A Large Scale Synthetic Dataset for MULTImodaL hATE "*MULTILATE*" with Text and Images and Adversarial Samples

Anonymous ACL submission

Abstract

Nowadays, one of the main problems our society struggles with is fighting online hate. In other words, as social media explodes with multimodal hate speech content, we require scalable multimodal hate speech detection systems. Thus, we present MULTILATE, a MULTImodaL hATE 2.6 million sample dataset for cross-modal hate speech classification and additional explanation through 3W Question Answering. Key features of the dataset include (1) textual utterances, (2) synthesized pictures produced by Stable Diffusion, (3) pixel-level temperature maps meant for explaining a picture-text interface, (4) question-answer triples addressing *"who", "what", and "why"* components of the statement,(5) Adversarial examples of both text and images. MULTILATE is aimed at creating and assessing interpretable multimodal hate speech classifiers.

1 Introduction

A prevalent sociological problem currently is online hate speech that 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, Facebook 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, thereby normalising prejudice (Wachs et al., 2022). Nowadays, we can find many forms of multimedia communication on social networks, such as mixing text, photos, and video. For example, most prior work on hate speech detection has concentrated only on text despite the abundance of multimodal hate content on the internet (Kumar et al., 2018).

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

Researchers have recently emphasized the creation of multi-modal hate speech detection systems that can perform at large scales, especially on platforms such as Facebook, Twitter, and Youtube (Gomez et al., 2020). Nevertheless, development of this domain has experienced limited due to the absence of large-scale cross-modal hate speech datasets. To circumvent this weakness, we propose to present a novel multimodal dataset named *"MULTILATE"* to facilitate mass-scale assessment of multimodal hate speech classification. The project MULTILATE consists of 2.6 million instances, including text-based statements and machine-produced visually appealing pictures created utilizing Stable Diffusion (SD) (Rombach et al., 2021). Further, every instance includes Pixel-Level HeatMaps for visual interpretability and Question-Answer (QA) pairs, which address "*who*", "*what*", and "*why*". 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.

A new, unique source of research data about interpretable multimodal hate speech classifiers is provided by the MULTILATE. It enables a mixed-mode model that integrates text, images, visual descriptions, and QA into systems where they can give understandable reasons for their forecasts. The dataset and benchmark for our study will spark innovations in the field of growing significance at the interface of computer vision, natural language processing, and ethics.

- **First large-scale multimodal hate speech dataset with 2.6 million examples**: However, previous hate speech datasets were small in size and modality-limited, thus stifling progress in multimodal detection. This work is based on an unprecedentedly huge dataset of 2.6 million samples with supporting documentation as in heat maps for the Images and

QA pairs for the Text.

- **Contains textual statements and synthetically generated hateful images**: We generated images paired with various hate-speech statements using Stable Diffusion models. Thus, it is possible to study the interaction between linguistic and visual modes of hate speech involving multi-modal perception.

- **Incorporates Question-Answering for explainability**: For instance, we extend the case with relevant information relating to who, what, and why. The model should also be interpretive and able to explain its predictions. This is what makes QA a model of rationality.

- **Includes adversarial examples for robustness**: We also give adversarial text and images to make the system resilient towards real examples. This allows for testing model limits and increases generality because it uncovers "*blind spots*" for more robust detection of multimodal hate speech.

To conclude, these contributions position our dataset as the one that could be used as a basis to progress on interpretable and reliable multimodal hate speech detection with its unparalleled scope, multimodality, justification, and adversarial. These data will significantly help address one of the most paramount social problems.

2 Realted Work

Hate speech refers to discrimination due to race, ethnic background, religion, gender, and sexual orientation. It has severe consequences, which include prejudice and violence in society. The classification of hate speech has mainly involved machine learning models such as 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 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) Generation of sets like TOXIGEN 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 deep learning 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 huge 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 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[2]. 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

[2] https://huggingface.co/spaces/Towhidul/5WQA

QA pairs are employed to check some elements of specific evidentiary documents for precise identification of falsity in claims.

In conclusion, the literature survey provides a nuanced understanding of the multifaceted challenges and advancements in hate speech and sexism classification, emphasizing the role of machine learning models in addressing these issues. From language-specific approaches to creating specialised datasets and exploring novel frameworks, the research showcased in the survey contributes to a growing body of knowledge aimed at mitigating the harmful effects of hate speech in various contexts.

3 Data

In this section, we have discussed the creation of a dataset called MULTILATE. As the name suggests, this dataset has been specifically designed to identify instances of hate speech, particularly sexism and racism, in online content. MULTILATE is a unique dataset we created that contains a total of 2.6 million examples extracted from 11 different datasets on sexism and racism. The labels used in this dataset are *"Hate"* and *"Not Hate"* for binary classification and *"Sexist"*, *"Racist"*, and *"Neither"* for Multiclass classification. This dataset can serve as a valuable resource for researchers and developers working on automated techniques for identifying and addressing instances of hate speech online. In the following sections, we will provide more details on creating and curating this dataset.

3.1 Data Sourcing

Data availability is one major thing that any model can benefit from. The classification job includes labelled data, which trains the model to obtain dependable accuracy. How correctly the characteristics are identified or retrieved directly affects how well the machine learning algorithms function. After the normalising text, detection tasks are carried out using classification algorithms. The efficacy of a model on a mixture of numerous datasets is always better than training on a specific dataset (Chiril et al., 2022). While creating the MULTILATE dataset, we extensively searched for published, public, and privately available datasets that contained instances of hate speech, as shown in Figure 1, which represents the flow of the process for creating the dataset. Tables 1, 2, and 3 provide valuable insights into the composition of the datasets from which text was extracted.

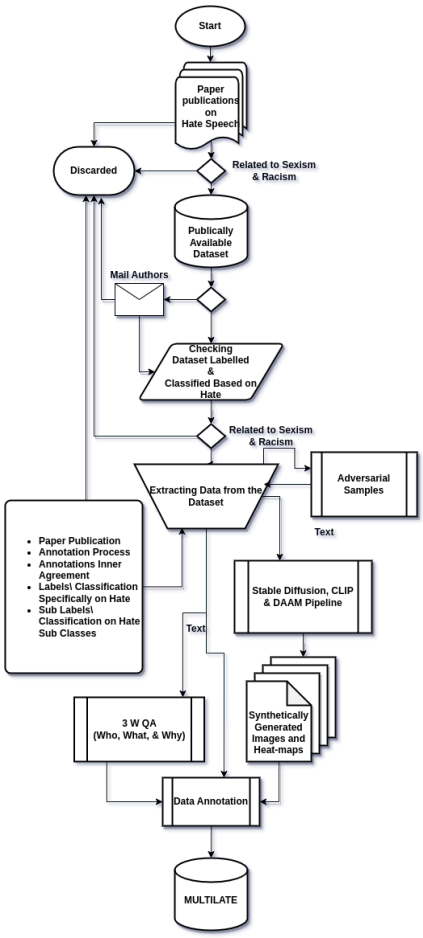

Figure 1: Flow diagram and pipeline of the MULTILATE creation.

Additionally, we contacted the authors of privately available datasets to request access to their data. In total, we were able to collect 69 datasets that contained examples of hate speech. To ensure that our dataset covered both sexism and racism, we only included datasets labelled and classified based on gender, race, ethnicity, sexist-racist slurs, stereotypes, and related features. Datasets that did not meet these criteria were excluded from our analysis. Through this rigorous approach, we aimed to create a comprehensive and representative dataset that could serve as a valuable resource for researchers.

3.2 Data Creation, Annotation, & Validation

Creating appropriate images constitutes a crucial part of the data pipeline for illustrating textured statements. Stable Diffusion, a current-generation text-conditioned image synthesis model, achieves this. Using multiple candidate images generated using Stable Diffusion for every textual claim gives different visual interpretations of the claims' text.

Table 1: Datasets for Sexist Classification

Dataset	Sexist	Not Sexist
CMSD (Samory et al., 2021)	1809	11822
EDOS (Kirk et al., 2023)	15330	44670

Table 2: Dataset for Racist Classification

Dataset	Racist	Not Racist
WSF (de Gibert et al., 2018)	1196	9507

Table 3: Datasets for both Sexist and Racist Classification with Adversarial Samples

Datasets	Sexist	Racist	Neither	Extracted
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**

The ranking of the candidates is done by using Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) that scores the images in terms of their suitability for the textual claim. Furthermore, for fine-grained visual explainability, Diffusion Attention Attribution Maps (DAAM) (Tang et al., 2022) are used to produce pixel-level heatmaps pointing out the parts of the made picture related to the words mentioned in the text. Combining these steps will supplement textual claims with appropriate Stable Diffusion images and fine-grained heatmaps that link textual concepts with visual parts.

A question-generating module is used to provide automatic textual "*who*", "*what*", and "*why*" QA pairs as complements to visual explainability for each claim. To begin with, semantic role labelling identifies pertaining textual spans that cover the major topics of the query. ProphetNet (Qi et al., 2020a) is a transformer-based QA model that generates natural language questions that can be answered from the claim text using these extractions. These questions will be answered automatically by employing another QA model, providing fine-grain textual explanations of the main actors, circumstances, and motivations to support each case.

The dataset introduces adversarial samples that exploit weaknesses to enhance model robustness. Through the backpropagation of tiny perturbations for images causing misclassification. Specific paraphrasing of the content and edits using character level for texts occur to ensure that semantics and fluency are maintained. This is achieved when they are combined because they are purposely designed to simulate real-world noise and distortions. Resilience to errors and ambiguities is built into such models through training on such adversarial cases.

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 by the MULTILATE dataset. The validation step includes the study of how well this model works on that basis, measuring its accuracy and generalizing characteristics. The results of the validation set used here are carefully evaluated and will be described in detail below under Results. This specialized subsample allows for a targeted assessment of the dataset's performance in detecting cases of hate speech, particularly with regard to sexism and racism online. By including both the textual and image levels as part of this validation process, our 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 Image - Stable Diffusion

While textual diversity in hateful statements is shared, the associated visual aspects also exhibit variability. To capture this, we utilize Stable Diffusion 2.1 (Rombach et al., 2021) to generate hateful images paired with textual statements. Stable

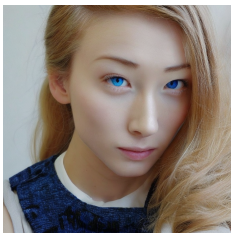

Figure 2: An image created for an example text "*how can you be chinese with blond hair and blue eyes -* ***Hate***", using Stable Diffusion.

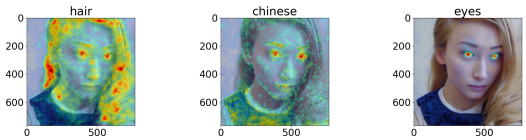

Figure 3: Heat maps generated for the Figure 2.

Figure 4: 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***".

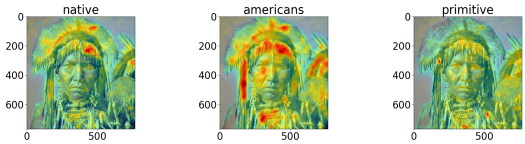

Figure 5: Heat maps generated for the Figure 4.

Diffusion's AI-based text-to-image generation capabilities allow the synthesis of diverse visual interpretations of hate speech.

Stable Diffusion is an open-source text-to-image model that can generate high-quality images conditioned on textual prompts. The latent diffusion process induces randomness, producing different results across generations. We generate three images per text statement and rank them to select the best pairing, as detailed next. Figures 2 and 4 are examples of images SD created for the respective text.

4.1 Re-ranking of Generated Images

To assess the generated images quantitatively, we use Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) to score each image based on the textual prompt. 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.

4.2 Pixel-level Image Heatmap

We utilize Diffusion Attention Attribution Maps (DAAM) (Tang et al., 2022) to generate pixel-level attribution maps highlighting which image regions correspond to the words in the associated hate speech prompt. This provides visual explainability into the generated multimodal pairing as shown in figure 6. The heatmaps are obtained by aggregating and upsampling cross-attention activations in Stable Diffusion's latent diffusion denoising model.

5 3W QA

The MULTILATE dataset provides a fine-grain explanation of the model outcome by supporting text justification. We develop a QA pipeline that asks multiple questions about each hate speech sentence and gets responses directly from the input data such as (Rani et al., 2023). We develop *"Who", "What", and "Why"* QA pairs pertaining to core semantics and protagonists of the statement.

5.1 3W Semantic Role Labelling (SRL)

QA generation is done in multiple stages and takes advantage of the latest neural semantic parsing and generative language model developments. Our first approach is to train an off-the-shelf neural SRL system that identifies spans from the input text to match predefined frame elements. Our targeted taxonomy uses ontology mapping to transform these semantic frames into Who, What, and Why roles.

5.2 Automatic 3W QA Pair Generation

These SRL extracted spans are fused with the input text and fed to a generative QA model called ProphetNet (Qi et al., 2020b). The unique n-stream self-attention mechanism helps deliberate planning ahead of predicting future tokens in ProphetNet, an encoder-decoder architecture. Language models are pre-trained on big corpora. ProphetNet generates well-formed and coherent questions that directly inquire about either who, what, or why (element) in terms of aspect extractions condition in the input text.

5.3 QA Pair Answering

We use the fine-tuned QA version of the T5 (Raffel et al., 2020) text-to-text transformer model for populating relevant solutions to these queries. T5 uses input statements and the Prophetnet's generated questions to answer by choosing the relevant extracted text with the solution. Quantified evaluation of multiple answers revealed that model T5 was the most accurate when extracting answers.

Lastly, we verify QA responses with evidence documents to ascertain whether input statements are logical. T5 models are then used to produce the final answer by combining the questions with the extracted evidence snippets into one input string. These responses and the original answers give high-resolution clues about which semantics of the statement are and which are not, backed by external data sources.

6 Adversarial Samples

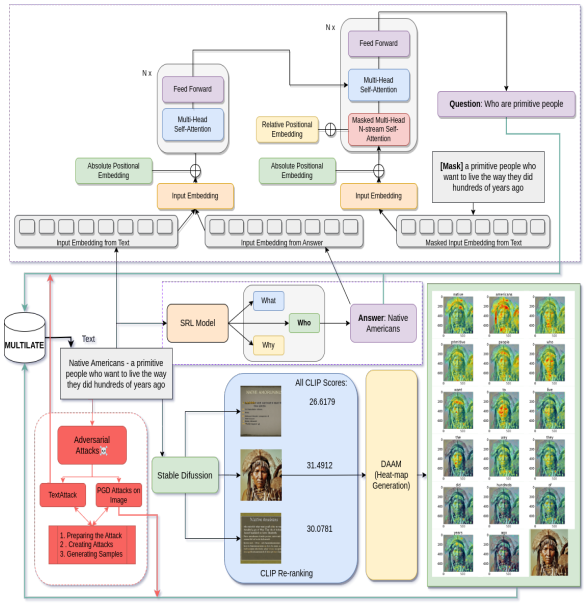

Figure 6: Overview of MULTILATE Framework - Integration of Stable Diffusion, SRL and T5 Models, and Adversarial Attack Setup for generating synthetic multimodal Hate Speech data.

We augment MULTILATE with adversarial examples for both text and images to improve model robustness and generalisation.

6.1 Adversarial Text

We generate adversarial text for 20% of MULTILATE using a TextAttack (Morris et al., 2020) model. The attack model inserts minimal perturbations into the original text that cause classifier errors but do not significantly alter human perception. For example, character manipulations like swaps, insertions, and deletions fool the model but look innocuous. The perturbed text retains semantics and fluency but fools hate speech classifiers.

6.2 Adversarial Images

We generate adversarial images for 20% of MULTILATE using projected gradient descent (PGD) (Deng and Karam, 2020) targeted attack on CLIP. PGD iteratively adjusts the image to maximize prediction error under an imperceptibility constraint. This finds small noise patterns that alter CLIP's prediction when added to the image. The noise is imperceptible but fools CLIP's multimodal hate speech judgments.

Training on these adversarial examples improves model resilience to semantic and visual perturbations. The adversarial augmentations in MULTILATE expose blindspots in current models, providing a challenging benchmark for developing robust multimodal hate speech detection.

7 Baseline Classification Models

7.1 CNN-BiLSTM

This research utilizes the CNN-BiLSTM (Fazil et al., 2023) model, combining convolutional and recurrent layers to improve feature extraction and sequencing. The model is implemented using the Keras library, and it accepts an ordered series of words converted into dense vector representations created from the GLOVE vector learnt on the MULTILATE dataset. After that, embedded words go through several convolutional layers that capture local features. The bidirectional LSTM layer is then used to obtain complete sequencing information. After that, the data from the second layer of LSTM is forwarded to the dense layer for binary classification.

During the experiment setup, several amendments were included to improve the model's efficiency. Secondly, a dropout mechanism was applied with a dropout rate of 0.2 to reduce the overfitting. Finally, the batch size of 128 was chosen for optimal processing speed. The 5-fold cross-section validation method was used to avoid the risk of overfitting and get unbiased estimations about the model's generalization accuracy. Each time, the dataset is split into five parts, with the model being trained on four of these parts and validated against the final part. This is done five times, with each

fold serving as the validation set precisely once. Finally, the average of all folds' final performance metrics is computed.

7.2 ResNet50

Using the MULTILATE dataset, the ResNet50 (Macrayo et al., 2023) architecture served as the baseline for image classification. The image inputs were processed using a pre-trained ResNet50 network and initialized with ImageNet weights in this Keras implementation. This enabled the model, to offer vital information necessary, especially when extracting image features. After that, the baseline ResNet50 was trained to fit the MULTILATE dataset's specific characteristics.

The ResNet50 model was upgraded by adding two more dense layers to improve its discriminant abilities. In this case, the first dense layer was fed 128 samples, ensuring the model could learn complicated patterns. This layer included a dropout rate of 0.2, whereby some neurons were randomly deactivated during training to avoid overfitting. This was followed by another thicker layer of density having a size of 64 and a slightly reduced dropout ratio of 0.1. The objective of this configuration was to provide a balance between model complexity and regularization that would lead to better generalization to fresh images in the MULTILATE dataset. A compelling image classifier could be achieved through extensive architecture whereby ResNet50 was used as the foundation with additional dense layers and suitable regularization.

7.3 Baseline Multimodal

The baseline multimodal classifier uses the CNN-BiLSTM model combined with ResNet50 to process text and images. The CNN-BiLSTM works well on sequential data and gives context information along a sequence of words based on GloVe embeddings. At the same time, ResNet50 pre-trained on ImageNet provides substantial image feature extraction to a multimodal architecture. Thus, the output from these two modalities is combined through the weighted Product fusion technique, providing an optimal unified representation that optimally complements the advantages of each of these models. Classifiers with two dense layers at 128 and 64 batches and dropouts of 0.2 and 0.1 that prevent overfitting improve discriminating power. The main goal is to utilize this combined multimodal approach. In particular, it provides an initial insight into how this dataset can be used to improve future multimodal classification techniques.

8 Results

A subset of the MULTILATE dataset consisting of 1004 pieces of text as a basis for creating and evaluating a baseline classification model. 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 table 4. The CNN-biLSTM was used for the categorization function in text information concerning category descriptors. This model offers an open baseline using available data before the full release of the MULTILATE corpus. These preliminary results show that the classification task can be carried out on the currently available data, which serves as a basis for more precise evaluations once larger volumes of the data sets from the MULTILATE project become available. The source codes and the train-test splits will be available for the public to compare with other research that can improve the models.

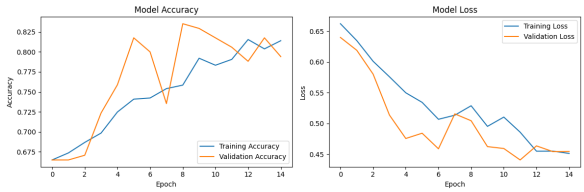

Figure 7: Training Accuracy and Loss on Binaryclass Text Classification.

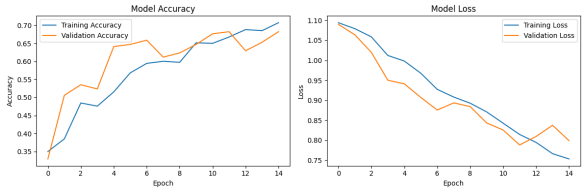

Figure 8: Training Accuracy and Loss on Multiclass Text Classification.

8.1 Results Analysis

On the subset of the MULTILATE dataset utilized in these pilot studies, the results show encouraging performance for hate speech identification with 0.84 accuracies on binary classification using BERT & RoBERTa and 0.69 accuracy for multiclass labelling; the CNN-BiLSTM text classification model obtained good metrics on multimodal classification, the model's accuracy and loss graphs are shown in the figures 7 and 8. The confusion matrices show how to distinguish between infor-

Table 4: Baseline models Precision (P), Recall (R), F1 Scores (F1) and Accuracy (Acc) on Binary Classification and Multiclass Classification.

Class	Modality	Model	P	R	F1	Acc
Binary	Text	CNN-BiLSTM	0.77	0.79	0.79	0.79
Binary	Text	BERT	0.90	0.58	0.70	**0.84**
Binary	Text	RoBERTa	0.82	0.66	0.73	**0.84**
Binary	Image	VGG16	0.61	0.60	0.61	0.60
Binary	Image	ResNet50	0.60	0.60	0.61	0.60
Binary	Image	CNN	0.64	0.63	0.63	**0.64**
Binary	Multimodal (Text+Image)	CNN-BiLSTM+VGG16	0.67	0.67	0.68	0.68
Binary	Multimodal (Text+Image)	CNN-BiLSTM+ResNet50	0.69	0.69	0.68	**0.70**
Multiclass	Text	CNN-BiLSTM	0.69	0.68	0.69	**0.69**
Multiclass	Text	BERT	0.70	0.68	0.67	0.68
Multiclass	Text	RoBERTa	0.71	0.69	0.70	**0.69**
Multiclass	Image	VGG16	0.40	0.40	0.41	0.40
Multiclass	Image	ResNet50	0.41	0.40	0.41	**0.41**
Multiclass	Image	CNN	0.39	0.39	0.40	0.39
Multiclass	Multimodal (Text + Image)	CNN-BiLSTM+VGG16	0.52	0.51	0.52	0.54
Multiclass	Multimodal (Text + Image)	CNN-BiLSTM+ResNet50	0.53	0.54	0.52	**0.55**

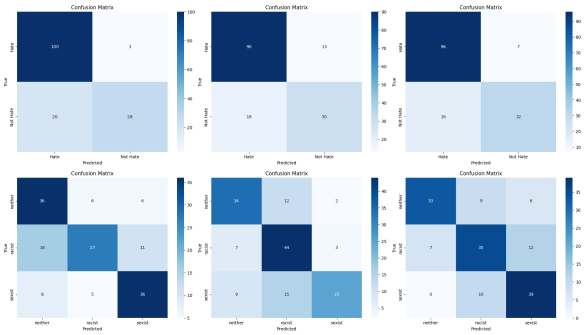

Figure 9: Confusion Matrix on Binary Text Classification in the first row, and Multiclass Text Classification in the second row, CNN-BiLSTM (left), BERT (middle), and RoBERTa (right).

mation that is sexist, racist, or neither, as well as how to discriminate between the hate and non-hate classifications effectively, as shown in figure 9.

Less skill is shown by the ResNet50 image classifier, suggesting that more customized architectures are required and that visual hate speech identification is a more difficult task but performs better than the unimodal variation, demonstrating the importance of integrating visual and textual information. These models will be further optimized and evaluated at scale using more extensive MULTILATE data in future studies. However, these initial results validate the feasibility of the hate speech detection task on this novel multimodal dataset.

9 Limitations

In the context of image synthesis, Stable Diffusion demonstrates impressive results but appears to have certain weaknesses in processing particular text inputs. Notably, the model faces challenges with extremely long texts – more than 65 words or consisting of a few sentences. Additionally, issues are often encountered while dealing with metaphorically altered text in terms of processing difficult linguistic formations. An interesting note to mention is that the model silently ignores tokens above 77, whereby tokens represent words or text equivalent groups. Yet, this behaviour can be regarded as a possible limitation of the maximum input length that the model can handle effectively in terms of computational and memory limitations. Such problems can be mitigated with the proposed approach of dividing input text into smaller segments that may allow a model to process longer and more complex descriptions.

10 Conclution

This study presents MULTILATE, an unprecedented large-scale dataset that uses produced visual and various textual assertions to further multimodal hate speech analysis. Adding adversarial examples and fine-grained explainability annotations provides unique capabilities for resilient and interpretable models. We outline the rigorous data-gathering methods that went into creating MULTILATE and benchmark categorization performance.

The most excellent resource currently accessible for increasing our understanding of multimodal hate speech is the 2.6 million sample MULTILATE corpus. It encourages crucial advances in computer vision, natural language processing, and ethics. The scale, diversity, interpretability, and integrity of MULTILATE enable this study to lay the groundwork for significant future endeavours. We believe that this dataset's baseline and ongoing advancements will help find answers to the problem of online animosity.

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.

Mohd Fazil, Shakir Khan, Bader M Albahlal, Reemiah Muneer Alotaibi, Tamanna Siddiqui, and Mohd Asif Shah. 2023. Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction. *IEEE Access*, 11:16801–16811.

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.

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.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. 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.

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.

2020a. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020b. 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.

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*.

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.