MULTILATE: A Synthetic Dataset on AI-Generated MULTImodaL hATE Speech

Advaitha Vetagiri[†1], Eisha Halder[1], Ayanangshu Das Majumder[1], Partha Pakray[1], Amitava Das[2,3]
[1] National Institute of Technology Silchar, Silchar, Assam, India, 788010.
[2] Artificial Intelligence Institute of UofSC, Columbia, South Carolina, USA.
[3] Wipro AI Lab, Bangalore, Karnataka, India.
[†]advaitha21_rs@cse.nits.ac.in

Abstract

One of the pressing challenges society faces today is the rapid proliferation of online hate speech, exacerbated by the rise of AI-generated multimodal hate content. This new form of synthetically produced hate speech presents unprecedented challenges in detection and moderation. In response to the growing presence of such harmful content across social media platforms, this research introduces a groundbreaking solution: *"MULTILATE"*. This initiative represents a concerted effort to develop scalable, multimodal hate speech detection systems capable of navigating the increasingly complex digital landscape. It contains 2.6 million text samples designed to classify multimodal hate speech, and these text-based statements are used to generate AI images created through Stable Diffusion. The dataset features pixel-level temperature maps, which are crucial for understanding the nuanced relationship between textual and visual components, thereby enhancing the interpretability of hate speech detection models. Additionally, MULTILATE includes 3W Question-Answer pairs that address the *"who"*, *"what"*, and *"why"* aspects of hate speech, providing deeper insights into the motivations and contexts behind such content. To further strengthen detection capabilities, the dataset also incorporates adversarial examples across textual and visual domains, ensuring robustness against adversarial attacks and enhancing the reliability of multimodal hate speech detection systems.

1 Introduction

A prevalent sociological problem currently is online hate speech, where Meta (*formerly Facebook*) has reported removed 18 million hate content articles [1] in the second quarter of 2023 which is more than the 10.7 million ones it deleted during the first quarter of 2023. Between April and June 2021, Meta took down more than 31 million posts containing hate speech. The spread of this hateful speech results in considerable emotional anguish, particularly within vulnerable minority groups, thus normalizing prejudice (Wachs et al., 2022). A new development to this is the continued appearance of hate speech generated by artificial intelligence (AI) (Xu et al., 2024), which is even more problematic regarding recognition and elimination. With the help of many new sophisticated language models, it is possible to create essentially fake texts that cannot be distinguished from the texts produced by humans, thus propagating hatred ideologies quickly and on a large scale. In addition, the complexity of AI-generated posts can outsmart existing content moderation techniques and ultimately let toxic messages spread. One egregious example involves a historical speech by Adolf Hitler, altered by AI to deliver antisemitic remarks in English. This manipulated video, shared by an influencer, quickly garnered over 15 million views on X (*formerly Twitter*) in March 2024 [2]. Such incidents underscore the growing concern among researchers and monitoring organizations about the proliferation of AI-generated hate and the urgent need for critical evaluation and robust detection mechanisms to combat this emerging threat.

To cater for this, new and improved detection mechanisms must be employed on social media platforms to teach the system to detect AI-driven hate speech along with regular human moderation and partnering with organizations focusing on AI ethics and safety online. The fight against online hate thus depends on how these new technological forms are met and how a good containment approach is developed that protects threatened groups against the emerging threat of hate speech via Artificial intelligence.

Researchers have recently emphasized the cre-

[1] https://www.statista.com/statistics/1013804/facebook-hate-speech-content-deletion-quarter/

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

ation of multimodal hate speech detection systems that can perform at large scales, especially on platforms such as Meta, X, and Youtube (Gomez et al., 2020). Nevertheless, the development of AI-generated hate has been limited due to the absence of large-scale cross-modal hate speech datasets specifically designed for Human-generated hate. To circumvent this weakness, a novel multimodal dataset named *"MULTILATE"* is proposed to facilitate the mass-scale assessment of AI-generated multimodal hate speech. This dataset consists of 2.6 million instances of text, and these text-based statements are used to generate AI images created through Stable Diffusion (SD) (Rombach et al., 2021). Further, every instance includes Pixel-Level HeatMaps for visual interpretability and Question-Answer (QA) pairs, which incorporate "*who*", "*what*", and "*why*" for explainability. Adversarial examples of text (Morris et al., 2020) and images (Deng and Karam, 2020) are also included to promote more robust multimodal hate speech detection in the field of growing significance at the interface of Natural Language Processing (NLP) and Artificial intelligence.

- **Large-scale AI-generated multimodal hate speech dataset with 2.6 million samples**: This work proposed a dataset of 2.6 million samples of text image modalities with their supporting documentation as in heat maps for the *"Images"* and QA pairs for the *"Text"*.

- **Incorporates AI-generated explainability**: Pixel-level heatmaps are generated for every token in the text, offering detailed visual insights into the model's focus areas during classification. Additionally, a 3W Question-Answering (QA) system is implemented to address "who," "what," and "why" for each text instance, ensuring that the model's predictions are both accurate and interpretable.

- **Includes adversarial examples for robustness**: Adversarial text and images are included to enhance the system's resilience against real-world examples. These examples test the model's limits and improve its generalization by revealing "blind spots", ultimately leading to more robust detection of multimodal hate speech.

2 Related Work

Hate speech refers to discrimination due to race, ethnic background, religion, gender, and sexual orientation. It has severe consequences, including prejudice and violence in society. The classification of hate speech has mainly involved machine learning models such as the Support Vector Machine (SVM) and Random Forest (Chhabra and Vishwakarma, 2023; MacAvaney et al., 2019). However, these challenges remain like conflict or overlapping definitions of emotions, availability of datasets, and algorithmic methodology (Chiril et al., 2022).

The widespread nature of online sexism has made researchers interested in sexism classification and has subsequently led to the emergence of automated recognition technologies. In studies, sexism is identified using deep learning (DL) architectures such as convolutional neural network (CNN) and Bidirectional Encoder Representations from Transformers (BERT) applied in social media conversations (Sharifirad and Jacovi, 2019; Rodríguez-Sánchez et al., 2020; Chiril et al., 2021; Vetagiri et al., 2023b). The Generation of datasets like TOXIGEN (Hartvigsen et al., 2022) helps in the improvement of toxic language detection calling for massive and uniform datasets. Moreover, there are developments on sexism detection in machine learning using data augmentation methods and ensembles of state-of-the-art language embeddings like BERT or Roberta (Ahuir et al., 2022).

Furthermore, the research explores the application of DL models such as BiLSTMs, BERT, and GPT-2 in sexism classification, demonstrating promising results (Abburi et al., 2021; Rodríguez-Sánchez et al., 2020; Vetagiri et al., 2023a). Challenges of resource-constrained languages such as Urdu hate speech for detection; traditional models outperform DL-based approach owing to class imbalance and data scarcity, a case study (Saeed et al., 2023). However, the concluding remarks emphasize the need for future work that addresses the challenges of improving current models' discrimination capabilities and exploring user-based features (Ahuir et al., 2022; Huang et al., 2022).

(Gomez et al., 2019) proposed the new problem of multi-modal hate speech detection with text and image. They constructed the MMHS150k dataset for annotated tweet images. They try out textual kernel-based fusion approaches such as (Gao et al., 2018) among other unimodal and multimodal models and showcase that images are helpful sources

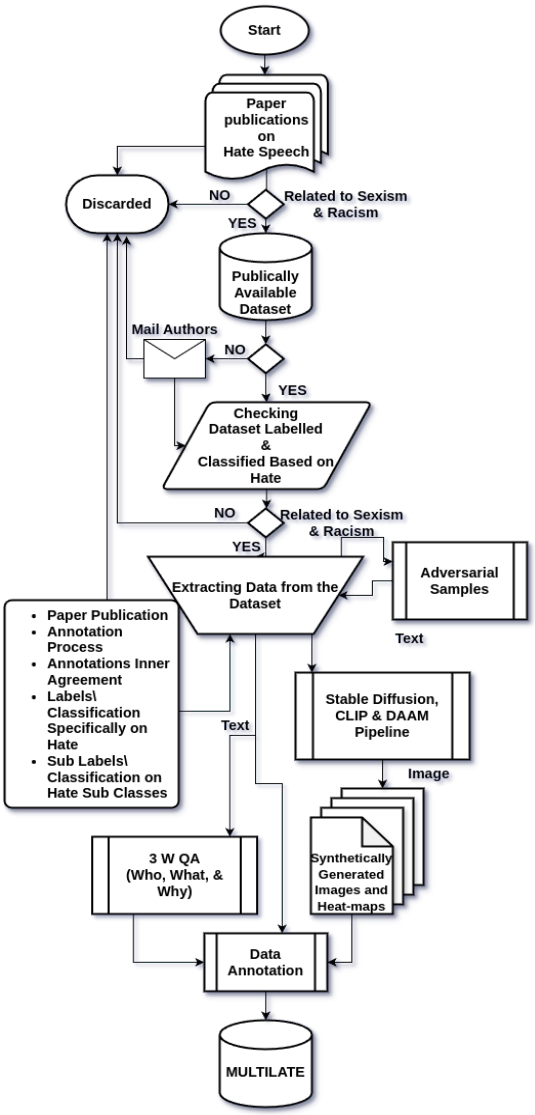


Figure 1: Detailed representation of the MULTILATE dataset creation pipeline. The process begins with the extensive search and acquisition of hate speech datasets from public and private sources, focusing on datasets labelled sexist and racist. The data undergoes a rigorous selection and screening phase to filter out irrelevant datasets. The flow continues with applying classification algorithms and integrating Stable Difussion, pixel-level heatmaps, 3W QA systems, and adversarial examples.

of information. Nevertheless, it is incredibly challenging in terms of data as well as the multimodal nature of the problem. However, concurrent modelling of the textual and visual information presents a potential for detecting hate speech as a critical open area for supporting content moderation. They generally create the beginnings of multimodal hatred utterance research on the study grounds.

Lastly (Rani et al., 2023) describes a five-factor, issue-based, question-answering system for a more intelligible explanation of automated fact-checking machines[3]. Using this method, the authors develop the FACTIFY-5WQA dataset of more than 390,000 textual claims in which they label each sentence's five semantics roles and pair them with appropriate questions that can be used as queries. Validated QA pairs are employed to check some elements of specific evidentiary documents for precise identification of falsity in claims.

3 Data

This section outlines the creation of the dataset called MULTILATE[4]. Specifically designed to identify instances of hate speech, particularly sexism and racism, in online content, MULTILATE is a unique dataset containing a total of 2.6 million samples extracted from 11 different datasets. The dataset includes synthetically generated images created using advanced AI models to enhance its scope and applicability. These AI-generated hate samples are particularly useful for training models that can classify and identify AI-generated hate speech in real-world settings. The dataset features labels for binary classification, such as *Hate"* and *Not Hate"*, as well as multiclass classification labels, including *Sexist"*, *Racist"*, and *"Neither"*. By providing a comprehensive and diverse collection of examples, this dataset is a valuable resource for researchers and developers working on automated techniques for detecting and mitigating hate speech online.

3.1 Data Sourcing

Data availability is a crucial factor that significantly benefits any model's performance. It is well documented that the efficacy of a model trained on a mixture of diverse datasets surpasses that of a model trained on a single dataset (Chiril et al., 2022). In creating the MULTILATE dataset, an extensive search was conducted for published, public, and privately available datasets containing instances of hate speech. As illustrated in Figure 1, this process represents the workflow for developing the dataset, and Table 1 offers valuable insights into the composition of the datasets from which text was extracted (Vetagiri et al.). Moreover, efforts were made to contact the authors of privately available datasets to request access to their data. A total of 69 datasets [5] containing examples of English hate

[3] https://huggingface.co/spaces/Towhidul/5WQA
[4] https://github.com/advaithavetagiri/MULTILATE
[5] https://hatespeechdata.com/

speech were collected. To ensure comprehensive coverage of both sexism and racism, only datasets labelled and classified based on gender, race, ethnicity, sexist-racist slurs, stereotypes, and related features were included. Datasets not meeting these criteria were excluded from the analysis, and a total of 11 datasets were finalized, as shown in figure 1.

3.2 Data Creation

3.2.1 Image - Stable Diffusion

AI-generated posts can evade existing content moderation systems, enabling the unchecked spread of toxic messages. To capture this, SD 2.1 was employed (Rombach et al., 2021) to generate hateful images paired with textual statements. Stable Diffusion's AI-based text-to-image generation capabilities allow the synthesis of diverse visual interpretations of hate speech. SD is an open-source text-to-image model that can generate high-quality images conditioned on textual prompts, and SD 2.1 is one of the newest text-to-image models from StabilityAI. A pipeline has been created that generates three images from the Stable Diffusion model for each text.

Re-ranking of Generated Images: To assess the generated images quantitatively, Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021), a model that scores the images based on how well they match with the text. This CLIP score indicates the match between text encoding and image encoding. Based on the CLIP score, we re-rank the images per prompt and select the top-ranked image as the best visual interpretation of the given hate speech statement.

Pixel-level Image Heatmap: For validating the best-generated image with the corresponding text, the Diffusion Attention Attribution Maps (DAAM) (Tang et al., 2022) is used to create heatmaps that highlight the image's areas corresponding to specific words in the text. This provides visual explainability into the generated multimodal pairing. Moreover, the Multilate pipeline requires a substantial amount of computational power and resources, particularly for AI image creation using models like SD. Due to these high computational demands, the dataset will be released in batches to manage the resource-intensive nature of generating AI-based images. By combining these steps, pairing text with suitable images and detailed heatmaps that show how the words relate to different parts of the image, as illustrated in Figure 2. Figures 3a and 3b are examples of images created from the pipeline for the respective text.

3.2.2 3W Question Answering

To enhance textual explanations for each text, a question-generating module is employed to automatically create "*who*", "*what*", and "*why*" question-answer pairs as demonstrated in related works like (Rani et al., 2023). The process of generating QA pairs involves 3 stages. This process begins with semantic role labelling (SRL), which identifies key phrases in the text that cover the main topics. Then ProphetNet (Qi et al., 2020), a transformer-based model, then uses these identified phrases to generate natural language questions that can be answered based on the content of the text. Finally, a question-answering (QA) model called T5 (Raffel et al., 2020), a transformer model, automatically answers these questions, offering detailed textual explanations about the main actors, situations, and motivations relevant to each text.

3W Semantic Role Labelling: The process of generating QA pairs involves multiple stages and leverages the latest advancements in neural semantic parsing and generative language models. The initial step involves training a neural SRL system to identify text spans corresponding to predefined semantic elements. These identified spans are then mapped onto the "*who*", "*what*", and "*why*" categories using a targeted ontology framework.

Automatic 3W QA Pair Generation: The spans extracted through SRL are combined with the input text and fed into a generative QA model called ProphetNet (Qi et al., 2020). ProphetNet, which employs a unique n-stream self-attention mechanism, allows for advanced planning in predicting future tokens. As an encoder-decoder architecture pre-trained on extensive corpora, ProphetNet generates coherent and well-formed questions that specifically address the "*who*", "*what*", or "*why*" aspects identified in the input text.

QA Pair Answering: A fine-tuned QA version of the T5 (Raffel et al., 2020) transformer model is utilized to answer the generated questions. The T5 model processes the input statements and questions from ProphetNet to produce answers by selecting the most relevant text extracts. Standardized evaluations of multiple answers have demonstrated that the T5 model is the most accurate method for extracting answers. Finally, the QA responses are verified against evidence documents by picking 5000 random texts to ensure the logical consis-

Table 1: Datasets for Sexist and Racist Classification with Adversarial Samples

Dataset	Sexist	Racist	Neither	Extracted
CMSD (Samory et al., 2021)	1809	-	11822	13631
EDOS (Kirk et al., 2023)	15330	-	44670	60000
WSF (de Gibert et al., 2018)	-	1196	9507	10703
ConvAbuse (Cercas Curry et al., 2021)	285	27	671	983
Measuring Hate Speech (Kennedy et al., 2020)	17230	28360	86283	131873
DGHD v0.2.3 (Vidgen et al., 2021b)	3786	5375	18969	28130
HateCheck (Röttger et al., 2021)	1145	757	1242	3144
Nuanced (Borkan et al., 2019)	133152	138966	1264764	1536882
MMHS150K (Gomez et al., 2019)	16243	49906	81074	147223
CAD (Vidgen et al., 2021a)	1352	963	20903	23218
Toxigen (Hartvigsen et al., 2022)	19073	88780	108940	216793
Adversarial Samples	41881	62866	329769	434516
Our Dataset (MULTILATE)	**251286**	**377196**	**1978614**	**2607096**

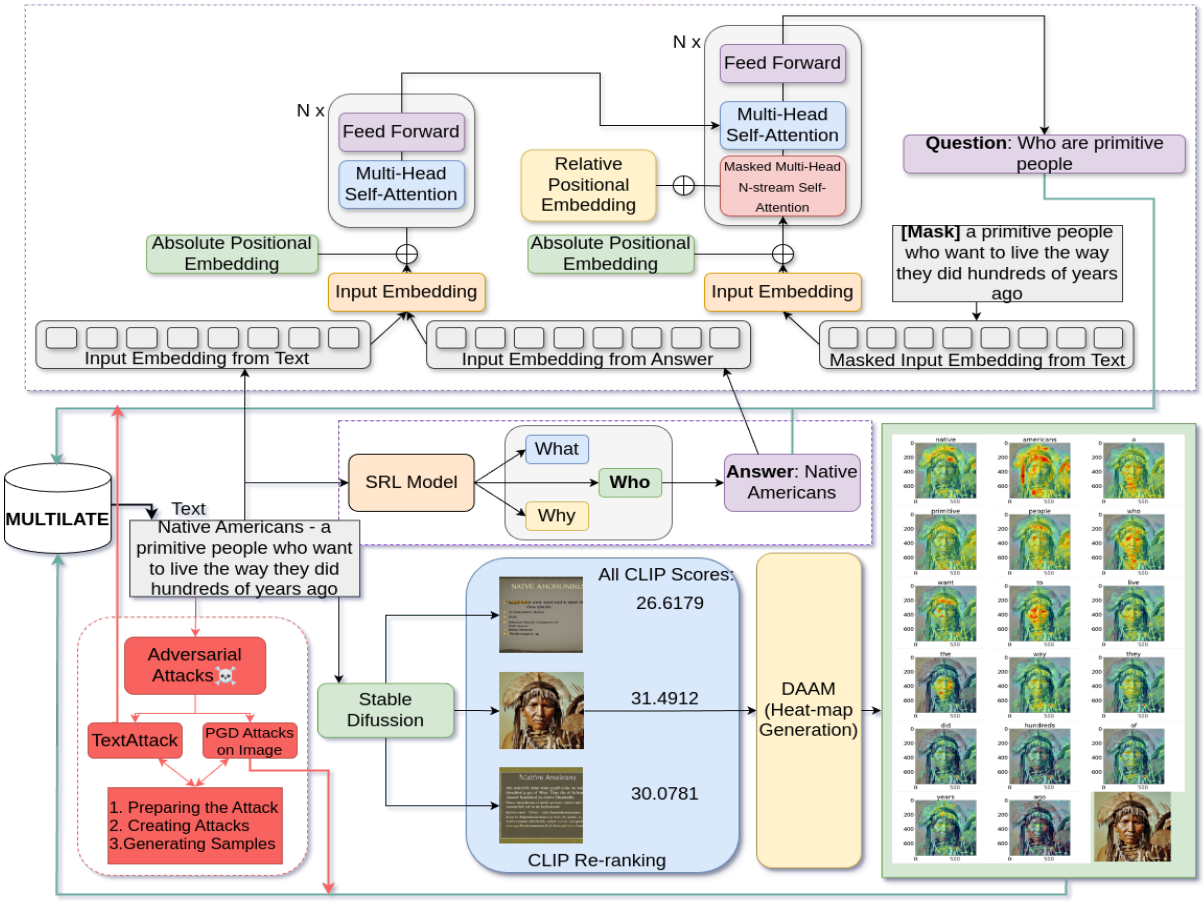
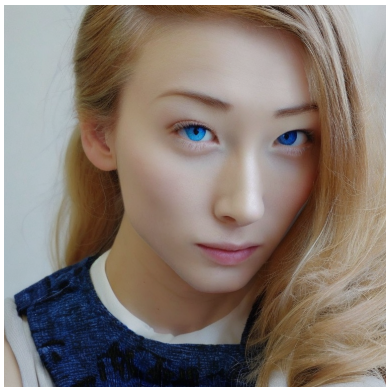
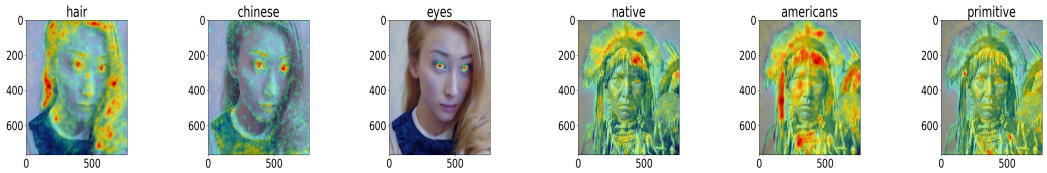


Figure 2: An in-depth overview of the MULTILATE framework, illustrating the integration of Stable Diffusion, SRL, and T5 models, along with an adversarial attack setup, to generate synthetic multimodal hate speech data. The process begins with stable diffusion, which generates synthetic images that capture the essence of various hate speech scenarios. Semantic Role Labeling (SRL) is then applied to the text, extracting and defining the roles and relationships between different textual elements. The T5 model is used to generate text that is coherent and aligned with the images, creating a seamless multimodal dataset. An adversarial attack setup is implemented to further enhance the dataset's applicability and challenge the robustness of detection models, introducing carefully crafted adversarial examples.



(a) An image created for an example text "*how can you be chinese with blond hair and blue eyes* - ***Hate***", using Stable Diffusion.

(b) Another image created for an example text "*Native Americans - a primitive people who want to live the way they did hundreds of years ago,* - ***Hate***".



(c) Heat maps generated for Figure 3a.

(d) Heat maps generated for Figure 3b.

Figure 3: Images generated through Multilate pipeline.

tency of the input statements through human-in-the-loop. The T5 model is then used to generate the final answers by integrating the questions with the extracted evidence snippets into a single input. These final responses and initial answers provide high-resolution insights into which aspects of the statement are supported or contradicted by external data sources.

3.2.3 Adversarial Samples

The dataset includes adversarial samples specifically designed to exploit vulnerabilities in multimodal hate speech classifiers. Unlike conventional methods, the proposed adversarial attack leverages contextual features from both text and images to generate sophisticated perturbations. These samples target cross-modal dependencies, challenging the robustness of the models. By incorporating these adversarial cases, models are tested against realistic and complex scenarios, paving the way for improving their resilience to multimodal adversarial threats.

Adversarial Text Adversarial text is generated for 20% of the MULTILATE dataset using the proposed attack, which integrates contextual features derived from the associated images. The attack identifies high-importance words in the text, guided by DAAMs, to correlate with salient regions in the accompanying image. Subtle perturbations, such as character-level modifications, are applied to these high-priority words to disrupt the model's predictions. These modifications are designed to maintain the semantic integrity and fluency of the text while exploiting the model's reliance on multimodal feature interactions.

Adversarial Images Adversarial images are crafted for 20% of the MULTILATE dataset using a method that utilizes contextual gradients from the text modality to guide pixel-level perturbations in the image. The attack employs a Projected Gradient Descent (PGD) (Deng and Karam, 2020) framework to iteratively modify high-importance regions identified through heatmaps generated by DAAM. These perturbations are imperceptible to the human eye but disrupt the multimodal decision boundaries by aligning noise with the most influential textual features. This targeted manipulation ensures that the adversarial examples exploit the interplay between modalities, rendering the model vulnerable to multimodal attacks.

3.3 Data Annotation

Data annotation is used to ensure the quality and reliability of the dataset. A random selection of 1004 samples from the dataset was chosen for annotation, employing a human-in-the-loop approach. This method involved three expert annotators care-

fully examining each sample to provide accurate labels and annotations. Annotators were trained to identify instances of hate speech, sexism, racism, or other forms of online toxicity present in the samples. Each sample was meticulously reviewed, with annotators providing detailed annotations to capture the nuances and context of the content. Following the annotation process, a thorough comparison was conducted against the original annotations, and the results yielded a Kappa score (McHugh, 2012) of 0.65, validating the accuracy and reliability of annotations. This meticulous verification process underscores the robustness of the dataset, affirming its suitability for training and evaluating hate speech detection models.

3.4 Data Validation

In terms of validation, a subset containing 1004 samples between textual and image data was used to test the accuracy level and reliability consistency for integrity provided in MULTILATE dataset. The validation step includes studying how well this model works on that basis, measuring its accuracy and generalizing characteristics. The validation set's results are carefully evaluated and described in detail in the Results section 5. This specialized subsample allows for a targeted assessment of the dataset's performance in detecting cases of hate speech, particularly about sexism and racism online. By including both the textual and image levels as part of this validation process, assessment contributes to a more complete understanding of whether or not data is suitable for use in training and testing techniques that rely on automation.

4 Baseline Classification Models

Several models were employed to establish baseline performance on the MULTILATE dataset, covering text, image, and multimodal classification. These models include CNN-BiLSTM, ResNet50, BERT, RoBERTa, VGG16, and a basic CNN model.

4.1 Text Classification

CNN-BiLSTM: For text classification, the CNN-BiLSTM (Vetagiri et al., 2024) model combines convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks. It uses GloVe embeddings trained on the MULTILATE dataset to convert text into dense vector representations. The CNN layers capture local text features, while the BiLSTM layers extract sequential information. The final dense layers perform binary classification. Regularization was implemented using a dropout rate of 0.2 to prevent overfitting, and a batch size of 128 was used. The model was evaluated using 5-fold cross-validation to ensure robust performance metrics.

BERT: BERT (Kalita et al., 2023), known as Bidirectional Encoder Representations from Transformers, was used as a text classifier to benchmark the dataset's performance. Fine-tuned on the MULTILATE dataset, BERT leverages its deep bidirectional transformer architecture to understand context and semantics, achieving high accuracy in binary text classification tasks.

RoBERTa: RoBERTa (A Robustly Optimized BERT Pretraining Approach) is another transformer-based model fine-tuned on the MULTILATE dataset for text classification. Similar to BERT, it benefits from a robust training regimen and achieves comparable performance, particularly excelling in capturing nuanced context, resulting in high accuracy for binary classification.

4.2 Image Classification

ResNet50: ResNet50 (Macrayo et al., 2023), a deep convolutional neural network pre-trained on ImageNet, was fine-tuned for image classification on the MULTILATE dataset. Additional dense layers were added to the architecture to improve its discriminative abilities. A dropout rate of 0.2 was applied to the first dense layer and 0.1 to the second, balancing model complexity and overfitting prevention. ResNet50 served as a strong baseline for visual feature extraction, although its performance indicated the need for more specialized architectures for visual hate speech detection.

VGG16: VGG16, a convolutional neural network architecture, was also fine-tuned for image classification tasks on the MULTILATE dataset. Despite being less deep than ResNet50, VGG16 provides a strong baseline for comparison. Its performance, while moderate, helped identify the challenges associated with visual hate speech detection and underscored the need for more tailored image classification models.

CNN: A simple CNN model was implemented as an additional baseline for image classification. Despite its straightforward architecture, the CNN model offered insights into the effectiveness of basic convolutional networks for visual hate speech detection. Its moderate performance highlighted

the need for more complex models to capture subtle visual cues.

4.3 Baseline Multimodal

The baseline multimodal classifier combines the CNN-BiLSTM for text and ResNet50 for images. The fusion of these two modalities was achieved using a weighted product fusion technique, which combines the strengths of both models. This approach was more effective than unimodal models, demonstrating the potential benefits of integrating textual and visual information for hate speech detection. The multimodal model was further enhanced with dense layers and dropout regularization to improve its discriminative power and prevent overfitting.

5 Results

A subset of the MULTILATE dataset consisting of 1004 pieces of text as a basis for creating and evaluating a baseline classification model before the full release of the MULTILATE corpus. The first subset included 853 samples for training and validation, whereas another subset of 151 samples was reserved for testing and the baseline results are shown in the tables 2, 3 & 4. Table 2 highlights the performance of text-based models (CNN-BiLSTM, BERT, and RoBERTa) for binary and multiclass classifications, reporting metrics such as Precision (P), Recall (R), F1 Score (F1), and Accuracy (Acc). The results illustrate that while BERT and RoBERTa achieve higher accuracy in binary classification, CNN-BiLSTM performs better for multiclass classification due to its ability to capture sequential dependencies in text. Table 3 presents the performance of image-based models (VGG16, ResNet50, and CNN), showing that ResNet50 provides marginally better results for both binary and multiclass tasks, emphasizing its strength in visual feature extraction. Table 4 compares multimodal models combining CNN-BiLSTM for text and ResNet50 or VGG16 for images. The results demonstrate that the multimodal approach outperforms unimodal models by leveraging complementary information from text and images, with CNN-BiLSTM+ResNet50 achieving the highest accuracy. These tables collectively highlight the effectiveness and limitations of different models, providing a benchmark for future studies on multimodal hate speech detection.

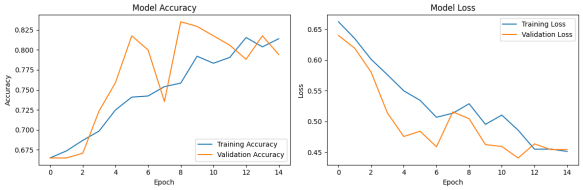


Figure 4: Training Accuracy and Loss on Binaryclass Text Classification.

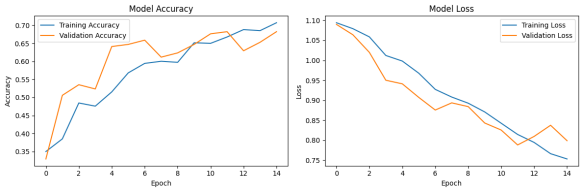


Figure 5: Training Accuracy and Loss on Multiclass Text Classification.

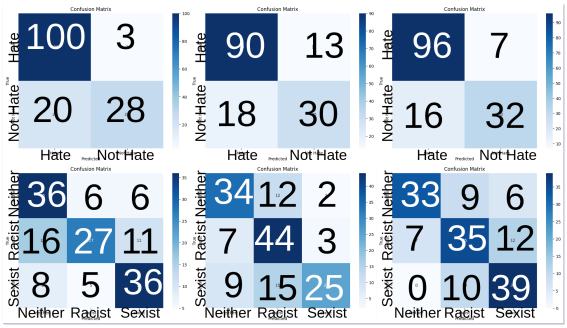


Figure 6: Confusion Matrix on Binary Text Classification in the first row, and Multiclass Text Classification in the second row, CNN-BiLSTM (left), BERT (middle), and RoBERTa (right).

Class	Model	P	R	F1	Acc
Binary	CNN-BiLSTM	0.77	0.79	0.79	0.79
Binary	BERT	0.90	0.58	0.70	**0.84**
Binary	RoBERTa	0.82	0.66	0.73	**0.84**
Multiclass	CNN-BiLSTM	0.69	0.68	0.69	**0.69**
Multiclass	BERT	0.70	0.68	0.67	0.68
Multiclass	RoBERTa	0.71	0.69	0.70	**0.69**

Table 2: Baseline models' performance for text modality: Precision (P), Recall (R), F1 Scores (F1), and Accuracy (Acc) for Binary and Multiclass Classification.

Class	Model	P	R	F1	Acc
Binary	VGG16	0.61	0.60	0.61	0.60
Binary	ResNet50	0.60	0.60	0.61	0.60
Binary	CNN	0.64	0.63	0.63	**0.64**
Multiclass	VGG16	0.40	0.40	0.41	0.40
Multiclass	ResNet50	0.41	0.40	0.41	**0.41**
Multiclass	CNN	0.39	0.39	0.40	0.39

Table 3: Baseline models' performance for image modality: Precision (P), Recall (R), F1 Scores (F1), and Accuracy (Acc) for Binary and Multiclass Classification.

Class	Model	P	R	F1	Acc
Binary	CNN-BiLSTM+VGG16	0.67	0.67	0.68	0.68
Binary	CNN-BiLSTM+ResNet50	0.69	0.69	0.68	**0.70**
Multiclass	CNN-BiLSTM+VGG16	0.52	0.51	0.52	0.54
Multiclass	CNN-BiLSTM+ResNet50	0.53	0.54	0.52	**0.55**

Table 4: Baseline models' performance for multimodal (text + image) modality: Precision (P), Recall (R), F1 Scores (F1), and Accuracy (Acc) for Binary and Multiclass Classification.

5.1 Error Analysis

The initial analysis of the MULTILATE dataset shows promising results for hate speech identification. BERT and RoBERTa achieved 0.84 accuracy in binary classification, while the CNN-BiLSTM model performed well in multimodal classification, with accuracy and loss metrics detailed in Figures 4 and 5. The confusion matrices in Figure 6 illustrate the model's ability to differentiate between sexist, racist, and neutral information, as well as between hate and non-hate categories. The ResNet50 image classifier demonstrated lower performance, indicating the need for more specialized architectures. Despite this, integrating visual and textual data has shown advantages over unimodal approaches. Further optimization and evaluation with larger MULTILATE datasets are planned for future studies. These preliminary results confirm the feasibility of hate speech detection using this multimodal dataset.

6 Conclusion

This study introduces MULTILATE, a comprehensive dataset designed to advance multimodal hate speech analysis by incorporating both synthetic visual content and diverse textual statements. The inclusion of adversarial examples and detailed interpretability annotations provides an essential resource for developing robust and explainable models. The dataset is constructed using rigorous data collection methods and benchmarked to assess classification performance across multiple modalities. Given the substantial computational requirements for AI-generated image content, the MULTILATE dataset will be released in batches. This phased-release approach is intended to manage the high resource demands effectively while ensuring accessibility and utility for future research. These efforts aim to enable significant advancements in understanding and mitigating online hate speech through multimodal analysis.

7 Limitations

Stable Diffusion demonstrates strong performance in image synthesis but has limitations when processing longer texts (over 65 words) and complex linguistic structures. It often ignores tokens beyond a certain length due to computational constraints, which can limit its effectiveness. Segmenting longer texts into smaller parts could help mitigate this issue. The 3W QA model, while helpful for understanding the "who", "what", and "why" of hate speech, may struggle with vague or ambiguous language, resulting in incomplete or inaccurate outputs. The model's reliability depends heavily on the clarity and quality of the input text. Adversarial examples, crafted to evaluate model robustness, do not always transfer effectively across different models or real-world scenarios, which limits their practical use. To improve model resilience, it is crucial to generate diverse adversarial examples for both text and image modalities, though challenges with transferability remain. Moreover, the Multilate pipeline requires a substantial amount of computational power and resources, particularly for AI image creation using models like Stable Diffusion. Due to these high computational demands, the dataset will be released in batches to manage the resource-intensive nature of generating AI-based images.

Acknowledgment

We appreciate the Department of Computer Science and Engineering (CSE) at the National Institute of Technology Silchar for allowing us to pursue our research and experimentation in CNLP and AI laboratories. We are grateful for the supportive research atmosphere that enhances our academic pursuits.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2021. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Science and Engineering*, 6(4):359–379.

Vicent Ahuir, José Ángel González, and Lluís-Felip Hurtado. 2022. Enhancing sexism identification and categorization in low-data situations.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, pages 1–28.

Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2021. "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Yingpeng Deng and Lina J Karam. 2020. Universal adversarial attack via enhanced projected gradient descent. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1241–1245. IEEE.

Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven CH Hoi, and Xiaogang Wang. 2018. Question-guided hybrid convolution for visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 469–485.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications. *CoRR*, abs/1910.03814.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Yucheng Huang, Rui Song, Fausto Giunchiglia, and Hao Xu. 2022. A multitask learning framework for abuse detection and emotion classification. *Algorithms*, 15(4):116.

Gyandeep Kalita, Eisha Halder, Chetna Taparia, Advaitha Vetagiri, and Partha Pakray. 2023. Examining hate speech detection across multiple indo-aryan languages in tasks 1 & 4. In *FIRE (Working Notes)*, pages 474–485.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Genevive Macrayo, Wilfredo Casiño, Jerecho Dalangin, Jervin Gabriel Gahoy, Aaron Christian Reyes, Christian Vitto, Mideth Abisado, Shekinah Lor Huyoa, and Gabriel Avelino Sampedro. 2023. Please be nice: A deep learning based approach to content moderation of internet memes. In *2023 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–5. IEEE.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Anku Rani, S.M Towhidul Islam Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10421–10440, Toronto, Canada. Association for Computational Linguistics.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Ramsha Saeed, Hammad Afzal, Sadaf Abdul Rauf, and Naima Iltaf. 2023. Detection of offensive language and its severity for low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 573–584.

Sima Sharifirad and Alon Jacovi. 2019. Learning and understanding different categories of sexism using convolutional neural network's filters. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 21–23, Florence, Italy. Association for Computational Linguistics.

Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Advaitha Vetagiri, Prottay Adhikary, Partha Pakray, and Amitava Das. 2023a. CNLP-NITS at SemEval-2023 task 10: Online sexism prediction, PREDHATE! In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 815–822, Toronto, Canada. Association for Computational Linguistics.

Advaitha Vetagiri, Prottay Kumar Adhikary, Partha Pakray, and Amitava Das. 2023b. Leveraging gpt-2 for automated classification of online sexist content. *Working Notes of CLEF*.

Advaitha Vetagiri, Prateek Mogha, and Partha Pakray. 2024. Cracking down on digital misogyny with multilate a multimodal hate detection system. *Working Notes of CLEF*.

Advaitha Vetagiri, Partha Pakray, and Amitava Das. A deep dive into automated sexism detection using fine-tuned deep learning and large language models. *Available at SSRN 4791798*.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Sebastian Wachs, Manuel Gámez-Guadix, and Michelle F Wright. 2022. Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, Behavior, and Social Networking*, 25(7):416–423.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. Ai for social science and social science of ai: A survey. *Information Processing & Management*, 61(3):103665.