps generated using Diffusion Attention Attribution Maps for the AI-generated figure 2.

data and has widely adopted machine learning [1] and deep learning [53] methods such as logistic regression, Naïve Bayes classifier, and different neural networks. Despite having some levels of success, these methods can generally have issues with generalisation and stability, especially when they encounter fairly ambiguous and complex hate speech instances.

This is a new development as Artificial Intelligence (AI) + hate speech continues to pop up everywhere [56], and it is exceptionally hard to recognize and eradicate them. Using many new sophisticated language models, it is now possible to generate essentially fake texts that cannot be distinguished from texts produced by humans, which, in turn, rapidly and on a very large scale convey hatred ideologies. Due to the individual nature of sharing and its ability to pinpoint specific and highly unique messages on a selected target, this technology can be used to deliver hate speech that is more personal and devastating to the targeted parties.

In addition to that, AI-generated posts can also outsmart existing content moderation techniques and, in the long run, allow toxic messages to spread. One glaring service of the AI involves changing a historical speech by Adolf Hitler to say antisemitic things in English. This manipulated video, shared by an influencer, quickly garnered over 15 million views on X (*formerly Twitter*) in March 2024 [1]. Figure 1 shows a few AI-generated images. These incidents highlight the growing worries of researchers and monitoring groups regarding the expansion of AI, which has produced hate and a pressing requirement for critical evaluation and strong detection systems to counter the expanding hazard.

Modern developments include working with multimodal media [3], combining textual and visual data for better results and more accurate identification of hate speech. This integration takes advantage of the synergies between various types of data because the strengths of one type can compensate for the shortcomings or weaknesses of another type. However, some limitations exist when extracting and integrating features from text and images to construct specific and precise classifiers [16]. An example of an AI-generated image from the MULTILATE dataset is shown in figure 2, and its heatmaps are shown in figure 3.

In view of the state-of-the-art multimodal hate speech detection systems, the present study proposes a new Multilate classifier for Multimodal hate to improve hate speech detection by incorporating sophisticated feature extraction methodologies and multiform data in the given context. For the text modality, the approach involves the 3WQA (who, what, and why question answering) [38] and sentiment analysis together with the

[1] https://tinyurl.com/4rf59cru

Convolutional Neural Networks - Bi-directional Long Short-Term Memory (CNN-BiLSTM) [50] architecture to capture the detailed features of the text. In the case of the image input layer, an input layer is created and connected to a Residual Neural Network (ResNet50) [39] such that only heat maps created by the Diffusion Attention Attribution Maps (DAAMs) [47] of the most crucial words in the equivalent text are passed to the network. This operationalization strategy uses a combination of modalities to enhance the classifier's performance in detecting hate speech with high precision and recall. To assess the effectiveness of the proposed Multilate classifier, numerous experiments were performed on the MultiOFF dataset [46], which compared the new model with several SOTA models based on several key performance indicators, such as precision, recall, and the F1 score. The findings highlight specific advancements in performance gains and hint at the potential of multimodal integration and superior feature extraction in the fight against hate speech online.

The main novelties and contributions of this article are as follows:

1. A novel CNN-BiLSTM with ResnNet50-assisted MultiLate classifier has been proposed for a multimodal text and image-based hate speech detection system.
2. An effective feature fusion pipeline has been designed and developed using 3WQA and sentiment analysis for text modality. The pipeline selectively processes DAAMs-generated heatmaps of the most important words, enhancing performance and reducing computational time for the imaging modality.
3. A benchmark image and text-based multi-modal AI-generated hate speech detection dataset is formulated using stable diffusion images with 3WQA and DAAMs for enhanced explainability.
4. A detailed and comprehensive multimodal fusion strategy combining textual and visual features has been identified to demonstrate significant performance improvements.

The remaining section of this article is organised as follows - Section 2 will consist of the background work and relative work done in the field, Section 3 is the data sampling and training-test data splits, and section 4, is the system overview of the models and was demonstrated. Section 5 explains the experimental setting of the proposed models. Finally, Section 6 will explain how these models fared on multimodal classification with SOTA comparison and some drawbacks.

2 Literature Survey

The detection of hate speech has improved dramatically, driven by the rise of abusive content on digital devices. Traditionally, search methods are mainly based on text-based classification, using machine learning and deep learning techniques. However, while hate speech increasingly incorporates more features—such as images with text, these traditional models face limitations. This literature review examines the evolution of hate speech identification models, highlighting the shift toward multiple perspectives and their associated challenges.

2.1 Hate Speech Detection in Text

Hate speech detection in text involves identifying and classifying offensive, discriminatory, or harmful language in written content using natural language processing [51] and machine learning techniques [18]. Asogwa *et al.* [4] proposed an optimized hate-speech detection using multiple machine learning models and an intense data pre-processing pipeline. The proposed model incorporates Support Vector Machines (SVM) and Naive Bayes (NB) as the primary classifier for recognising patterns involved in processed shape speech data and achieved an optimal performance accuracy of 99.37% with the SVM classifier. In comparison, The authors in [11] proposed HateBERT, which incorporates Bidirectional Encoder Representations from Transformers (BERT) [26]. The proposed model utilizes transformer architecture to understand the context of words in a sentence more effectively by considering both preceding and following words instead of just

processing text sequentially, allowing it to achieve superior performance.

HateBERT is a fine-tuned variant of BERT that is trained on a large dataset of Reddit comments from banned communities that were known for offensive content; this enhances the BERT's ability to detect offensive text, allowing HateBERT to outperform BERT models in the detection of hate speech. *Rajput et al.* [36] incorporated static BERT embeddings for effective pattern generation and classification of hate speech. The proposed model generates contextualized embeddings for each word in the given text using BERT, representing the words' general meaning across different contexts. Subsequently, the extracted static embeddings are arranged into an embedding matrix representing words as fixed vectors and fed to a Convolution Neural Network (CNN) to detect patterns and features in the text.

2.2 Hate Speech Detection in Image

The rise of visual content on social media platforms has made images a dominant mode of communication due to their ability to convey messages faster and transcend language barriers. With the rise in the image as the information sharing modality, multiple forms of hate speech are also inscribed in them, making them more communal and malicious. Researchers [32, 29] around the globe deal with image-based hate speech detection using various tools and techniques to ensure a safer online environment.

Putra *et al.* proposed a CNN-based deep learning model for detecting hate speech on images. The proposed model classifies whether the image is hateful or not but does not capture the image's caption and ignores the other features of the images, making them vulnerable to satire or ironic content. Similarly, multiple pre-trined techniques are used in detecting hate speech in images by replacing CNNs with finely-tuned VGG16 and Xception [29]. However, as effective as these models are in detecting hateful text or images, they are designed exclusively for textual or image analysis and cannot handle multi-modal content such as memes, etc. This limitation underlines the need for multimodal approaches integrating various data types to comprehensively understand hate speech across different media. additionally, in the case of static embedding, their effectiveness is also subject to the quality and balance of the training data, which can impact the model's ability to generalize to diverse or unseen data.

2.3 Multimodal Hate Speech Detection

Multimodal hate-speech detection has emerged as a significant research area in recent years, catalyzed by Facebook's Hateful Meme Detection Challenge in 2020. The top three winners of this challenge utilized transformer models to combine image and text features, achieving notable performance improvements. A particularly effective approach involved an ensemble of four distinct Vision-Language (VL) transformer models: VL-BERT [45], UNITER [13], VILLA [15], and ERNIE-Vil [58]. These models, tailored for the task of detecting hateful memes, were shown to enhance meme classification accuracy by integrating text and image features [63].

This strategy also incorporated additional data preprocessing steps, such as Google Vision Web Entity Detection for contextual information and the FairFace classifier for race and gender identification.

VL-BERT was extended to link image regions with external text tokens, UNITER retained its Image-Text Matching (ITM) head to utilize pre-trained alignment tasks, and ERNIE-Vil employed scene graph information for a more nuanced understanding. This ensemble method demonstrated improved model generalization and performance in hateful meme detection tasks. Subsequently, Niklas *et al.* proposed an ensemble of 12 Vision+Language (V+L) transformer models, including VisualBERT, UNITER, and ERNIE-ViL, which was employed for meme classification, combining image and text analysis to detect hate speech [27]. The models were fine-tuned with enhancements to boost performance, and the final predictions were averaged using various methods to enhance accuracy.

Early-fusion multimodal models, such as LXMERT, UNITER, and Oscar, were employed by

Riza *et al.* [49] to process images and text together, aiming to capture their combined meaning. These models were initially trained on extensive datasets of images and text and subsequently fine-tuned on the Hateful Memes dataset. Techniques such as confounder upsampling, loss re-weighting, cross-validation ensemble optimization, and margin ranking loss were utilized to improve classification accuracy.

Furthermore, integrating object detection tags with YOLO9000 [40] allowed the models to better identify hate speech by recognizing target groups in images. Jafar *et al.* [6] combined a ResNet-50 for extracting image features, BERT for generating text embeddings, and an LSTM for capturing sequential relationships in text, integrating these components to effectively detect and classify hateful content in memes. Various state-of-the-art have also employed combinations of multiple machine learning models, such as Support Vector Machines and Naive Bayes, with convolutional neural network architectures like VGG16 and Xception [43], and the most optimal performance was achieved with an LSTM combined with VGG16. Multiple benchmark techniques have been also developed to separately identify whether images are not safe for work (NSFW) and detect hate speech in text [8].

Gokul *et al.* [22] utilized Contrastive Language-Image Pre-training (CLIP) encoders to represent images and text within a shared feature space, facilitating the capture of relationships between text and image features. To enhance the interaction between these features, a Feature Interaction Matrix (FIM) was employed to model cross-modal interactions explicitly. Following the flattening of the CLIP output layers, a simple classifier was applied to the cross-align fusion of CLIP features, achieving a commendable accuracy of 85.8 on the Hateful Memes Challenge (HMC) dataset. Recent efforts have also extended hate speech detection to videos [54], although some approaches focus solely on the textual components, overlooking other video features.

An alternative method by Aneri *et al.* [37] extends beyond textual analysis by also considering available video features, such as the subject's emotional states, and employs various techniques, including sentiment analysis and speech emotion recognition, to identify emotions.

Despite the advancements brought by cross-modal interaction techniques, SOTA models face several limitations. For instance, the Hate-CLIPper model [22] presents a significant challenge in terms of computational complexity due to the high dimensionality of the Feature Interaction Matrix (FIM) [24]. With dimensions such as n=1024 and m=1024, the model must manage a parameter space in the order of $O(n^2m)$, which is computationally demanding. Moreover, these models are susceptible to adversarial attacks [44]. Adversarial attacks involve making subtle alterations to input images or text, which can mislead the models into incorrect classifications. For example, minor modifications to a meme's visual or textual content can interfere with the feature extraction process, resulting in misinterpretation. The HateCLIPper model, which relies on a Feature Interaction Matrix to capture cross-modal interactions, is particularly vulnerable to sophisticated attacks targeting these interactions. Consequently, despite their strong performance, these models require robust defence mechanisms to ensure reliability in real-world applications.

Synthesis: The present literature shows a shift from traditional text-based hate speech detection to more complex multimodal approaches. While early models focused solely on text, such as providing foundational insights, they are now challenged by the growing complexity of hate speech that combines text and images. Current multimodal models, including those using transformers, CNNs and pre-trained architectures, have made some advances but often struggle with modern internet content's subtle and contextually rich nature. The MultiLate classifier is proposed to address these challenges, a new approach to integrating text and image modalities for robust hate speech detection. The MultiLate classifier combines advanced text analysis using CNN-BiLSTM and 3WQA with image analysis through pre-trained ResNet50 and Diffusion Attention Attribution Maps. This approach enhances the classifier's ability to handle complex multimodal content and improves detection accuracy.

3 Dataset

This study utilised two datasets: the MULTILATE dataset [52] and the MultiOFF dataset [46]. Each dataset provides valuable resources for analyzing multimodal hate speech and offensive content detection.

3.1 MULTILATE Dataset

The MULTILATE dataset is a comprehensive resource designed to identify instances of hate speech, particularly sexism and racism, in online content. This dataset comprises a total of 2.6 million examples extracted from 11 different datasets on sexism and racism, but a total of 1004 samples have been released, which are divided into training and testing splits. The dataset includes labels for binary classification ("Hate" and "Not Hate") as shown in table 1 and multiclass classification ("Sexist", "Racist", and "Neither") as shown in the table 2. Statistical details about the dataset are shown in figure 4a for binary classes and figure 4b for multi classes. The MULTILATE dataset was created to provide a diverse multimodal collection of hateful statements paired with visually generated images. For the image component, Stable Diffusion (SD 2.1) [41] was employed to generate hateful images that match textual prompts, using its advanced text-to-image synthesis capabilities. A pipeline was designed to create three images per text prompt, which were then ranked using CLIP [34], a model that scores images based on text alignment, to select the most representative image for each prompt. Additionally, Diffusion Attention Attribution Maps (DAAM) [48] were used to produce heatmaps highlighting areas of the image relevant to specific words in the text, enhancing visual explainability.

The dataset also incorporates a 3W (who, what, why) Question-answering (QA) component, enriching textual explanations with automatic QA pairs for each text. Using semantic role labelling (SRL), key phrases are extracted, and ProphetNet [33] generates relevant questions based on these phrases. T5 [35] is then applied to produce accurate answers, verified by human reviewers, offering deeper insight into the motivations and actors within each statement. This comprehensive process ensures that MULTILATE provides rich, multimodal data for hate speech analysis. Details about the data creation, annotation, and validation process can be found in [52].

Table 1. MULTILATE Dataset - Binary Classification

Splits	Hate	Not Hate	Total Count
Train	573	280	853
Test	103	48	151

Table 2. MULTILATE Dataset - Multiclass Classification

Splits	Sexist	Racist	Neither	Total Count
Train	296	277	280	853
Test	49	54	48	151

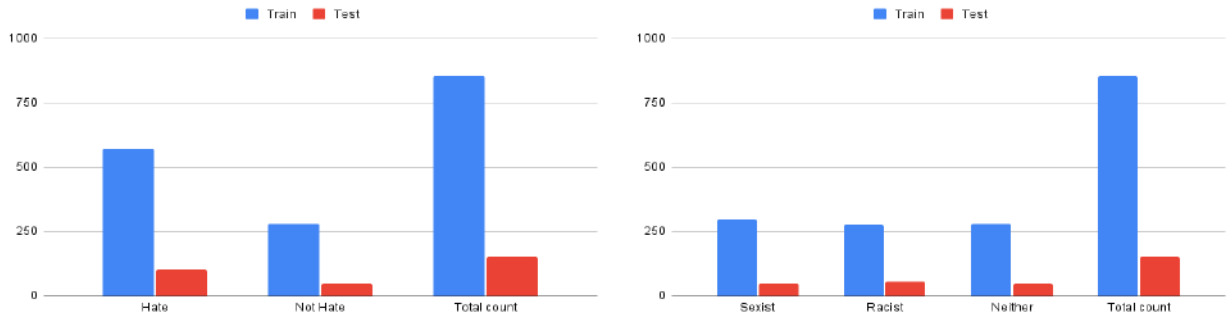
3.2 MultiOFF Dataset

The MultiOFF dataset is a multimodal meme created to identify offensive content in both image and text modalities. This dataset includes 743 memes, manually annotated with labels indicating whether the content is offensive or not; more information can be seen in the table 3 and the dataset statistics are shown in the figure 5.

The memes were collected from social media platforms like Reddit, Facebook, Twitter, and Instagram, including image URLs and embedded text. The dataset provides a valuable resource for developing classifiers capable of detecting offensive content by combining textual and visual information. The data collection and annotation process details are described in [46].

Table 3. MultiOFF Dataset

Splits	Offensive	Not Offensive	Total
Train	187	258	445
Test	59	90	149
Val	59	90	149



(a) Statistics of binary classes.

(b) Statistics of multi classes.

Fig. 4. MULTILATE dataset statistics for binary and multi-class classification

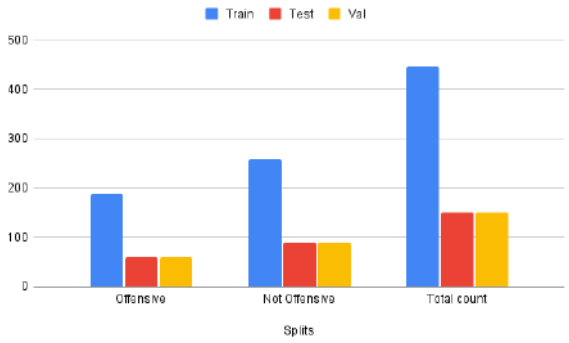


Fig. 5. MultiOFF dataset statistics for binary classification

4 Proposed Methodology

The problem of hate speech detection in the text has gained importance as a subtask in natural language processing, as it focuses on the problem of filtering out the unfavourable and discriminating content in the information space. This is because one of the main problems in the current state of the domain is in feature extraction and representation of the input data, which can often be complex and can include text, image, and other modalities [14, 28, 60, 55].

To overcome this challenge, the present study suggests a new deep learning model that combines the best features of the convolutional neural networks (CNNs) [20] and bidirectional long short-term memory (BiLSTM) [7] for text data and the CNNs for image data [19]. The proposed model is named MultiLate, which first extracts text features using CNNs and then later exploits sequence modelling of BiLSTM for text classification, whereas ResNet-50 is used for image classification. The features obtained from both modalities are then fused to perform the hate speech classification at the multimodal level. The entire pipeline of the MultiLate is shown in figure 6, and the architecture is shown in figure 7.

4.1 Textual Modality

For the text data, the CNN-BiLSTM model is defined as taking a sequence of words as input. The sequence of words is then embedded. Embedding converts words into continuous space, where words in similar contexts are closer in the vector space [14, 60]. Further, it also has features like 3WQA (who, what, why) to focus on the reasonability and context of the sentence, as well as sentiment to detect positive and negative sentiments [60].

Subsequently, to embedding, the obtained vectors and extra features go through consecutive layers of convolution, where features in the input sequence are learned by applying a dot product between the input sequence and a filter in Windows. The layers' output is a set of feature maps; each map captures some local interaction pattern in the input sequence [14, 60].

Subsequently, all the feature maps of the convolution layer are transferred to the BiLSTM

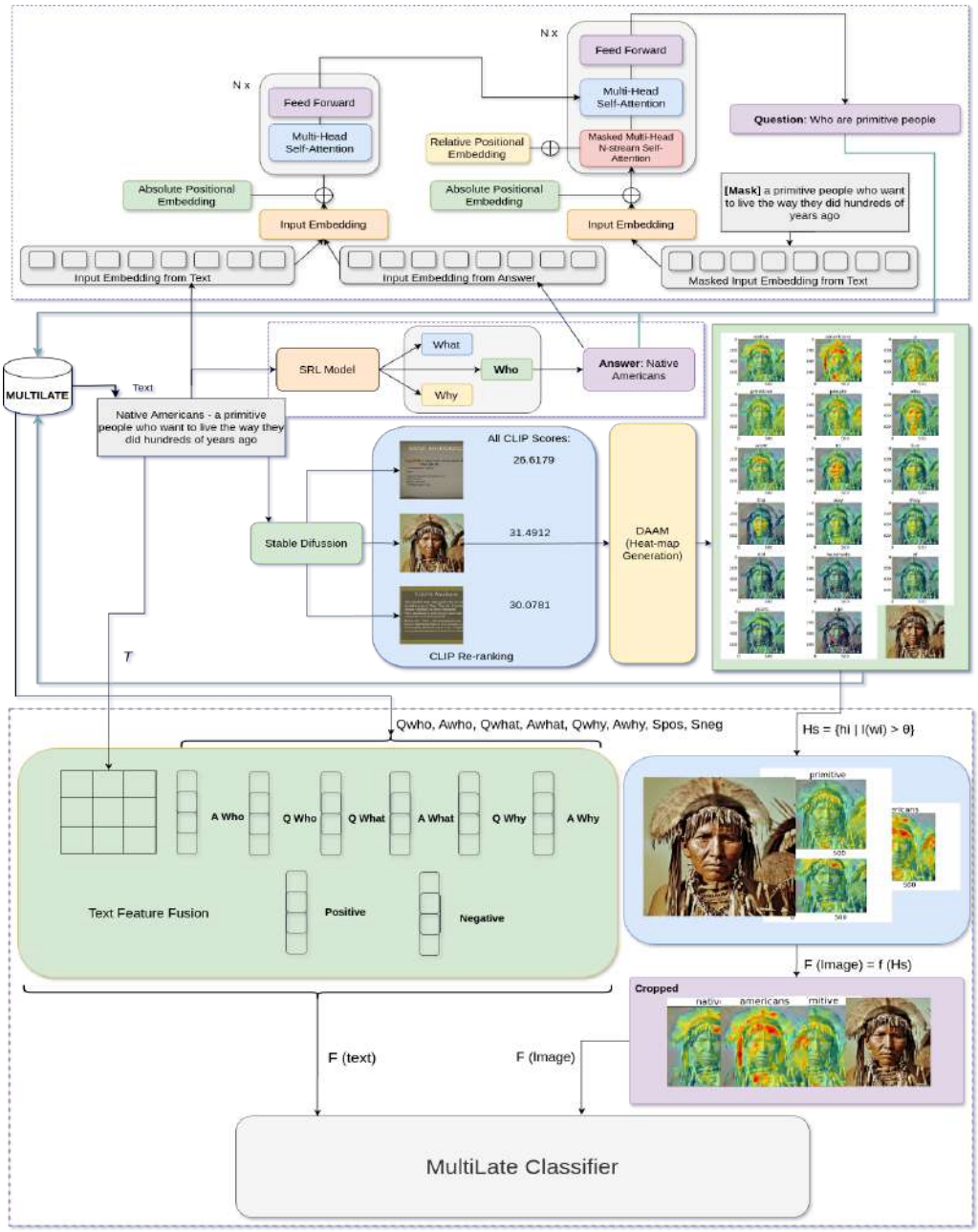


Fig. 6. Overview of MultiLate feature fusion framework for classification - which integrates with MULTILATE dataset contains Stable Diffusion, SRL and T5 Models, used for generating synthetic multimodal Hate Speech data

layer. BiLSTM is a recurrent layer that has a memory of previous inputs it has seen. It is composed of two LSTMs. The first operates in the forward direction, and the second in the backward direction on the input sequence. The output of each LSTM is the hidden state, which is

further concatenated to give the final output, which contains information from both directions of the sequence [60, 62].

Three-W Question Answering (3WQA)

$$Q_{\text{who}} = \text{QA}_{\text{who}}(T), \tag{1}$$

$$Q_{\text{what}} = \text{QA}_{\text{what}}(T), \tag{2}$$

$$Q_{\text{why}} = \text{QA}_{\text{why}}(T). \tag{3}$$

Here, $Q_{\text{who}}, Q_{\text{what}}$, and Q_{why} are the outputs of the question-answering model for the "who," "what," and "why" questions, respectively, applied to the text T.

Sentiment Analysis

$$S_{\text{pos}}, S_{\text{neg}} = \text{SentimentAnalysis}(T). \tag{4}$$

Here, S_{pos} and S_{neg} are the positive and negative sentiment scores obtained from the sentiment analysis model applied to the text T.

Feature Fusion for Text

$$F_{\text{text}} = [Q_{\text{who}}, A_{\text{who}}, Q_{\text{what}}, A_{\text{what}}, Q_{\text{why}}, A_{\text{why}}, S_{\text{pos}}, S_{\text{neg}}]. \tag{5}$$

The feature vector F_{text} is obtained by concatenating the 3WQA outputs and sentiment scores.

CNN-BiLSTM Classification

$$H_{\text{CNN}} = \text{CNN}(F_{\text{text}}), \tag{6}$$

$$H_{\text{BiLSTM}} = \text{BiLSTM}(H_{\text{CNN}}), \tag{7}$$

$$\hat{y}_{\text{text}} = \text{Dense}(H_{\text{BiLSTM}}). \tag{8}$$

Here, H_{CNN} is the feature map obtained from the CNN layer, H_{BiLSTM} is the hidden state from the BiLSTM layer, and $\hat{y}_{\text{text}}$ is the final classification output for the text modality.

4.2 Image Modality

For image data, the study employs one of the most potent ConvNet models, ResNet-50, with deep architecture that helps address the vanishing gradient problem by applying the residual learning technique. ResNet-50 uses several convolutional layers implemented in ResNet architecture with skip connections to transform inputs into high-level features, vital in completing image classification [39].

A pipeline has been created using DAAMs to generate pixel-level attribution heatmaps highlighting which image regions correspond to the words in the associated hate speech prompt [10]. In this pipeline, instead of extracting features from all heatmaps generated based on the input words in the text, it has been modified to work on the heatmaps of the most important words in the sentence using the techniques of Contextual Importance like Term Frequency-Inverse Document Frequency (TF-IDF) [59] and Attention Mechanisms [30]. This process increases the strength and shortens computational time. These heatmaps highlight the regions of the image most relevant to the associated textual content, and hence, the features obtained are contextually meaningful for classification.

After generating and integrating the heatmaps, the features are extracted using the pre-trained weights of ResNet-50. These extracted features are then passed through two dense layers, further refining the feature representation by learning complex patterns and interactions within the data. The dense layers are followed by a flattened layer, which transforms the multi-dimensional feature maps into a one-dimensional vector suitable for the final classification stage.

Formulas for Selective Processing of DAAMs-Generated Heatmaps

Let $T = \{w_1, w_2, \ldots, w_n\}$ be the set of words in the text modality and $H = \{h_1, h_2, \ldots, h_n\}$ be the corresponding DAAMs-generated heatmaps. The importance score $I(w_i)$ of each word w_i is computed using TF-IDF and Attention Mechanisms.

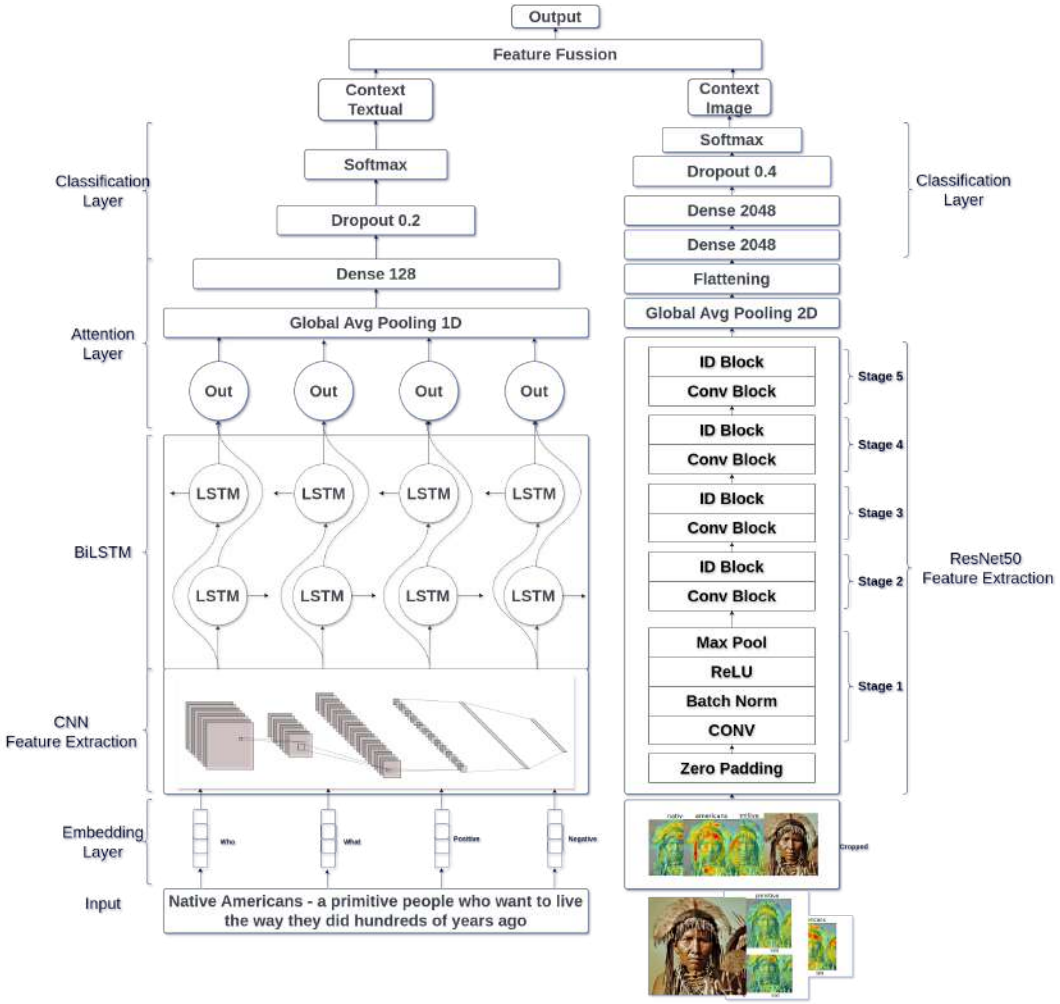


Fig. 7. MULTILATE Classification Architecture for Multimodal Hate

The selected subset of heatmaps H_s is given by:

$$H_s = \{h_i \mid I(w_i) > \theta\}, \tag{9}$$

where θ is a predefined threshold for word importance.

The feature vector F for the image modality is then:

$$F = f(H_s), \tag{10}$$

where f is the feature extraction function applied to the selected heatmaps H_s. This process enhances the model's accuracy and reduces computational time by focusing on the most contextually significant features.

4.3 Multimodality

The multimodality explores integrating textual and visual data in the proposed MultiLate classifier to enhance hate speech detection. By leveraging complementary information from different modalities, the classifier aims to capture the nuanced and context-dependent nature of multimodal hate speech [17].

This section discusses the process of feature extraction and fusion, demonstrating how the combined modalities lead to more accurate and robust classification results compared to single-modality approaches. Integrating these

Algorithm 1 Feature Fusion Pipeline for Text and Image Modalities

Input:

— Textual input $T = \{w_1, w_2, \ldots, w_n\}$ with 3WQA and sentiment analysis results.

— Image input I with DAAMs-generated heatmaps $H = \{h_1, h_2, \ldots, h_n\}$.

Output:

— Combined feature vector F for multimodal classification.

1: Compute importance scores $I(w_i)$ for each word w_i in T using Contextual Importance techniques.
2: Select the most important words based on $I(w_i)$.
3: Generate a subset of DAAMs-generated heatmaps $H_s = \{h_i \mid I(w_i) > \theta\}$, where θ is a threshold.
4: Extract textual features F_{text} using CNN-BiLSTM model with T.
5: Extract image features F_{image} using ResNet50 model with H_s.
6: Concatenate F_{text} and F_{image} to form the combined feature vector F.
7: **return** F

advanced techniques underscores the potential of multimodal frameworks in effectively addressing complex hate speech scenarios.

The outputs from the image processing pipeline, now represented as a one-dimensional feature vector, are fused with those from the text processing pipeline. This fusion has been achieved through a concatenation technique [42]. The fused features are subsequently passed through a dense layer for further classification, where the model applies a sigmoid activation function to obtain a probability score, indicating the likelihood of the input being classified as hate speech.

The algorithm 1 demonstrates how the multimodal fusion and the fused features are sent to the classification model for classification.

5 Experimental Setup

The experimental setup is meticulously designed to evaluate the efficacy of the proposed MultiLate classifier. The study focuses on extracting semantic, syntactic, and linguistic features [57] from textual data using advanced NLP techniques while leveraging pre-trained deep-learning models for visual feature extraction. The textual features are enriched with Global Vectors (GloVe) [31] word embeddings and further augmented by a feature fusion pipeline incorporating 3WQA and sentiment analysis from SentiWordNet [5]. Concurrently, image features are extracted using a pre-trained ResNet50 model and enhanced with DAAMs for contextually meaningful heatmaps. These processed features from both modalities are then integrated to perform multimodal classification. The proposed models are rigorously tested using repeated k-fold cross-validation [2], ensuring robust evaluation through precision, recall, and F1-score metrics. This comprehensive experimental framework underscores the robustness and generalization capability of the MultiLate classifier.

5.1 Feature Extraction

5.1.1 Text Modality

Semantic, syntactic, and linguistic features are extracted from textual data using advanced Natural Language Processing (NLP) techniques.

Specifically, word embeddings are generated using GloVe for Word Representation, which encapsulates semantic relationships among words in a high-dimensional space [25]. Furthermore, additional features are incorporated, including 3W's questions and corresponding answers, along with sentiment polarity (positive and negative sentiments) derived from SentiWordNet. These additional features are seamlessly integrated with the primary textual features through sophisticated feature fusion methodologies, thereby enriching the representational capacity of the textual data.

5.1.2 Image Modality

Visual feature extraction from images is accomplished using a pre-trained ResNet50 model, which leverages transfer learning from the ImageNet dataset to capture complex visual patterns. These extracted visual features are subsequently fused with pixel-level attribution heatmaps generated by DAAMs. These heatmaps identify the most pertinent regions of the image that correspond to the significant words in the associated hate speech text. The selective processing of heatmaps based on contextual importance enhances the relevance of the features and reduces computational overhead.

5.2 Classification

5.2.1 Text Classification

A Convolutional Neural Network integrated with a Bidirectional Long Short-Term Memory network is employed to classify textual data.

The architecture comprises a single BiLSTM layer with 128 hidden units, capturing the text's bidirectional dependencies. This is followed by a GlobalAveragePooling1D layer to reduce the dimensionality, a dense layer with 128 neurons to learn complex feature interactions and a final softmax layer for output classification. The summary of the parameters used is shown in tables 4, 5 and table 6, the summary of hyperparameters and performance metrics for different runs and figure 8 shows the visualization of the same.

5.2.2 Image Classification

The classification of image data is performed using a pre-trained ResNet50 model. The architecture includes a GlobalAveragePooling2D layer to aggregate the spatial dimensions, followed by two dense layers with 2048 neurons each, which enable the model to learn intricate patterns in the visual data. The classification is finalized with a softmax layer, providing probabilistic outputs for each class. The summary of the Parameters used is shown in the table 7.

5.2.3 Multimodal Classification

The contextually enriched textual and visual features are fused for multimodal classification to create a comprehensive feature representation.

These fused features are then input into a classification model that leverages the combined strengths of both modalities, enhancing the robustness and accuracy of the classification.

The proposed models are rigorously evaluated using the repeated k-fold cross-validation technique, with the results averaged over 5 folds to ensure robustness and generalizability. The evaluation metrics employed include precision, recall, and F1-score, providing a comprehensive assessment of the model performance across different aspects of classification accuracy.

6 Results and Discussion

The results and discussion section delves into the performance metrics of the proposed MultiLate classifier, providing a detailed analysis of its efficacy in multimodal hate speech detection. The classifier's performance is assessed on binary and multiclass classification tasks across the MULTILATE and MultiOFF datasets. Key metrics such as precision, recall, and F1-score are compared against baseline models and state-of-the-art approaches to underscore the improvements brought by the proposed methods. The section also examines the impact of integrating advanced feature extraction techniques, such as 3WQA and DAAMs, on the classifier's accuracy and computational efficiency. The results demonstrate the model's robust generalization capabilities and superior performance, highlighting its potential in effectively combating hate speech across diverse online platforms.

6.1 Performance Analysis on Multilate Dataset Using proposed Multilate classifier

The performance of the proposed MultiLate classifier was rigorously evaluated on the MultiLate dataset, which includes diverse text and image modalities.

Table 4. Summary of Hyperparameters and Performance Metrics for Runs 1 to 7

Hyper-parameters	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7
Con1D	32	32	64	64	**64**	64	64
kernel_size	2	3	2	2	**2**	3	2
LSTM	64	64	64	128	**128**	128	256
dropout	0.1	0.1	0.1	0.1	**0.2**	0.2	0.2
Dense	128	128	128	128	**128**	128	128
activation (first)	relu	relu	relu	relu	**relu**	relu	relu
dropout (second)	0.1	0.1	0.1	0.1	0.2	0.2	0.2
activation (second)	softmax	softmax	softmax	softmax	**softmax**	softmax	softmax
optimizer	adam	adam	adam	adam	**adam**	adam	adam
batch_size	128	128	128	128	**128**	128	128
epochs	25	25	15	15	**15**	15	15
Accuracy	0.7817	0.7608	0.7875	0.7605	**0.8029**	0.7637	0.7946
loss (Train)	0.0906	0.1383	0.7607	0.7494	**0.0188**	0.0767	0.0758
Test Accuracy	0.77	0.75	0.77	0.75	**0.79**	0.75	0.77

Table 5. Summary of Hyperparameters and Performance Metrics for Runs 8 to 13

Hyper-parameters	Run 8	Run 9	Run 10	Run 11	Run 12	Run 13
Con1D	64	64	64	64	64	64
kernel_size	2	2	2	2	2	2
LSTM	256	256	256	256	512	512
dropout	0.3	0.2	0.2	0.2	0.2	0.4
Dense	128	128	256	128	128	128
activation (first)	relu	relu	relu	relu	relu	relu
dropout (second)	0.3	0.2	0.2	0.1	0.1	0.1
activation (second)	softmax	softmax	softmax	softmax	softmax	softmax
optimizer	adam	adam	adam	adam	adam	adam
batch_size	128	256	256	256	512	128
epochs	15	15	15	25	5	5
Accuracy	0.7871	0.78	0.7858	0.7683	0.7417	0.7767
loss (Train)	0.0726	0.0726	0.0726	0.0781	0.1954	0.1562
Test Accuracy	0.77	0.76	0.76	0.74	0.73	0.75

Table 6. Text Classification Parameters

Parameter	Setting
Word Embedding	GloVe 300d
Batch Size	128
Epochs	15
Learning Rate	0.001
Optimizer	Adam
Loss Function	Categorical Cross Entropy

The evaluation metrics, including precision, recall, and F1 score, were used to compare the performance of baseline models against the proposed model for both binary and multiclass classification tasks.

The results reveal significant improvements brought about by the proposed approach.

6.1.1 Baseline Model Analysis

CNN-BiLSTM (Text Modality)- Binary Classification: The baseline CNN-BiLSTM model achieved a training precision, recall, and F1 score of 0.72,

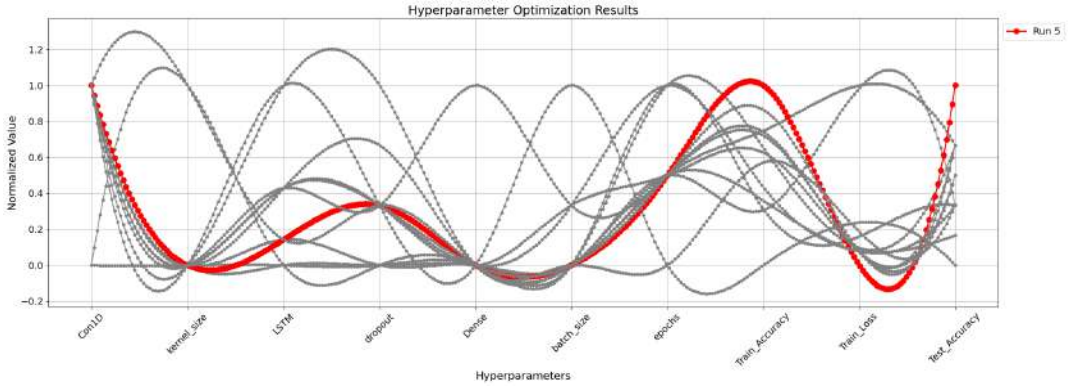


Fig. 8. Visualization of Hyperparameters and Performance Metrics Across Different Runs. Run 5 is Highlighted in Red to Indicate the Best Performance. The Initial Activation Function is ReLU, the Loss Function is Categorical Crossentropy, the Final Activation Function is Softmax, and the Optimizer Used is Adam

Table 7. Image Classification Parameters

Parameter	Setting
Image Size	512×512
Batch Size	32
Epochs	Avg 21
Learning Rate	0.00001
Optimizer	Adam
Loss Function	Categorical Cross Entropy
Rescale	1/255

0.71, and 0.72, respectively. For testing, the precision, recall, and F1 score were slightly lower at 0.71, 0.70, and 0.70.

This small decrease in testing performance indicates good generalization from training to testing data.

For Multiclass Classification, the model's training precision, recall, and F1 score were 0.59, 0.58, and 0.59, respectively. Testing metrics were similar, with precision, recall, and F1 scores of 0.59, 0.59, and 0.58, indicating consistent performance across datasets.

VGG16 (Image Modality)-Binary Classification: The VGG16 model demonstrated a training precision, recall, and F1 score of 0.62, 0.60, and 0.61. The testing results closely matched precision, recall, and F1 scores at 0.62, 0.60, and 0.61, showing stable performance. Multiclass Classification: Training results were 0.41, 0.40, and 0.41 for precision, recall, and F1 score, respectively. Testing metrics were slightly lower at 0.38, 0.40, and 0.39, suggesting room for improvement in handling more complex classification tasks.

ResNet50 (Image Modality)-Binary Classification: Training precision, recall, and F1 scores were 0.62, 0.61, and 0.61, with testing results at 0.60, 0.60, and 0.61. This shows that ResNet50 consistently performs. Multiclass Classification: Training precision, recall, and F1 scores were 0.41, 0.40, and 0.41, while testing results were 0.39, 0.40, and 0.40, indicating stable yet modest performance.

Multimodal - Binary Classification: The multimodal approach achieved training metrics of 0.69, 0.68, and 0.69 for precision, recall, and F1 score. Testing results were similar, at 0.68 for all three metrics, reflecting a well-generalized model.

Multiclass Classification: The training precision, recall, and F1 scores were 0.53, 0.51, and 0.53, with testing scores slightly lower at 0.52 for all metrics.

6.1.2 Proposed Model Analysis

CNN-BiLSTM (Text Modality)-Binary Classification: The proposed model achieved a precision, recall, and F1 score of 0.78 across both training

and testing datasets, demonstrating a substantial improvement over the baseline model and excellent generalization. Multiclass Classification: Training metrics were 0.64, 0.64, and 0.63, while testing results were 0.64, 0.62, and 0.61, indicating a robust performance with minor drops in testing accuracy.Training accuracy and loss of text classification in the binary setting are given in figure 10 while figure 11 show the same in the multiclass setting, figure 12b and figure 12b shows the confusion matrix for both binary and multiclass settings respectively.

ResNet50 (Image Modality)-Binary Classification: Training precision, recall, and F1 scores were 0.78, 0.77, and 0.74, respectively. Testing metrics were close, with precision, recall, and F1 scores at 0.72, 0.72, and 0.72, showcasing consistency and reliability. Multiclass Classification: The model had training precision, recall, and F1 scores of 0.64, 0.64, and 0.63. Testing metrics were slightly lower at 0.61, 0.62, and 0.60, demonstrating a good balance between training and testing performance. Training accuracy and loss of image classification in the binary setting are given in figure 13 while figure 14 show the same in the multiclass setting, figure 15a and figure 15b shows the confusion matrix for both binary and multiclass settings respectively.

Multimodal-Binary Classification: The multimodal proposed model achieved a training precision, recall, and F1 score of 0.75, 0.74, and 0.75. Testing results were also robust at 0.74 for all metrics, confirming strong generalization capabilities. Multiclass Classification: The training metrics were 0.70, 0.69, and 0.70 for precision, recall, and F1 score. Testing results were very similar at 0.68, 0.68, and 0.68, indicating the model's efficacy in handling complex classification tasks. Table 8 summarizes the stated results.

6.1.3 K-Fold Validation Results

The performance of the proposed MultiLate classifier was further validated using K-fold cross-validation, which provided additional insights into the model's consistency and reliability across different folds of the dataset. The results are summarized in the table 9.

These K-fold validation results further affirm the consistency and robustness of the proposed MultiLate classifier. The average metrics across folds indicate stable and reliable performance for both text and image modalities in binary and multiclass settings. This consistency is crucial for real-world applications, where models must perform reliably across diverse and unseen data.

The high precision, recall, and F1 scores demonstrate the proposed model's enhanced capability in effectively handling the complexities associated with hate speech detection and classification in multimodal datasets.

6.1.4 Analysis Summary

The proposed MultiLate classifier significantly enhances performance metrics when benchmarked against baseline models. The MultiLate classifier achieves superior precision, recall, text and image modalities and is pivotal in improving its heightened efficacy in accurately detecting and categorizing hate speech.

The minimal variance in metrics between training and testing datasets, typically within a 1-2% margin, underscores the model's robust generalization capabilities and reliability across different data subsets. Incorporating multimodal data, leveraging both text and image modalities, plays a pivotal role in this improved performance.

The text modality benefits from advanced feature extraction techniques, including 3WQA (Who, What, and Why Question Answering), for reasoning and sentiment analysis, enriching the context and interpretability of the text data. Using DAAM-generated heatmaps linked with corresponding sentence words for the image modality ensures that the most salient features are emphasized, enhancing the classifier's accuracy and computational efficiency.

Fusing these sophisticated feature extraction techniques within the multimodal framework enables the MultiLate classifier to better understand the input data, significantly elevating its performance metrics. This methodological advancement not only boosts the classifier's ability to generalize across varied datasets but also positions it as a

Table 8. Performance comparison of different models on hate speech classification

Baseline	Modality	Training			Testing		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
CNN-BiLSTM	Binary	0.72	0.71	0.72	0.71	0.70	0.70
	Multiclass	0.59	0.58	0.59	0.59	0.59	0.58
VGG16	Binary	0.62	0.60	0.61	0.62	0.60	0.61
	Multiclass	0.41	0.40	0.41	0.38	0.40	0.39
ResNet50	Binary	0.62	0.61	0.61	0.60	0.60	0.61
	Multiclass	0.41	0.40	0.41	0.39	0.40	0.40
MULTIMODAL	Binary	0.69	0.68	0.69	0.68	0.68	0.68
	Multiclass	0.53	0.51	0.53	0.52	0.52	0.52
Proposed CNN-BiLSTM	Binary	0.79	0.78	0.79	0.78	0.78	0.78
	Multiclass	0.64	0.64	0.63	0.64	0.62	0.61
Proposed ResNet50	Binary	0.78	0.77	0.74	0.72	0.72	0.72
	Multiclass	0.64	0.64	0.63	0.61	0.62	0.60
Proposed MULTIMODAL	Binary	0.75	0.74	0.75	0.74	0.74	0.74
	Multiclass	0.70	0.69	0.70	0.68	0.68	0.68

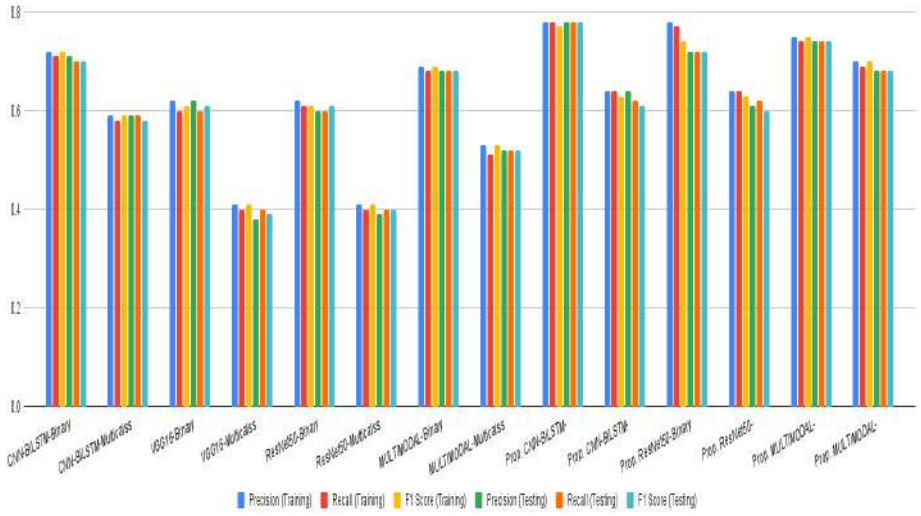


Fig. 9. Statistics on the baseline and proposed models on Multilate dataset

highly effective tool in the fight against hate speech in diverse online environments.

6.2 Performance Analysis on MultiOFF Dataset Using proposed Multilate classifier

The performance analysis of the proposed MultiLate classifier on the MultiOFF dataset demonstrates its effectiveness and superiority over several SOTA models. The detailed performance metrics, including precision, recall, and F1 score, are summarized in the following table 10. For text classification, the proposed model achieves a precision of 0.63, a recall of 0.61, and an F1 score of 0.62.

Table 9. K-Fold Validation Results for Proposed Model

Proposed Model	Folds	Precision	Recall	F1 Score
Text	Binary 1	0.77	0.78	0.77
	Binary 2	0.80	0.80	0.80
	Binary 3	0.77	0.75	0.75
	Binary 4	0.79	0.78	0.79
	Binary 5	0.77	0.78	0.76
	Binary Avg	**0.79**	**0.78**	**0.79**
	Multiclass 1	0.62	0.62	0.62
	Multiclass 2	0.64	0.64	0.64
	Multiclass 3	0.65	0.63	0.63
	Multiclass 4	0.62	0.63	0.62
	Multiclass 5	0.68	0.66	0.66
	Multiclass Avg	**0.64**	**0.64**	**0.63**
Image	Binary 1	0.69	0.70	0.64
	Binary 2	0.83	0.83	0.81
	Binary 3	0.78	0.77	0.75
	Binary 4	0.80	0.76	0.72
	Binary 5	0.81	0.79	0.76
	Binary Avg	**0.78**	**0.77**	**0.74**
	Multiclass 1	0.47	0.45	0.45
	Multiclass 2	0.69	0.69	0.69
	Multiclass 3	0.61	0.60	0.60
	Multiclass 4	0.69	0.69	0.69
	Multiclass 5	0.75	0.74	0.74
	Multiclass Avg	**0.64**	**0.64**	**0.63**

Table 10. Performance Analysis on MultiOFF Dataset Using Proposed MultiLate Classifier

Modality	Classifier	Precision	Recall	F1 Score
Text	LR	0.58	0.40	0.48
	NB	0.52	0.45	0.49
	DNN	0.47	0.54	0.50
	Stacked LSTM	0.39	0.42	0.40
	BiLSTM	0.42	0.23	0.30
	CNN	0.39	0.84	0.54
	Proposed	**0.63**	**0.61**	**0.62**
Image	VGG16	0.41	0.16	0.24
	Proposed	**0.63**	**0.63**	**0.63**
Multi	Stacked LSTM + VGG16 [46]	0.40	0.66	0.50
	BiLSTM + VGG16 [46]	0.40	0.44	0.41
	CNNText + VGG16 [46]	0.38	0.67	0.48
	DisMultiHate [23]	0.645	**0.651**	0.646
	Proposed	**0.651**	0.648	**0.650**

The training accuracy and loss of the model can be seen in figure 17. These results surpass traditional classifiers such as Logistic Regression (LR), Naive Bayes (NB), and advanced deep learning models like DNN and Stacked LSTM. The proposed model's performance is particularly

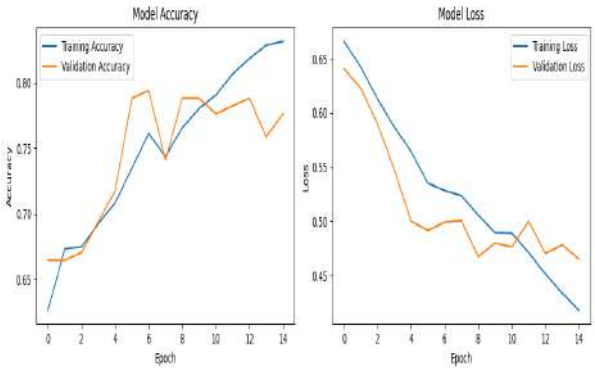


Fig. 10. Training accuracy and loss of text classification in binary setting

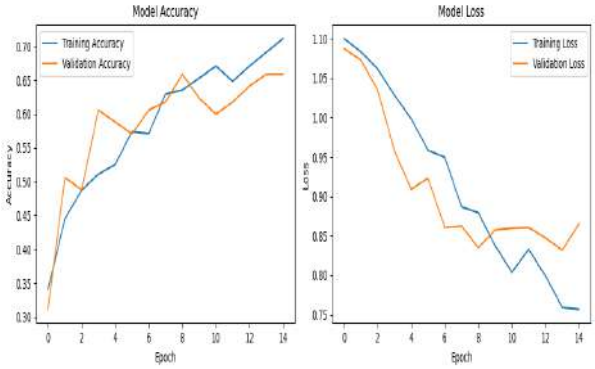


Fig. 11. Training accuracy and loss of text classification in multiclass setting

noteworthy compared to CNN and BiLSTM models, which demonstrate superior balance in precision and recall.

The proposed model significantly outperforms the VGG16 model in the image classification task, achieving a precision, recall, and F1 score of 0.63 across all metrics, the training accuracy and loss of the image classification can be seen in the figure 18. This highlights the efficacy of the proposed model's feature extraction and processing techniques in handling visual data.

When evaluating multi-modal classification, the proposed MultiLate classifier shows a marked improvement over combined models such as Stacked LSTM + VGG16, BiLSTM + VGG16, and CNNText + VGG16.

The proposed model attains a precision of 0.651, recall of 0.648, and F1 score of 0.650, outperforming the DisMultiHate model, one of the SOTA models in this domain. The confusion matrix of both text and image classification on the multiOFF dataset is shown in figure 19a for text and figure 19b for the image.

6.3 Limitations

Although the proposed MultiLate classifier shows important improvements in multimodal hate speech classification, mentioning the following limitations is worthwhile. The first potential pitfall is that the given model may be computationally expensive due to the design of the computational graph. Compound Word2Vec and CNN-BiLSTM with Textual Features complemented by ResNet50 with Visual Features necessitate computing and memory resources in the form of DAAMs. This can be problematic, especially for real-time applications or environments with low processing power. Moreover, the training of the MultiLate classifier strongly depends on the quality and the proportion of training data. It must also be noted that the data used in this work is rather large, but it does not capture all possible types of hate speech in real-life conditions. This limitation is mainly applicable when the training dataset draws only from a sample population or within a limited domain, which may limit the ability of the model to work accurately with different data from different diverse or underrepresented populations.

However, it is imperative to note that integrating deeper feature fusion procedures such as the 3WQA and sentiment analysis in the model's working also adds complexity and may bring interpretability issues. Although these techniques improve the model's performance in detecting hate speech, they also obscure the steps or the decision-making process, which is important in achieving a more favourable opinion on using artificial intelligence. Another issue that has been identified concerns the vulnerability of the model to adversarial perturbations. The slightest input change can alter the model's classification even when the input data are textural or actual images. This weakness highlights the

(a) Binary. **(b)** Multiclass.

Fig. 12. Confusion matrix of both binary and multiclass of the test set in text

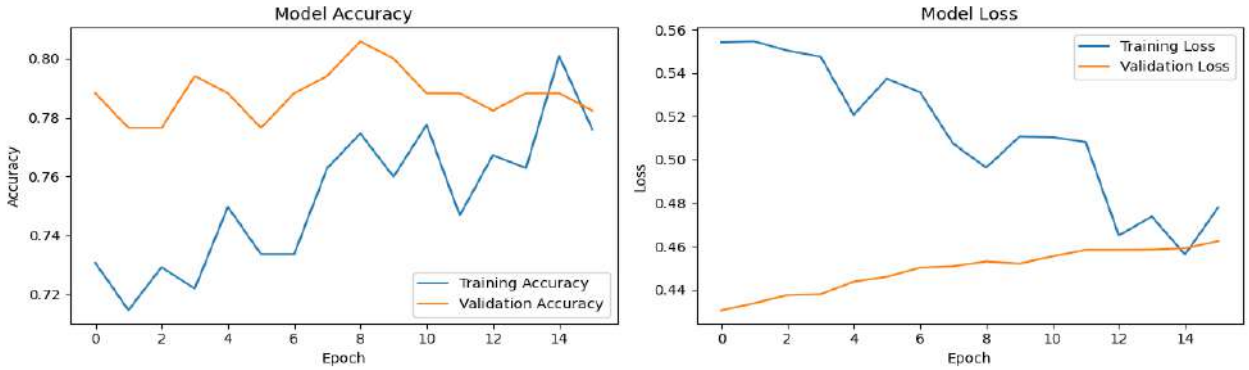


Fig. 13. Training accuracy and loss of image classification in binary setting

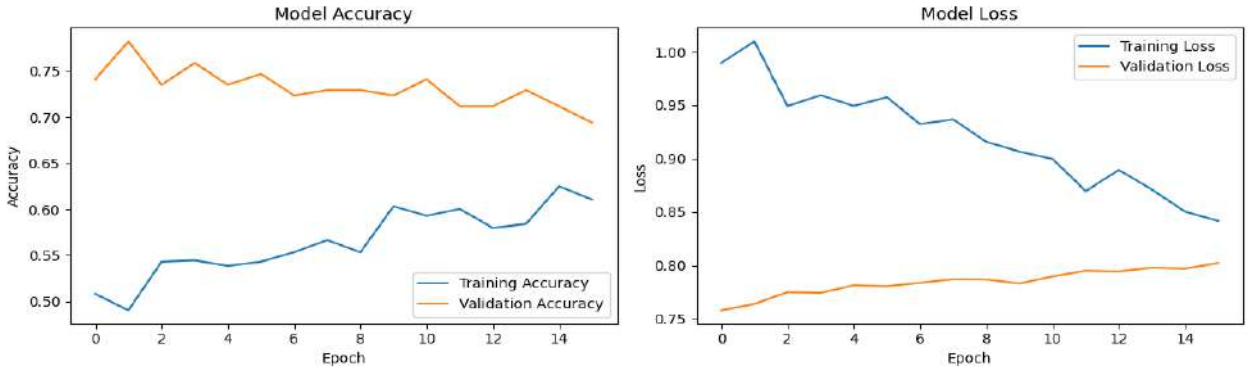


Fig. 14. Training accuracy and loss of image classification in multiclass setting

importance of having strong protective measures to maintain reliability in practical use. Lastly, while the proposed model's efficiency is improved for classifying multimedia content, there is still some potential for improvement in processing and analysis. While the current approach to solving this problem appears to be sufficient, there is room for further optimization, which involves improving the algorithm to reduce the computational time required while maintaining accuracy.

(a) Binary. **(b)** Multiclass.

Fig. 15. Confusion matrix of both binary and multiclass of the test set in the image

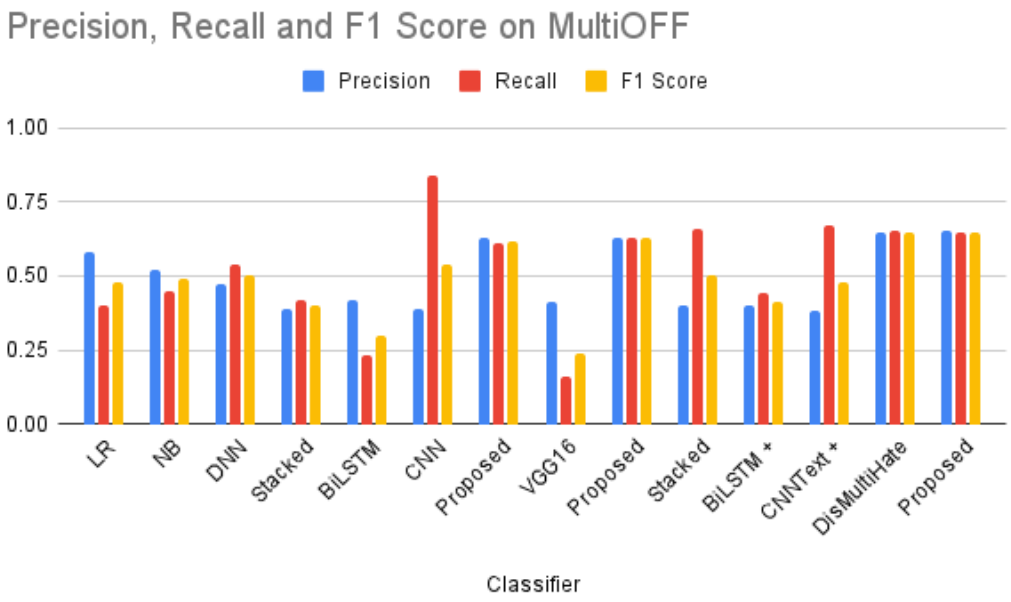


Fig. 16. State-of-the-art comparison of the proposed models on MultiOFF dataset

7 Conclusion

The proposed MultiLate classifier significantly advances detecting multimodal hate speech by effectively integrating textual and visual data. This classifier demonstrates robust performance across multiple datasets by applying a CNN-BiLSTM framework for textual analysis and a ResNet50 architecture for visual feature extraction, enhanced by DAAMs. Incorporating advanced feature fusion techniques, such as 3WQA and sentiment analysis, further strengthens the model's capability to accurately identify hate speech.

By selectively processing the most contextually relevant heatmaps generated from DAAMs, the proposed approach improves classification accuracy and reduces computational overhead, making it more efficient. Extensive evaluations on the MULTILATE and MultiOFF datasets highlight the superior performance of the MultiLate classifier compared to baseline models.

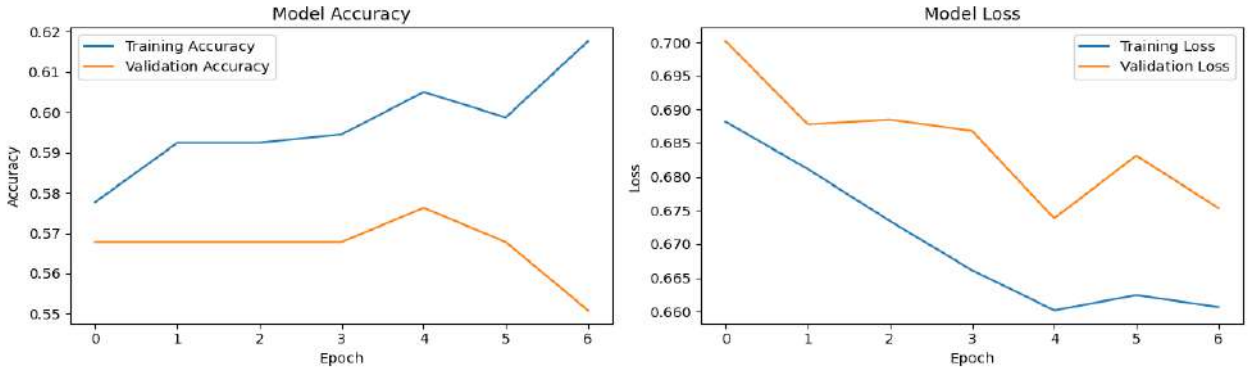


Fig. 17. Training accuracy and loss of text classification in MultiOFF dataset

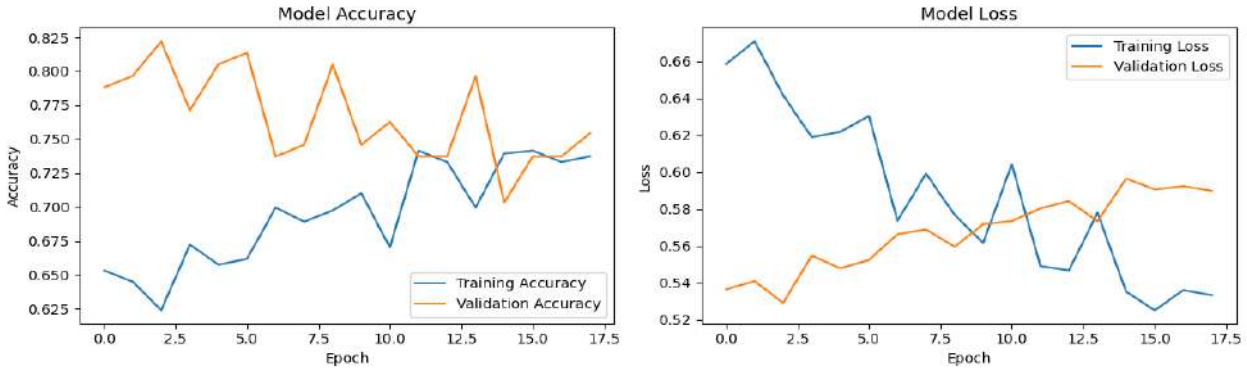


Fig. 18. Training accuracy and loss of image classification in MultiOFF dataset

(a) Text. **(b)** Image.

Fig. 19. Confusion matrix of both text and image classification on MultiOFF dataset

The model consistently achieves higher precision, recall, and F1 scores, underscoring its robustness and generalization capabilities. Comparative analysis with state-of-the-art models reaffirms the efficacy of the proposed methodology in handling complex and varied instances of hate speech.

In conclusion, the MultiLate classifier offers a comprehensive solution to multimodal hate speech detection challenges, combining sophisticated

neural architectures with innovative feature fusion techniques.

This framework not only enhances detection accuracy but also optimizes the utilization of computational resources, providing a practical and scalable approach for real-world applications in online content moderation. Future work will further refine the model and explore its applicability to other domains requiring multimodal analysis.

Acknowledgments

We extend our gratitude to the Department of Computer Science and Engineering (CSE) at the National Institute of Technology Silchar for allowing us to conduct our research and experimentation in the CNLP and AI laboratories.

We deeply appreciate the supportive research environment that fosters and enriches our academic endeavours.

References

1. **Agarwal, S., Sonawane, A., Chowdary, C. R. (2023).** Accelerating automatic hate speech detection using parallelized ensemble learning models. Expert Systems with Applications, Vol. 230, pp. 120564. DOI: https://doi.org/10.1016/j.eswa.2023.120564.
2. **Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., et al. (2012).** The'k'in k-fold cross validation. ESANN, Vol. 102, pp. 441–446.
3. **Arya, G., Hasan, M. K., Bagwari, A., Safie, N., Islam, S., Ahmed, F. R. A., De, A., Khan, M. A., Ghazal, T. M. (2024).** Multimodal hate speech detection in memes using contrastive language-image pre-training. IEEE Access.
4. **Asogwa, D. C., Chukwuneke, C. I., Ngene, C., Anigbogu, G. (2022).** Hate speech classification using svm and naive bayes. arXiv preprint arXiv:2204.07057.
5. **Baccianella, S., Esuli, A., Sebastiani, F., et al. (2010).** Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. Lrec, Valletta, Vol. 10, No. 2010, pp. 2200–2204.
6. **Badour, J., Brown, J. A. (2021).** Hateful memes classification using machine learning. 2021 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, pp. 1–8.
7. **Balouchzahi, F., Butt, S., Sidorov, G., Gelbukh, A. (2023).** Reddit: Regret detection and domain identification from text. Expert Systems with Applications, Vol. 225, pp. 120099. DOI: https://doi.org/10.1016/j.eswa.2023.120099.
8. **Baranwal, A., Gohil, V., Dahat, H., Salunke, A. (2024).** Hate speech and nsfw image classification using bert and resnet-34 model. 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), IEEE, pp. 388–394.
9. **Beniwal, R., Saraswat, P. (2024).** A hybrid bert-cpso model for multi-class depression detection using pure hindi and hinglish multimodal data on social media. Computers and Electrical Engineering, Vol. 120, pp. 109786. DOI: https://doi.org/10.1016/j.compeleceng.2024.109786.
10. **Cao, R., Lee, R. K.-W., Chong, W.-H., Jiang, J. (2022).** Prompting for multimodal hateful meme classification. **Goldberg, Y., Kozareva, Z., Zhang, Y.**, editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 321–332. DOI: 10.18653/v1/2022.emnlp-main.22.
11. **Caselli, T., Basile, V., Mitrović, J., Granitzer, M. (2021).** HateBERT: Retraining BERT for abusive language detection in English. **Mostafazadeh Davani, A., Kiela, D., Lambert, M., Vidgen, B., Prabhakaran, V., Waseem, Z.**, editors, Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational

Linguistics, Online, pp. 17–25. DOI: 10.18653/v1/2021.woah-1.3.

12. **Chauhan, D. S., Singh, G. V., Arora, A., Ekbal, A., Bhattacharyya, P. (2022).** An emoji-aware multitask framework for multimodal sarcasm detection. Knowledge-Based Systems, Vol. 257, pp. 109924. DOI: https://doi.org/10.1016/j.knosys.2022.109924.
13. **Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J. (2020).** Uniter: Universal image-text representation learning. European conference on computer vision, Springer, pp. 104–120.
14. **Djuric, N., Zhou, J., Morris, R. K., Grbovic, M., Radosavljević, V., Bhamidipati, N. (2015).** Hate speech detection with comment embeddings. DOI: 10.1145/2740908.2742760.
15. **Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., Liu, J. (2020).** Large-scale adversarial training for vision-and-language representation learning. Advances in Neural Information Processing Systems, Vol. 33, pp. 6616–6628.
16. **Gomez, R., Gibert, J., Gomez, L., Karatzas, D. (2020).** Exploring hate speech detection in multimodal publications. Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 1470–1478.
17. **Jaafar, N., Lachiri, Z. (2023).** Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. Expert Systems with Applications, Vol. 211, pp. 118523. DOI: https://doi.org/10.1016/j.eswa.2022.118523.
18. **Kalita, G., Halder, E., Taparia, C., Vetagiri, A., Pakray, P. (2023).** Examining hate speech detection across multiple indo-aryan languages in tasks 1 & 4. FIRE (Working Notes), pp. 474–485.
19. **Karayiğit, H., Çiğdem İnan Acı, Akdağlı, A. (2021).** Detecting abusive instagram comments in turkish using convolutional neural network and machine learning methods. Expert Systems with Applications, Vol. 174, pp. 114802. DOI: https://doi.org/10.1016/j.eswa.2021.114802.
20. **Khan, M. S., Malik, M. S. I., Nadeem, A. (2024).** Detection of violence incitation expressions in urdu tweets using convolutional neural network. Expert Systems with Applications, Vol. 245, pp. 123174. DOI: https://doi.org/10.1016/j.eswa.2024.123174.
21. **Kibriya, H., Siddiqa, A., Khan, W. Z., Khan, M. K. (2024).** Towards safer online communities: Deep learning and explainable ai for hate speech detection and classification. Computers and Electrical Engineering, Vol. 116, pp. 109153. DOI: https://doi.org/10.1016/j.compeleceng.2024.109153.
22. **Kumar, G. K., Nandakumar, K. (2022).** Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. arXiv preprint arXiv:2210.05916.
23. **Lee, R. K.-W., Cao, R., Fan, Z., Jiang, J., Chong, W.-H. (2021).** Disentangling hate in online memes. Proceedings of the 29th ACM international conference on multimedia, pp. 5138–5147.
24. **Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., Sun, G. (2018).** xdeepfm: Combining explicit and implicit feature interactions for recommender systems. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1754–1763.
25. **Liu, S., Bremer, P.-T., Thiagarajan, J. J., Srikumar, V., Wang, B., Livnat, Y., Pascucci, V. (2017).** Visual exploration of semantic relationships in neural word embeddings. IEEE transactions on visualization and computer graphics, Vol. 24, No. 1, pp. 553–562.
26. **Mozafari, M., Farahbakhsh, R., Crespi, N. (2020).** Hate speech detection and racial bias mitigation in social media based on bert model. PloS one, Vol. 15, No. 8, pp. e0237861.
27. **Muennighoff, N. (2020).** Vilio: State-of-the-art visio-linguistic models applied to hateful memes. arXiv preprint arXiv:2012.07788.

28. **Mullah, N. S., Zainon, W. M. N. W. (2021).** Advances in machine learning algorithms for hate speech detection in social media: A review. DOI: 10.1109/access.2021.3089515.

29. **Nahin, A. S. M., Roza, I. I., Nishat, T. T., Sumya, A., Bhuiyan, H., Hoque, M. M. (2024).** Bengali hateful memes detection: A comprehensive dataset and deep learning approach. 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), IEEE, pp. 01–06.

30. **Niu, Z., Zhong, G., Yu, H. (2021).** A review on the attention mechanism of deep learning. Neurocomputing, Vol. 452, pp. 48–62. DOI: https://doi.org/10.1016/j.neucom.2021.03.091.

31. **Pennington, J., Socher, R., Manning, C. (2014).** GloVe: Global vectors for word representation. **Moschitti, A., Pang, B., Daelemans, W.**, editors, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

32. **Putra, B. P., Irawan, B., Setianingsih, C., Rahmadani, A., Imanda, F., Fawwas, I. Z. (2022).** Hate speech detection using convolutional neural network algorithm based on image. 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), IEEE, pp. 207–212.

33. **Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., Zhang, R., Zhou, M. (2020).** ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. **Cohn, T., He, Y., Liu, Y.**, editors, Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, pp. 2401–2410. DOI: 10.18653/v1/2020.findings-emnlp.217.

34. **Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021).** Learning transferable visual models from natural language supervision. International Conference on Machine Learning, PMLR, pp. 8748–8763.

35. **Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J. (2020).** Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, Vol. 21, No. 1, pp. 5485–5551.

36. **Rajput, G., Punn, N. S., Sonbhadra, S. K., Agarwal, S. (2021).** Hate speech detection using static bert embeddings. Big Data Analytics: 9th International Conference, BDA 2021, Virtual Event, December 15-18, 2021, Proceedings 9, Springer, pp. 67–77.

37. **Rana, A., Jha, S. (2022).** Emotion based hate speech detection using multimodal learning. ArXiv, Vol. abs/2202.06218.

38. **Rani, A., Tonmoy, S. T. I., Dalal, D., Gautam, S., Chakraborty, M., Chadha, A., Sheth, A., Das, A. (2023).** FACTIFY-5WQA: 5W aspect-based fact verification through question answering. **Rogers, A., Boyd-Graber, J., Okazaki, N.**, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, pp. 10421–10440. DOI: 10.18653/v1/2023.acl-long.581.

39. **Rao, S., Verma, A. K., Bhatia, T. (2021).** A review on social spam detection: Challenges, open issues, and future directions. Expert Systems with Applications, Vol. 186, pp. 115742. DOI: https://doi.org/10.1016/j.eswa.2021.115742.

40. **Redmon, J., Farhadi, A. (2017).** Yolo9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271.

41. **Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2021).** High-resolution image synthesis with latent diffusion models.

42. **Saad, W., Shalaby, W. A., Shokair, M., El-Samie, F. A., Dessouky, M., Abdellatef,**

E. (2022). Covid-19 classification using deep feature concatenation technique. Journal of Ambient Intelligence and Humanized Computing, pp. 1–19.

43. **Sethi, A., Kuchhal, U., Katarya, R., et al. (2021).** Study of various techniques for the classification of hateful memes. 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), IEEE, pp. 675–680.
44. **Song, X., Xu, D., Peng, C., Zhang, Y., Xue, Y. (2024).** A two-stage frequency-domain generation algorithm based on differential evolution for black-box adversarial samples. Expert Systems with Applications, Vol. 249, pp. 123741. DOI: https://doi.org/10.1016/j.eswa.2024.123741.
45. **Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J. (2019).** Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.
46. **Suryawanshi, S., Chakravarthi, B. R., Arcan, M., Buitelaar, P. (2020).** Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. **Kumar, R., Ojha, A. K., Lahiri, B., Zampieri, M., Malmasi, S., Murdock, V., Kadar, D.**, editors, Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, pp. 32–41.
47. **Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F. (2023).** What the DAAM: Interpreting stable diffusion using cross attention. **Rogers, A., Boyd-Graber, J., Okazaki, N.**, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, pp. 5644–5659. DOI: 10.18653/v1/2023.acl-long.310.
48. **Tang, R., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Lin, J., Ture, F. (2022).** What the daam: Interpreting stable diffusion using cross attention. arXiv preprint arXiv:2210.04885.
49. **Velioglu, R., Rose, J. (2020).** Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. arXiv preprint arXiv:2012.12975.
50. **Vetagiri, A., Adhikary, P., Pakray, P., Das, A. (2023).** CNLP-NITS at SemEval-2023 task 10: Online sexism prediction, PREDHATE! Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, pp. 815–822. DOI: 10.18653/v1/2023.semeval-1.113.
51. **Vetagiri, A., Adhikary, P. K., Pakray, P., Das, A. (2023).** Leveraging gpt-2 for automated classification of online sexist content. Working Notes of CLEF, pp. 1107–1122.
52. **Vetagiri, A., Halder, E., Das Majumder, A., Pakray, P., Das, A. (2024).** Multilate: A synthetic dataset on ai-generated multimodal hate speech. Proceedings of the 21th International Conference on Natural Language Processing (ICON), NLP Association of India (NLPAI), Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India, pp. .
53. **Vetagiri, A., Pakray, P., Das, A. (2024).** A deep dive into automated sexism detection using fine-tuned deep learning and large language models. Available at SSRN 4791798.
54. **Wu, C. S., Bhandary, U. (2020).** Detection of hate speech in videos using machine learning. 2020 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, pp. 585–590.
55. **Wullach, T., Adler, A., Minkov, E. (2021).** Towards hate speech detection at large via deep generative modeling. IEEE Computer Society, Vol. 25, No. 2, pp. 48–57. DOI: 10.1109/mic.2020.3033161.
56. **Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., Han, X. (2024).** Ai for social science and social science of ai: A survey. Information Processing & Management, Vol. 61,

No. 3, pp. 103665. DOI: https://doi.org/10.1016/j.ipm.2024.103665.

57. **Yang, J., Yang, Z., Zhang, S., Tu, H., Huang, Y. (2021).** Sesy: Linguistic steganalysis framework integrating semantic and syntactic features. IEEE Signal Processing Letters, Vol. 29, pp. 31–35.
58. **Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., Wang, H. (2021).** Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. Proceedings of the AAAI conference on artificial intelligence, Vol. 35, No. 4, pp. 3208–3216.
59. **Zhang, W., Yoshida, T., Tang, X. (2011).** A comparative study of tf*idf, lsi and multi-words for text classification. Expert Systems with Applications, Vol. 38, No. 3, pp. 2758–2765. DOI: https://doi.org/10.1016/j.eswa.2010.08.066.
60. **Zhang, Z., Luo, L. (2019).** Hate speech detection: A solved problem? the challenging case of long tail on twitter. IOS Press, Vol. 10, No. 5, pp. 925–945. DOI: 10.3233/sw-180338.
61. **Zhong, S., Scarinci, A., Cicirello, A. (2023).** Natural language processing for systems engineering: Automatic generation of systems modelling language diagrams. Knowledge-Based Systems, Vol. 259, pp. 110071. DOI: https://doi.org/10.1016/j.knosys.2022.110071.
62. **Zhou, Y. (2020).** A review of text classification based on deep learning. Proceedings of the 2020 3rd international conference on geoinformatics and data analysis, pp. 132–136.
63. **Zhu, R. (2020).** Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. arXiv preprint arXiv:2012.08290.

Article received on 01/12/2024; accepted on 22/01/2025.
**Corresponding author is Partha Pakray.*