Team Oryu@CHiPSAL 2026:Integrating Text and Vision Transformers for Multimodal Hate Speech Detection in Memes

Noore Tamanna Orny, Joyeta Barua Moni, Md. Abtahee Kabir, Hasan Murad
Chittagong University of Engineering and Technology
Pahartali, Raozan, Chattogram, Bangladesh
{u2104092, u2104130, u2104089}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

With the proliferation of multimodal content on various social media platforms, automated hate speech detection has emerged as a challenge, especially in meme-based communication, where meaning arises from interactions between text and images. In these situations, unimodal techniques are inadequate in capturing semantics. In order to address such issues, a late-fusion-based multimodal hate speech detection framework has been proposed and implemented for the CHiPSAL shared task. In the proposed framework, multimodal content is processed by utilizing XLM-RoBERTa for multilingual text representation and a Vision Transformer (ViT) for visual representation. Both modal representations are fused using a fully connected classification head and are used for binary hate speech detection. The findings suggest that multimodal content effectively captures features from individual modalities and helps improve hate speech detection accuracy by obtaining a Macro F1-score of 0.66 and ranking 5th on the leaderboard. Also, transformer-based multimodal fusion performs effectively and acts as a reliable baseline for hate speech detection in low-resource multilingual meme-based communication scenarios.

Keywords: Multimodal Hate Speech Detection, Meme Analysis, Vision Transformer (ViT), XLM-RoBERTa, Multimodal Learning, Hate Meme Classification, Social Media Moderation, Macro F1 Optimization, Transformer-based Models, Multilingual NLP

1. Introduction

With the proliferation of multimodal content on various social media platforms, automated hate speech detection has emerged as a challenge, especially in meme-based communication, where meaning arises from interactions between text and images. In such scenarios, unimodal-based approaches have been found inadequate in capturing semantics from multiple modalities. In order to address such issues, a late-fusion-based multimodal hate speech detection framework was proposed and implemented for the CHiPSAL shared task (Thapa et al., 2026; Sarveswaran et al., 2026). In the proposed framework, multimodal content is processed by utilizing XLM-RoBERTa for multilingual text representation and a ViT for visual representation. Both modal representations are fused using a fully connected classification head and are used for binary hate speech detection.The major contributions of this work can be summarized as follows:

1. This work presents a transformer-based multimodal hate speech detection system specifically designed to handle the complexities of low-resource multilingual memes.

2. This work validates the effectiveness of the XLM-RoBERTa and Vision Transformer architectures in the context of the hate speech detection task, specifically through the application of the late-fusion strategy.

3. This work presents the application of class-weighted training, which is effective in reducing the issue of class imbalance for this task. This work presents competitive performance securing the 5th position on the leaderboard.

While the proposed approach is an improvement over existing transformer-based architectures, the major contribution is in the systematic analysis of the effectiveness of late fusion approaches, class imbalance handling, and the integration of multimodal features in a low-resource code-mixed meme environment. This study offers significant empirical evidence of the effectiveness of the proposed approaches and hence serves as a robust baseline for future studies.

2. Related Work

Due to the rapid growth of user-generated content on social media, there is a considerable increase in harmful and hateful content, which led to extensive research on automated hate speech detection systems. Initial research on this topic was mostly focused on text-based classification problems using traditional machine learning techniques, including Support Vector Machines and logistic regression models with handcrafted features. Although these models demonstrated considerable success, they failed to capture contextual nuances, sarcasm, and implicit hate speech, which are often encountered in real-world datasets (Maharjan et al., 2025).

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 284–291

Early studies have also dealt with the challenges that are encountered during the detection of hate speeches on social media platforms, particularly when using automated detection techniques. The research in this area has dealt with the detection of hate speeches on social media platforms with the help of natural language processing techniques. The various challenges that are encountered during the detection of hate speeches on social media platforms with the help of NLP-based detection mechanisms include contextual ambiguity, sarcasm, domain variability, and the difficulty of identifying implicit hateful expressions that are common in online discourse. These complex linguistic features are difficult to detect with the help of NLP-based detection mechanisms. Therefore, advanced detection mechanisms, such as deep learning, for hate speeches on social media platforms are required (Parihar et al., 2021).

Recently, with the introduction of deep learning techniques, transformer-based language models including BERT and its multilingual variants have demonstrated considerable success in improving the performance of hate speech detection systems. Multilingual variants, including XLM-RoBERTa, have demonstrated considerable potential for handling low-resource languages by applying cross-lingual transfer learning techniques. These models are able to learn contextualized word representations, which are more accurate than traditional word embeddings for semantic meaning representation (Das et al., 2024). However, these text-based systems are insufficient for meme-based analysis, as hate speech is often represented by interacting with images. In order to address this drawback, there has been a growing tendency towards multimodal-based hate speech detection. This method uses a combination of text and image features to better comprehend hate speech, as the latter often makes use of images to convey its intended meaning. ViT and CNN-based models have emerged as popular choices for image encoding in multimodal-based hate speech detection methods (Thapa et al., 2025a). These models facilitate the effective extraction of semantic features from images.

The process of multimodal learning relies heavily on the fusion mechanism employed. Early fusion, late fusion, and cross-modal attention-based methods have been explored in previous studies. It has been demonstrated that late fusion-based methods, which involve the independent encoding of each modality followed by their combination, can offer a robust yet computationally efficient solution for multimodal learning tasks (Qi and Wei, 2025). Recent studies have introduced the concept of gated fusion and attention-based methods to weigh the relative importance of each modality, thereby improving the robustness of the model in the presence of noisy or misleading modalities(Piot and Parapar, 2025)

Upon these developments, the present study employs a late fusion multimodal transformer model with integrated XLM-RoBERTa for text understanding and ViT for feature extraction from images. The motivation is to capitalize on the strengths offered by pre-trained multimodal representations while ensuring stability and efficiency for meme classification in resource-scarce environments.

3. Data Description

3.1. Dataset Overview

The experiments of this work were performed on a multimodal Nepali meme dataset that was created to perform hate speech detection (Thapa et al., 2025b,c). Each instance of the meme in the dataset contains both textual and visual information, representing the multimodal nature of internet memes. The problem is defined as a binary classification problem where the system needs to predict whether the meme is hateful or non-hateful.

Each instance of the meme contains information related to the image of the meme, textual information related to the meme, and a label related to the meme. The dataset is challenging since the hateful information is encoded based on the interaction of both textual and visual information of the meme. The annotation framework used in the dataset follows the multimodal hate speech annotation guidelines introduced in the CrisisHateMM dataset (Bhandari et al., 2023).

Label	Training	Evaluation	Test
Hate (1)	720	98	–
Non-hate (0)	348	35	–
Total	**1068**	**133**	**134**

Table 1: Label-wise Distribution of Training, Evaluation, and Test Data

3.2. Data Splits

The dataset is split into training, validation(evaluation) and test sets.The training split contains 1,068 samples, the validation(evaluation) split contains 133 samples, and the test split contains 134 samples. In the validation set, the hate class includes 98 samples and the non hate class includes 35 samples, which indicates a noticeable imbalance like the training set. The evaluation dataset was used only to perform model selection, monitoring of hyperparameters, and initial model performance evaluation, while the test data was used only to perform the final submission and leaderboard evaluation.

3.3. Label Distribution

From the dataset, we can see that there is a class imbalance issue, as the hate class dominates the dataset. In the training set, the hate class(label=1) has 720 instances, whereas the non-hate class(label=0), has only 348 instances. In the evaluation set, there are 98 hate class instances, whereas there are only 35 instances of the non-hate class.

3.4. Text Characteristics

The text data portion of the given data set includes small meme captions, as well as other textual data that is full of noisy, contextual, and informal language. The text data is full of slang, spelling mistakes, code-switching, emojis, and other contextual expressions, making it more complicated for the model to detect hate text automatically. The text data also includes many implicit expressions of hateful intent, such as sarcasm, humor, etc., rather than explicit hateful words or phrases. The text data portion is also too short, as it is part of the meme, making it more complicated for the model to classify the text data correctly.

3.5. Image Characteristics

The visual component of the dataset comprises meme images sourced from social media platforms, which have varying levels of diversity in their visual content and the nature of the image and content used in the memes. These images have varying resolutions and ratios, ranging from low-resolution compressed images to more visually appealing and professionally created images, although in the interest of model compatibility, all the images have been preprocessed to 224x224 pixels. These images have varying levels of text overlay, symbolization, and graphical content, which are visually representative of the memes and the context in which the memes are used. Additionally, the dataset has images that are natural and those that have been artificially created, including complex backgrounds and varying levels of color distribution in the images.

3.6. Preprocessing

For textual processing, meme text was tokenized using the XLM-RoBERTa tokenizer with a maximum sequence length of 128 tokens. For the visual modality, images were resized to 224×224 pixels, randomly horizontally flipped during training for data augmentation, normalized, and converted to tensor format. These preprocessing steps ensure compatibility with the pretrained transformer encoders while improving model generalization. As shown in Figure 1, the proposed framework consists of modality-specific encoders followed by a multimodal fusion module.

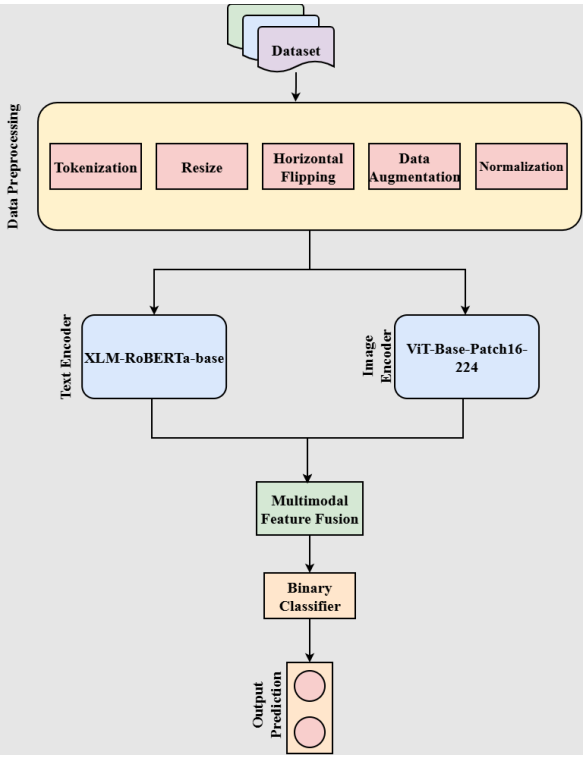

Figure 1: Proposed multimodal hate speech detection architecture.

4. Methodology

4.1. Task Definition

The problem this paper aims to solve is the multimodal hate speech detection in memes. In this case, the information available in the memes will be both textual and visual.

The problem can be formally defined as follows: given a meme with text information T and image information I, the aim of the model will be to learn the following function-

$$f(T, I) \rightarrow y$$

where y $\in$ (0 , 1) corresponds to hate/ non-hate.

4.2. Overall Architecture

The proposed system utilizes a late fusion multimodal transformer with four major components:

1. Multilingual Text Encoder: XLM-RoBERTa
2. Vision Transformer Image Encoder: ViT
3. Feature Fusion Module

4. Fully Connected Classification Head

This architecture utilizes modality-specific encoders and fuse the features using a concatenation method followed by a non-linear transformation.

4.3. Text Encoding

For textual semantics, we use the pre-trained model XLM-RoBERTa-base(Conneau et al., 2019), which is a transformer model pre-trained on cross-lingual large-scale data. This model has been chosen due to the multilingual content of the meme text in the dataset, which includes Nepali, English, and code-switched text.

The input sequence T, the text embedding of which we want to find, is first split into subword units using the tokenizer. Then the transformer model processes the sequence to find the text embedding

$$h_T = \mathrm{XLM{-}R}(T)$$

The text embedding has a dimensionality of 768 units.

4.4. Image Encoding

For visualization purposes, we will use Vision Transformer (ViT-Base-Patch16-224)(Wu et al., 2020) which has been pre-trained on ImageNet. Unlike convolutional neural networks, ViT uses a self-attention mechanism to consider global relationships in images.

Each image is resized to 224x224 pixels. This image is divided into 16×16 patches. The patches are then embedded and passed through the transformer encoder to obtain a vision representation:

$$h_I = \mathrm{ViT}(I)$$

The classification head is removed from the ViT. The 768-dimensional vector is used as the image embedding. ViT is used because:memes require global context understanding, hateful cues are spatially distributed, transformers are similar to text transformers

4.5. Multimodal Fusion

After getting the modality-specific features, the model applies late fusion using feature concatenation. The text and image embeddings are concatenated in the feature dimension which is a 1536-dimensional representation.

$$h_F = [h_T; h_I]$$

The late fusion approach is used because: pre-trained encoders can be utilized separately,it is robust with limited training data, it alleviates overfitting across modalities

4.6. Classification Head

The fused representation is then passed through a fully connected neural network to capture cross-modal interactions. The classification head is composed of the following layers:

Component	Configuration
Dropout	0.3
Linear Projection	$1536 \rightarrow 512$
Activation Function	ReLU
Dropout	0.3
Linear Layer	$512 \rightarrow 2$

Table 2: Classification Head Architecture

The mathematical representation of the classification head is given by:

$z = ReLU(W_1\ \mathsf{h}_F + \mathsf{b}_1\ ),\ \hat{y} = W_2 \mathsf{z} + \mathsf{b}_2$

The last layer generates logits for the binary hate classification problem.

This was done using the classification head as dimensionality reduction improves generalization, the ReLU(Nair and Hinton, 2010) activation function adds non-linearity to the cross-modal learning, and the dropout adds regularization to prevent overfitting as the dataset used was not very large.

4.7. Handling Class Imbalance

The dataset has a class imbalance problem, where one class has significantly more instances than the other. To tackle this issue, a class-weighted cross-entropy loss is used. Class weights are calculated using the following formula based on the balanced heuristic:

$$w_c = \frac{N}{K \cdot N_c} \tag{1}$$

where N_c is the number of instances in the class c.

The weighted loss is calculated as follows:

$$\mathcal{L} = -w_y \log p_y$$

As we address the class imbalance, it-

1. Improves minority class recall.
2. Stabilizes macro-F1.
3. Prevents majority bias.

4.8. Training Procedure

The proposed multimodal model was trained using an end-to-end approach to perform three epochs of training. In the case of each meme, the input text was tokenized, and the contextual text embedding was obtained using the pre-trained XLM-RoBERTa model, whereas the input image was passed through the ViT model to extract the visual

feature embedding. The obtained feature embedding from both the image and the text was concatenated to form a single feature vector, which was passed through a fully connected classification head to obtain the hate and non-hate class logits.

The model was optimized using the AdamW(Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5, whereas class-weighted cross-entropy loss was used to handle the class imbalance issue, and the model selection was carried out using the best model based on the validation set's Macro F1-score.

4.9. Implementation Details

The model was developed using PyTorch as a deep learning library, as well as Hugging Face's Transformers library for transformer models. Text features were represented using a pretrained XLM-RoBERTa-base model with a sequence length of 128 tokens. Visual features were represented using a pretrained Vision Transformer (ViT-Base-Patch16-224), which was accessed via the timm library, resizing images to 224x224 pixels and randomly flipping images horizontally. A fully connected fusion network was used to fuse the 768-dimensional text and image embeddings, with dropout set to 0.3 for a binary classification problem. AdamW was used as the optimizer with a learning rate of 2×10^{-5},as well as class-weighted cross-entropy loss to account for class imbalance. Three epochs were trained with a batch size of 4, and the model was selected based on Macro F1-Score on the evaluation set.The training configuration of the model has been shown in Table 3:

Parameter	Value
Framework	PyTorch
Text Model	XLM-RoBERTa-base
Image Model	ViT-Base-Patch16-224
Image Size	224×224
Max Text Length	128
Optimizer	AdamW
Learning Rate	2×10^{-5}
Dropout	0.3
Epochs	3
Batch Size	4
Hardware	NVIDIA GPU (Kaggle)

Table 3: Training Configuration

5. Results and Discussion

5.1. Evaluation Metric

The multimodal model was evaluated based on the official CHiPSAL shared task protocol. The main evaluation criteria used were Macro F1-score, which is mostly used for imbalanced classification problems since all classes are given equal weightage. In addition to Macro F1-score, Accuracy, Macro Precision, and Macro Recall were used to keep track of the model's performance.

Accuracy is defined as the proportion of samples that were predicted and actually belonged to the class predicted by the model.

Macro Precision is defined as the proportion of positive class predictions made by the model.

Macro Recall evaluates how well the model is able to recall positive samples across all classes. The Macro F1-score is the harmonic mean of Precision and Recall computed separately for each class and then averaged over all classes.

5.2. Quantitative Results

The performance of the proposed late fusion multimodal model on the test set is shown in Table 4.

Metric	Score
Macro F1	0.66
Accuracy	0.68
Precision (Macro)	0.64
Recall (Macro)	0.55

Table 4: Performance of the Proposed Model

To comprehend the contribution of each mode better, we also test the performance of the unimodal models which is shown in Table 5.

Model	Macro F1
Text-only (XLM-RoBERTa)	0.21
Image-only (ViT)	0.45
Proposed	0.52

Table 5: Comparison of Unimodal and Proposed Models using Validation Set

In this case, the text-only model uses XLM-RoBERTa, while the image-only model uses the Vision Transformer (ViT). The results show that the multimodal model has a better performance than the other models, illustrating the effectiveness of the fusion of textual and visual modalities for meme comprehension. All the approaches in Table 5 are conducted under consistent experimental settings, using the same dataset, trained for 3 epochs with a learning rate of 2×10^{-5}, and the same loss function. On the official Codabench leaderboard, our system has achieved: Macro F1: 0.66 , Leaderboard Rank: 5th place

5.3. Analysis of Model Performance

The performance of the multimodal hate speech detection system was enhanced gradually with the improvement of the model architecture and the training strategy. Initially, the multimodal baseline model achieved a Macro F1-score of 0.40, which implied that the model was not effective enough to handle the relationship between the hate speech and the visual content of the memes. The model was also challenged to make balanced decisions on the hate speech classification.

A gated fusion mechanism was incorporated to address the challenges associated with the initial multimodal model. With the enhanced model, the Macro F1-score was improved to 0.52. This was due to the incorporation of the gating mechanism that allowed the model to control the amount of information to be used from the textual and visual content of the hate speech memes. The model was improved to be effective in highlighting the importance of the modality of the hate speech content. The enhanced model was also challenged to be somewhat sensitive to noise and instability. Significantly high performance gain was achieved using the final late fusion transformer model, which attained a Macro F1-Score of 0.66. Several reasons contributed to the performance gain. First, the application of class-weighted cross-entropy loss effectively addressed the class imbalance problem, allowing the model to learn more balanced decision boundaries between the hate and non-hate classes. Second, the late fusion approach ensured greater stability in the model as the text encoder and image encoder learned strong modality-specific representations before fusion. Third, the effective handling of the data and the application of consistent preprocessing ensured the quality of the input data, which in turn ensured greater stability in the model.

The model attained its best performance within the first three epochs, which confirms the effectiveness of the pre-trained XLM-RoBERTa and Vision Transformer backends in the model. However, the slight variation in the performance metric across the epochs confirms that the model is still slightly sensitive to the presence of noisy data in the dataset and the class imbalance problem, which can be addressed in the future.

Model	Macro F1
Basic multimodal baseline	0.40
Gated fusion variant	0.52
Final late fusion transformer model	0.66

Table 6: Comparison of Model Variants

.

5.4. Error Analysis

Despite the strong performance in competitive scenarios, the model faces obstacles in a few challenging scenarios.

1. The model struggles with memes where hate is implied rather than stated. The reasons for this include a lack of deep cross-modal reasoning and insufficient world knowledge.

2. Sarcastic memes continue to be challenging due to the ambiguity of the text's sentiment and the image. The literal meaning of the image and the intended meaning of the meme can be very different.

3. The model uses the Vision Transformer for visual understanding, which inherently processes the textual patterns present in the images through patch embeddings. However, it does not explicitly include text extraction through dedicated mechanisms such as OCR. As a result, the model struggles with the hateful content present in the text of the images in the case of memes.

4. Regional memes can be challenging for the multilingual model, particularly when there is limited linguistic resource. As depicted in the confusion matrix in Figure 2, a substantial number of hate instances are correctly classified. Nevertheless, the system also experiences a considerable number of errors. To be precise, the system incorrectly classifies 26 instances of non-hate as hate, while it also incorrectly classifies 24 instances of hate as non-hate.

This also reveals that the system is highly sensitive to certain cues, either lexical or visual, which it perceives as indicative of hate. At the same time, it is also evident that certain hate instances are not as obvious for the model, which does not leverage multimodal reasoning.

Error Type	Count
False Positives	26
False Negatives	24

Table 7: Error Analysis on Validation Set

6. Conclusion

In this study, a multimodal architecture is proposed where text representation learning is accomplished using XLM-RoBERTa, while a Vision Transformer is used for image feature extraction. A late fusion mechanism is employed along with a fully connected classification network. To address class imbalance, class-weighted cross-entropy loss is

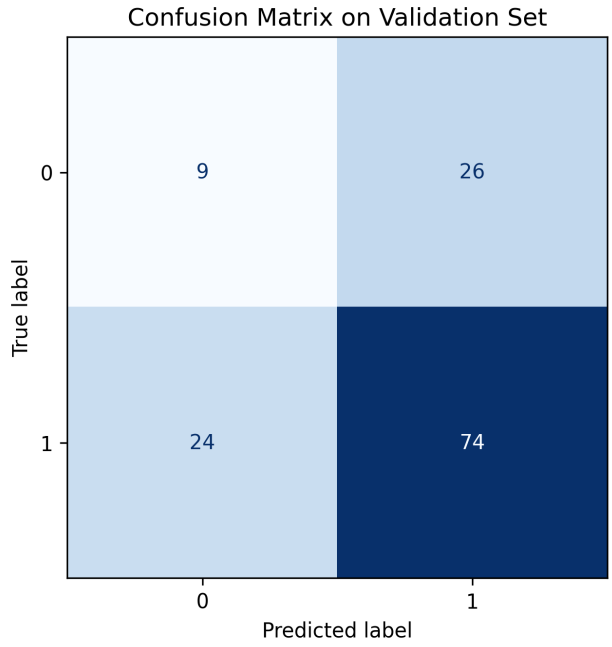

Figure 2: Confusion Matrix on Validation Set

incorporated, and image augmentation is applied to improve model performance. Based on the experimental results obtained using the CHiPSAL shared task dataset, the multimodal architecture effectively captures information from both modalities and maintains stable performance on the evaluation set. This demonstrates the effectiveness of transformer-based multimodal learning for meme hate speech detection while remaining computationally efficient through the late fusion design.

7. Limitations

Several avenues for further research exist for the proposed multimodal hate speech detection approach. First, the scope of the proposed work should be extended by incorporating more data for the training process, including more meme collections, for the detection of hate speech across different domains. Second, more sophisticated multimodal fusion methods, such as attention-based or region-token alignment, should be explored. Third, domain adaptive fine-tuning of the proposed pretrained transformers should be explored.

Furthermore, the proposed multimodal hate speech detection approach should be extended by incorporating more contextual information, such as user data, for the detection of implicit hate speech. Finally, more robust training methods should be explored for the proposed multimodal hate speech detection approach for the detection of hate speech across different domains as well as more efficient multimodal architectures.

Moreover, this work is centered on late fusion due to its stability and effectiveness in low-resource settings, alternative fusion strategies such as early fusion and cross-modal attention-based methods could further enhance interaction between modalities. Exploring these approaches may provide deeper multimodal alignment and is an important direction for future work.

8. Ethical Statement

In this research, a multimodal hate speech detection model is proposed based on the publicly available meme data from the shared task of the CHiPSAL dataset. Ethical considerations are also taken into account in the research. The data is publicly available and is curated for research purposes. There is no personal information available in the dataset. Therefore, the privacy of the data is not compromised in any way. To avoid the bias in the data available for the hate speech dataset, class-weighted cross-entropy loss is used in the model. The model is also pre-trained using the multilingual XLM RoBERTA model. However, the bias may still be present in the data because of the less diverse data available.The proposed system is strictly for research and decision support in content moderation and is not intended for fully autonomous use without human supervision because automated systems may also yield false positives or false negatives, which could impact the user. Data augmentation and balancing are also included in the research for fairness and generalization. However, detecting hate in multimodal content is difficult, especially for less represented cultures. The research follows the ethical principles of AI and the ethical guidelines set in the shared task. The model is used for classification of existing content only and does not produce content that is harmful.

9. Bibliographical References

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Susmita Das, Arpita Dutta, Kingshuk Roy, Abir Mondal, and Arnab Mukhopadhyay. 2024. A survey on automatic online hate speech detection in low-resource languages.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Sujal Maharjan, Astha Shrestha, Shuvam Thakur, and Rabin Thapa. 2025. Multimodal kathmandu@CASE 2025: Task-specific adaptation of multimodal transformers for hate, stance, and humor detection. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 107–114, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA. Omnipress.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Paloma Piot and Javier Parapar. 2025. Towards efficient and explainable hate speech detection via model distillation. In *European Conference on Information Retrieval*, pages 376–392. Springer.

Meng Qi and Chung-Lun Wei. 2025. Lamha: Efficient multimodal malicious meme classification via lora-tuned adaptation and attentive-mlp fusion. In *Proceedings of the 2025 7th International Conference on Control and Computer Vision*, ICCCV '25, page 62–68, New York, NY, USA. Association for Computing Machinery.

Kengatharaiyer Sarveswaran, Surendrabikram Thapa, Ashwini Vaidya, Tafseer Ahmed Khan, and Bal Krishna Bal. 2026. Findings of the second workshop on challenges in processing south asian languages (chipsal 2026). In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Siddhant Bikram Shah, Kritesh Rauniyar, Shuvam Shiwakoti, Surabhi Adhikari, Hariram Veeramani, Kristina T. Johnson, Ali Hurriyetoglu, Hristo Tanev, and Usman Naseem. 2025a. Multimodal hate, humor, and stance event detection in marginalized sociopolitical movements. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 20–31, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Kengatharaiyer Sarveswaran, Bal Krishna Bal, and Usman Naseem. 2026. Multimodal hate and sentiment understanding in low-resource text-embedded images for online safety and digital well-being. In *Proceedings of the Second Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*.

Surendrabikram Thapa, Hariram Veeramani, Liang Hu, Qi Zhang, Wei Wang, and Usman Naseem. 2025b. A multimodal prompt-based framework for analyzing code-mixed and low-resource memes. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1913–1923.

Surendrabikram Thapa, Hariram Veeramani, Imran Razzak, Roy Ka-Wei Lee, and Usman Naseem. 2025c. Cross platform multimodal retrieval augmented distillation for code-switched content understanding. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2042–2051.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision.