Multimodal Transformer Framework for Multilingual Harmful Meme Classification

Charmathi Rajkumar[1], **Malliga Subramaniyan**[1], **Bharathi Raja Chakravarthi** [2]
[1]Kongu Engineering College, Tamil Nadu, India
[2]Unit for Inclusive AI, School of Computer Science & Data Science Institute,
University of Galway, Ireland

Abstract

The rapid expansion of social media platforms has led to a significant increase in the spread of harmful content, including misogynistic, homophobic, and transphobic memes. Detecting such content is challenging because memes often combine textual and visual elements and frequently appear in multilingual and culturally diverse contexts. This study proposes a multimodal transformer-based framework for multilingual harmful meme classification that integrates textual and visual representations to improve detection performance. The proposed architecture employs XLM-RoBERTa for multilingual text encoding and the Swin Transformer for hierarchical visual feature extraction. A cross-attention fusion mechanism is introduced to enable meaningful interaction between textual and visual modalities. The fused representation is then processed through a classification layer to perform multi-class prediction. Experiments are conducted across multiple datasets covering eight languages and three harmful content categories: misogyny, homophobia/transphobia, and hate speech. The model is evaluated using the macro-F1 score and demonstrates consistent improvements over baseline multimodal systems across both high-resource and low-resource languages. The results highlight the effectiveness of transformer-based multimodal architectures in capturing implicit and contextual harmful signals present in memes. The study contributes to the development of robust multilingual systems for harmful content detection and supports efforts toward creating safer and more inclusive online environments.

1 Introduction

The rapid growth of social media platforms has transformed how people communicate, share opinions, and express identities. However, this growth has also led to an increase in harmful content, including misogynic, homophobic, and transphobic expressions (Chakravarthi et al., 2024). The detection of these expressions is very difficult since the content is often implicit and sarcastic in nature. Detecting them is particularly challenging when the content is multilingual or multimodal, as meaning often emerges from a combination of text, imagery, and social context. Therefore, building robust computational systems that can identify harmful content across languages and modalities has become an important research direction.

Recent studies highlight the difficulty of identifying misogynistic and harmful memes. For instance, Chakravarthi et al. (2025) introduced the ToxiCN-MM dataset for Chinese harmful memes and emphasized the importance of contextual knowledge in multimodal detection. Shared tasks such as those organized by Chakravarthi et al. (2024) have extended this research to low-resource Dravidian languages, contributing annotated datasets to support multilingual and code-mixed hate speech detection. These efforts collectively demonstrate that harmful content detection requires models capable of understanding both language-specific and cross-modal relationships.

Hate speech and related abusive expressions often target individuals or groups based on characteristics such as gender, caste, migration status, race, or sexual orientation and are frequently embedded within dynamic and multilingual online discourse. Recent efforts highlight the growing need for robust systems capable of handling culturally sensitive and multilingual harmful content (Rajiakodi et al., 2025). Furthermore, Kumaresan et al. (2025) emphasizes the importance of fine-grained modeling to identify harmful spans within social media text.

This paper proposes a transformer-based multimodal framework for harmful meme classification that integrates XLM-RoBERTa (XLM-R) for multilingual textual representation and Swin Transformer for visual feature extraction. The textual

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 99–107

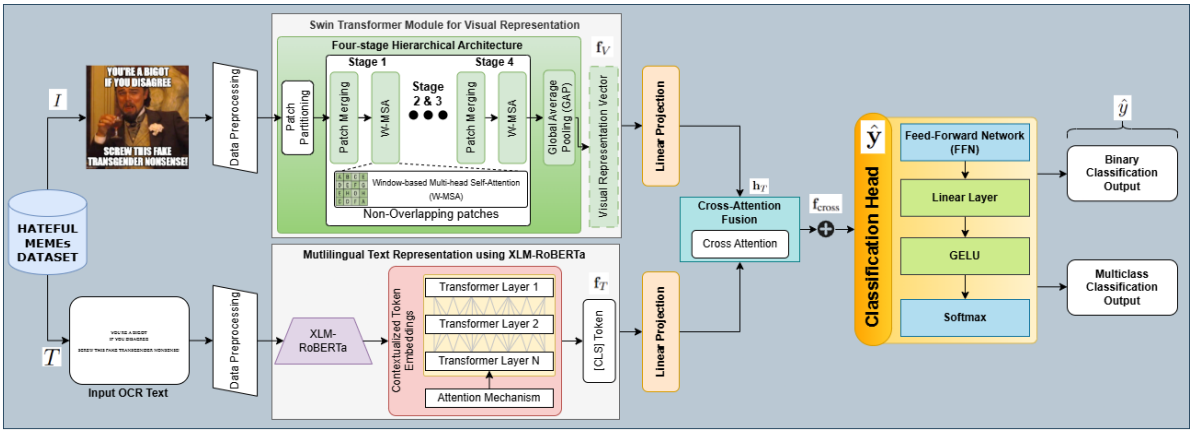

Figure 1: Our proposed architecture diagram

and visual embeddings produced by these models are combined using a cross-attention fusion mechanism to enable meaningful interaction between the two modalities. The fused multimodal representation is then passed to a classification layer to perform multi-class prediction. By jointly modeling linguistic context and visual semantics, the proposed architecture aims to capture both explicit and implicit harmful signals present in multilingual memes, enabling robust detection of abusive content across diverse languages and contributing toward safer and more inclusive online environments.

2 Related Work

In K and B (2025), ResNet-50 was employed for visual feature extraction and BERT for textual representation. Their results demonstrated that combining visual and textual representations improves performance over unimodal approaches. Rahman et al. (2025b) explored transformer-based architectures such as CharBERT and CLIP in different fusion strategies. Recent work Rajiakodi et al. (2026) highlights the importance of transformer-based models and balanced evaluation strategies for abuse detection in social media contexts.(Hossan et al., 2025) proposed a multimodal fusion-based framework for misogynistic meme detection in Tamil and Malayalam using machine learning, deep learning, and transformer-based architectures. Their study combined textual models such as BERT, MuRIL, and mBERT with visual encoders including ResNet50 and DenseNet121 through feature-level and decision-level fusion strategies. Experimental results showed that BERT+ResNet50 and MuRIL+ResNet50 achieved strong performance, demonstrating the effectiveness of multimodal transformer-based approaches for harmful meme classification in low-resource languages.

Rajiakodi et al. (2026) underscore the challenge of detecting women-targeted abusive content in Tamil social media. The study highlights the effectiveness of transformer-based models and macro-F1-based evaluation in low-resource settings. These findings reinforce the need for language-aware and context-sensitive abusive content detection systems. Sayma et al. (2025) focused on identifying misogynistic memes in Malayalam and achieving the macro F1 score of 0.8. In addition, Kumaresan et al. (2025) extended the scope of harmful content detection by focusing on fine-grained span-based detection of homophobic and transphobic expressions. This highlighted the importance of effective multilingual transformer-based models in dealing with nuances of online abuse expressions. Wang and Markov (2024) used RoBERTa for textual encoding and Swin Transformer V2 for visual feature extraction, followed by a multilayer perceptron (MLP) for feature fusion and classification. Choi et al. (2024) demonstrates that effective multimodal systems can be built through efficient knowledge transfer without large-scale multimodal pretraining, reducing computational cost while maintaining strong performance.

3 Proposed Method

We propose a multimodal transformer-based architecture for multilingual harmful meme classification that jointly models textual semantics and visual context through cross-attention fusion. Given a meme sample $\mathcal{M} = (T, I)$, where T denotes the textual component (OCR-extracted text) and I denotes the associated image, the objective is to

Task	Languages	Train	Dev	Test	Total
Misogyny	Tamil	1,136	284	356	1,176
	Malayalam	640	160	200	1,000
	Chinese	1,190	170	340	2,500
Homophobia/transphobia	Hindi	640	160	200	1,000
	English	450	110	140	700
	Chinese	760	190	240	1,190
Hate speech	Hindi	900	300	770	1,970
	Bodo	400	150	350	900

Table 1: Dataset distribution across Languages

predict a class label $\hat{y} \in \mathcal{Y}$. Here $\mathcal{Y}$ represents the task-specific label space corresponding to harmful meme categories (e.g., misogyny, homophobia, transphobia or hate speech), depending on the dataset and task formulation.

The overall prediction pipeline can be expressed as:

$$\hat{y} = \text{Classifier}\Big(\text{CrossAttn}\big(\text{XLM-R}(T),\ \text{SwinT}(I)\big)\Big) \tag{1}$$

The architecture consists of three principal modules: **(i)** multilingual textual encoding via XLM-RoBERTa, **(ii)** hierarchical visual representation via the Swin Transformer, and **(iii)** a cross-attention multimodal fusion mechanism followed by a classification head.

3.1 Multilingual Text Representation using XLM-RoBERTa

To encode multilingual textual information, we employ XLM-RoBERTa (Conneau et al., 2020), a transformer-based cross-lingual language model pretrained on large-scale multilingual corpora. Given an input token sequence $T = \{w_1, w_2, \ldots, w_n\}$, the model produces contextualized token embeddings through stacked self-attention layers. The sentence-level representation is obtained from the `[CLS]` token of the final encoder layer:

$$\mathbf{f}_T = \text{XLM-R}(T) \in \mathbb{R}^{d_T} \tag{2}$$

where d_T denotes the hidden representation dimension.

Self-attention within each encoder layer computes contextual interactions as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value projections of token embeddings, and d_k is the projection dimension. This mechanism allows the model to capture contextual dependencies such as sarcasm, implicit insults, and culturally grounded harmful expressions across languages.

3.2 Visual Representation using Swin Transformer

For the visual modality $I \in \mathbb{R}^{H\times W\times 3}$, we employ the Swin Transformer (Liu et al., 2021). The model partitions the image into non-overlapping patches and processes them through hierarchical transformer stages.

Self-attention is computed within local windows of size $M \times M$ patches using Window-based Multi-head Self-Attention (W-MSA). A learnable relative position bias $\mathbf{B}$ is incorporated to encode spatial relationships:

$$\text{W-MSA}(\mathbf{Z}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V} \tag{4}$$

Patch merging layers progressively reduce spatial resolution while increasing feature dimensionality, forming a four-stage hierarchical architecture that captures both local visual patterns and global semantic structure.

The final visual representation is obtained via global average pooling over the final stage output:

$$\mathbf{f}_V = \text{SwinT}(I) \in \mathbb{R}^{d_V} \tag{5}$$

3.3 Cross-Attention Multimodal Fusion

To model interactions between textual and visual modalities, both feature vectors are first projected into a shared latent space of dimension d_f:

$$\mathbf{h}_T = \mathbf{f}_T\mathbf{W}_T, \quad \mathbf{h}_V = \mathbf{f}_V\mathbf{W}_V \tag{6}$$

Parameter	Value
Epochs	20
Learning rate	2×10^{-5}
Batch size	16
Optimizer	AdamW
Dropout	0.1
Max sequence length	128
Image resolution	224×224

Table 2: Training hyperparameters

where $\mathbf{W}_T \in \mathbb{R}^{d_T \times d_f}$ and $\mathbf{W}_V \in \mathbb{R}^{d_V \times d_f}$ are learnable projection matrices.

Cross-attention is then applied where the textual representation acts as the *query*, while the visual representation provides the *key* and *value*. This allows the model to highlight visual features relevant to the textual semantics:

$$\mathbf{f}_{\text{cross}} = \text{softmax}\left(\frac{(\mathbf{h}_T\mathbf{W}^Q)(\mathbf{h}_V\mathbf{W}^K)^\top}{\sqrt{d_f}}\right)(\mathbf{h}_V\mathbf{W}^V) \tag{7}$$

where $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ denote the learnable query, key, and value projection matrices.

The cross-attended representation is combined with the textual representation using a residual connection and layer normalization to produce the final multimodal representation:

$$\mathbf{f}_{\text{fused}} = \text{LayerNorm}(\mathbf{h}_T + \mathbf{f}_{\text{cross}}) \in \mathbb{R}^{d_f} \tag{8}$$

3.4 Classification Head

The fused representation $\mathbf{f}_{\text{fused}}$ is passed through a feed-forward network with GELU activation to produce the final prediction:

$$\hat{\mathbf{y}} = \text{softmax}(\text{GELU}(\mathbf{f}_{\text{fused}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2) \tag{9}$$

The model is trained end-to-end using the standard cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{|\mathcal{Y}|} y_c \log \hat{y}_c \tag{10}$$

where y_c denotes the ground-truth indicator for class c and $\hat{y}_c$ represents the predicted probability. This formulation supports both binary and multi-class classification depending on the number of categories in $\mathcal{Y}$.

4 Experimental Setup

4.1 Datasets

We evaluate our framework across three harmful meme classification tasks spanning eight languages, covering both high-resource and low-resource settings. Each sample consists of a meme image paired with its associated text, and labels are assigned at the meme level for binary or multi-class classification.

MDMD: Misogyny Detection Meme Dataset (Ponnusamy et al., 2024) covers Tamil and Malayalam, and **CMMD: Chinese Misogynistic Meme Dataset** (Chakravarthi et al., 2025) covers Chinese, addressing harmful content targeting women across diverse linguistic and cultural contexts.The dataset was collected from different shared tasks and consists of meme samples annotated with binary labels such as misogynistic and non-misogynistic.

Homophobia/Transphobia Meme Detection Dataset spans Hindi, English, and Chinese, targeting memes that express discriminatory content toward LGBTQ+ communities.The dataset consists of meme samples annotated with labels such as homophobic/transphobic, and non-anti-LGBT categories. [1].

Hate Speech Meme Detection Dataset (Ghosh et al., 2026) covers Hindi and Bodo. Bodo represents an extremely low-resource language, which introduces additional challenges for cross-lingual generalization.

The full dataset statistics across all tasks and languages are summarized in Table 1. Note that the Chinese datasets used in the misogyny and homophobia/transphobia tasks originate from different task annotations.

4.2 Implementation Details

All experiments are implemented in PyTorch and executed on a single NVIDIA A100 GPU. For textual encoding, we initialize the model from the pretrained `xlm-roberta-base` checkpoint. For visual encoding, we use the pretrained `swin-base-patch4-window7-224` checkpoint. Both encoders are fine-tuned jointly during training.

Input images are resized to 224×224 pixels and normalized using ImageNet statistics. Text inputs are tokenized using the XLM-RoBERTa tokenizer with a maximum sequence length of 128 tokens.

[1] https://www.codabench.org/competitions/11335/

The shared multimodal projection dimension is set to $d_f = 768$.

The key training hyperparameters are summarized in Table 2.

4.3 Evaluation Metric

We report the **macro-averaged F1 score (Macro-F1)** as the primary evaluation metric across all tasks and languages.

Macro-F1 computes the F1 score independently for each class and then averages them with equal weight:

$$\text{Macro-F1} = \frac{1}{|\mathcal{Y}|} \sum_{c=1}^{|\mathcal{Y}|} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \qquad (11)$$

where P_c and R_c denote the precision and recall for class c, respectively. This metric is particularly suitable for imbalanced datasets because it assigns equal importance to all classes regardless of their frequency.

5 Results and Discussion

Tables 3, 4, and 5 present the macro-F1 scores of our proposed model in comparison with baseline systems across three datasets: misogyny, homophobia/transphobia, and hate speech.

5.1 Misogyny Detection

Our proposed model demonstrated strong performance across all three language tracks in the misogyny detection task mentioned in Table 5. For Chinese, our model achieved a macro-F1 score of **0.8715**, In the Malayalam track, our model achieved **0.856**, demonstrating competitive cross-lingual performance despite the domain gap between multilingual pretraining and English-specific meme semantics. The most significant improvement was observed in Tamil, where our model achieved **0.7619**, substantially outperforming the second-best system at 0.7351. The consistent gains across all three languages highlight the effectiveness of XLM-RoBERTa in capturing multilingual semantic context and Swin Transformer in extracting hierarchical visual features through an effective multimodal fusion strategy.

5.2 Homophobia Detection

In the homophobia detection task, our proposed model similarly demonstrated consistent and competitive performance across all language tracks mentioned in Table 4. For Chinese, the model achieved a macro-F1 score of **0.923**, effectively capturing subtle homophobic cues embedded in both textual and visual modalities. In English, our model achieved **0.7519**, reflecting the robust cross-lingual transfer capabilities of the multimodal framework. For Hindi, the model achieved **0.7825**, further confirming the strength of multilingual pretraining for low-resource language settings. Across all tracks, the fusion of XLM-RoBERTa and Swin Transformer proved effective in detecting homophobic content in memes.

5.3 Hate Speech Detection

Our proposed framework demonstrated superior performance across both language tracks in the hate speech detection task mentioned in Table 3. For Hindi, our model achieved a macro-F1 score of **0.7615**, outperforming all baseline systems. In the Bodo track, our model achieved **0.7830**, substantially surpassing all competing approaches. The consistent improvements across both languages highlight the advantage of cross-attention fusion over concatenation-based approaches, with particularly strong gains in Bodo, a low-resource language, demonstrating the robustness of our framework in limited data settings.

5.4 Error Analysis

Despite strong overall results, our model occasionally misclassified memes that relied on cultural or language-specific implicit cues not captured by surface-level text or visual features alone. Sarcastic or ironic content, where the text appears neutral but the visual context conveys offensive intent, posed a particular challenge across all language tracks. Future work may explore the incorporation of external cultural knowledge bases or cross-modal attention mechanisms to further address these limitations.

Conclusions

This study presents a multimodal transformer-based framework for multilingual harmful meme classification. The proposed architecture integrates XLM-RoBERTa for multilingual textual representation and the Swin Transformer for hierarchical visual feature extraction. A cross-attention fusion mechanism is applied to combine textual and visual embeddings, enabling the model to capture complex interactions between language and imagery in memes. Experimental evaluation across multiple datasets and eight languages demonstrates that

Language	Team Name	Models	Macro F1
Hindi	CSIS BITS Pilani	XLM-RoBERTa + CLIP (Concat)	0.56780
	NLP Fusion	Hindi-RoBERTa + ResNet-34 (Concat)	0.62400
	FiRC-NLP	LLM + XLM-R + SigLIP (Ensemble)	0.65710
	Our proposed method	**XLM-RoBERTa + Swin (CA)**	**0.76150**
Bodo	CSIS BITS Pilani	XLM-RoBERTa + CLIP (Concat)	0.59970
	FiRC-NLP	LLM + XLM-R + SigLIP (Ensemble)	0.62220
	NLP Fusion	mBERT + ResNet-34 (Concat)	0.63130
	Our proposed method	**XLM-RoBERTa + Swin (CA)**	**0.78300**

Table 3: Comparison of proposed method with existing methods in Hate speech task

Language	Team Name	Method	Macro F1
English	Susmitha	XLM-R + CLIP-ViT (Gated)	0.6121
	SigJBS	Qwen2-VL + LoRA	0.6396
	BiasBreakers	CLIP + Neural Classifier (Concat)	0.7384
	Our proposed method	**XLM-RoBERTa + Swin (CA)**	**0.7519**
Hindi	MemeSentinel	CLIP + Gated Fusion	0.6068
	MemeScouts	VLM Prompting + Random Forest	0.6426
	BiasBreakers	CLIP + Neural Classifier (Concat)	0.7385
	Our proposed method	**XLM-RoBERTa + Swin (CA)**	**0.7825**
Chinese	Susmitha	XLM-R + CLIP-ViT (Gated)	0.7371
	MemeScouts	VLM Prompting + Random Forest	0.7527
	MemeSentinel	CLIP + Gated Fusion	0.7535
	Our proposed method	**XLM-RoBERTa + Swin (CA)**	**0.923**

Table 4: Comparison of proposed method with existing methods in homophobia/transphobia Task

the proposed framework achieves strong performance in detecting harmful meme categories such as misogyny, homophobia, transphobia, and hate speech. The model consistently outperforms several baseline systems in terms of macro-F1 score across both high-resource and low-resource language settings. These results indicate that combining multilingual language models with transformer-based visual representations can effectively capture both explicit and implicit harmful signals embedded in memes. The findings emphasize the importance of multimodal and multilingual approaches for addressing harmful content in online environments. By jointly modeling visual and textual context, the proposed framework improves the ability to detect abusive expressions that may not be identifiable using single-modality methods. Future research can focus on enhancing multimodal fusion mechanisms, incorporating cultural knowledge representations, and developing more computationally efficient architectures to support large-scale real-world deployment.

Limitations and Future Work

Despite the promising performance of the proposed framework, several limitations remain. The current approach relies on a relatively simple fusion mechanism to combine textual and visual features, which may not fully capture complex interactions between the two modalities. Additionally, the model is built on large transformer-based architectures such as XLM-RoBERTa and Swin Transformer, which require significant computational resources for training and inference, potentially limiting their deployment in low-resource environments. The performance of the system also depends on the quality, balance, and diversity of the training data. If certain harmful patterns, cultural contexts, or language variations are underrepresented, the model may struggle to generalize to unseen content.

Future work can address these limitations by exploring more advanced multimodal fusion strate-

Language	Team name	Method	Macro_f1
Chinese	SSNCSE (K and B, 2025)	BERT + ResNet (Concat)	0.7034
	CUET_12033(Rahman et al., 2025b)	CharBERT + BiLSTM(GMU)	0.7089
	CVF_NITT(T and K, 2025)	CLIP (Early Fusion)	0.7362
	CUET's_White_Walkers(Rahman et al., 2025a)	BERT + ResNet(Early Fusion)	0.8542
	Our proposed method	**XLM-R + Swin(CA)**	**0.8715**
Malayalam	Code_Conquerors(Rao et al., 2025)	BERT + ViT(Concat)	0.7561
	Fired_from_NLP(Chowdhury et al., 2025)	EffNet + mBERT (Concat)	0.8037
	CUET-NLP_Big_O(Hossan et al., 2025)	MuRIL + EffNet (FC)	0.8253
	byteSizedLLM(Manukonda and Kodali, 2025)	XLM-R+ ResNet(BiLSTM)	0.8391
	Our proposed method	**XLM-R + Swin(CA)**	**0.856**
Tamil	Team_Strikers(Shanmugavadivel et al., 2025a)	LSTM + ResNet (CNN-LSTM)	0.6477
	Code_Conquerors(Rao et al., 2025)	BERT + CLIP (Concat)	0.6641
	InnovationEngineers(Shanmugavadivel et al., 2025b)	BERT + EffNet(VLM)	0.6878
	MNLP(Chauhan and Kumar, 2025)	XLM-R + ViT(Concat)	0.7351
	Our proposed method	**XLM-R + Swin (CA)**	**0.7619**

Table 5: Comparison of proposed method with existing methods in the Misogyny task

gies, such as deeper cross-modal attention mechanisms or end-to-end multimodal training frameworks. Incorporating data augmentation techniques and expanding multilingual datasets could further improve model robustness and generalization across diverse linguistic and cultural contexts. In addition, lightweight model architectures or efficient training strategies may help reduce computational requirements and enable broader real-world deployment.

References

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025. An overview of the misogyny meme detection shared task for Chinese social media. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 200–208, Naples, Italy. Unior Press.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of shared task on multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 139–144, St. Julian's, Malta. Association for Computational Linguistics.

Shraddha Chauhan and Abhinav Kumar. 2025. MNLP@DravidianLangTech 2025: Transformer-based multimodal framework for misogyny meme detection. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 248–253, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Dongha Choi, Jung-jae Kim, and Hyunju Lee. 2024. TransferCVLM: Transferring cross-modal knowledge for vision-language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16733–16746, Miami, Florida, USA. Association for Computational Linguistics.

Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury, Anik Mahmud Shanto, Jidan Al Abrar, and Hasan Murad. 2025. Fired_from_NLP@DravidianLangTech 2025: A multimodal approach for detecting misogynistic content in Tamil and Malayalam memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 459–464, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Koyel Ghosh, Mithun Das, Sumukh Patel, Nilotpal Bhandary, Alloy Das, Animesh Mukherjee, Sandip Modha, Debasis Ganguly, Utpal Garain, Sylvia Jaki, and Thomas Mandl. 2026. Overview of the hasoc track at fire 2025: Abusive meme identification — shadows behind the laughter. In *Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '25, page 28–31, New

S.No	Image	Translation	GT / Pred	Error Type	Explanation
1		“Why are so many people in relationships?”	Humorous / Neutral	Implicit Sarcasm Detection	Humor is conveyed primarily through Tamil reaction images rather than explicit sarcastic wording.
2		Girl asks for a maroon stone-work mask while shopping for Perunnal.	Misogynistic / Non-misogynistic	Gender Stereotype Misclassification	Indirect sexist humor is expressed through exaggerated shopping stereotypes.
3		“Two ladies, this way please.”	Homophobic / Non-hateful	Contextual LGBTQ+ Hate Misclassification	Hidden discriminatory intent is conveyed through contextual implication.
4		“If you are under me, then 50.”	Offensive / Neutral	Implicit Sexual Harassment Detection	Indirect sexually suggestive sarcasm was not correctly identified.
5		“Send a screenshot of your Chrome history.”	Neutral / Offensive	Contextual Humor Misclassification	Playful internet humor was incorrectly classified as offensive.

Table 6: Qualitative Error Analysis of Harmful Meme Classification. GT: ground truth, Pred: predicted label

York, NY, USA. Association for Computing Machinery.

Md. Refaj Hossan, Nazmus Sakib, Md. Alam Miah, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A multimodal fusion-based approach for identifying misogyny memes. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 427–434, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Sreeja K and Bharathi B. 2025. SSNCSE@LT-EDI-2025: Detecting misogyny memes using pretrained deep learning models. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 1–5, Naples, Italy. Unior Press.

Prasanna Kumar Kumaresan, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Paul Buitelaar, Malliga Subramanian, and Kishore Kumar Ponnusamy. 2025. Overview of homophobia and transphobia span detection in social media comments. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 229–234, Naples, Italy. Unior Press.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Durga Prasad Manukonda and Rohith Gowtham Kodali. 2025. byteSizedLLM@DravidianLangTech 2025: Multimodal misogyny meme detection in low-resource Dravidian languages using transliteration-aware XLM-RoBERTa, ResNet-50, and attention-BiLSTM. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 86–91, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya Rajiakodi, Prasanna Kumar Kumaresan, Sajeetha

Thavareesan, Bhuvaneswari Sivagnanam, Anshid K.A, Susminu S Kumar, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. From laughter to inequality: Annotated dataset for misogyny detection in Tamil and Malayalam memes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7480–7488, Torino, Italia. ELRA and ICCL.

Md. Mizanur Rahman, Jidan Al Abrar, Md. Siddikul Imam Kawser, Ariful Islam, Md. Mubasshir Naib, and Hasan Murad. 2025a. CUET's_White_Walkers@LT-EDI 2025: Racial hoax detection in code-mixed on social media data. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 63–67, Naples, Italy. Unior Press.

Mehreen Rahman, Faozia Fariha, Nabilah Tabassum, Samia Rahman, and Hasan Murad. 2025b. CUET_12033@LT-EDI-2025: Misogyny detection. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 127–132, Naples, Italy. Unior Press.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Shunmuga Priya Muthusamy Chinnan, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, Balasubramanian Palani, Kogilavani Shanmugavadivel, Abirami Murugappan, and Charmathi Rajkumar. 2025. Findings of the shared task caste and migration hate speech detection. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 215–221, Naples, Italy. Unior Press.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Kathiravan R., Rajalakshmi andPannerselvam, Bhuvaneshwari Sivagnanam, Jananayagan V, Charmathi Rajkumar, R Ramesh Kannan, and Bharathi Raja Chakravarthi. 2026. From Comments to Harm: A Findings Report on Abusive Tamil Text Targeting Women on Social Media - DravidianLangTech@ACL 2026. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Pathange Omkareshwara Rao, Harish Vijay V, Ippatapu Venkata Srichandra, Neethu Mohan, and Sachin Kumar S. 2025. Code_Conquerors@DravidianLangTech 2025: Multimodal misogyny detection in Dravidian languages using vision transformer and BERT. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 283–287, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech 2025: A multimodal transformer-based approach for detecting misogynistic memes in Malayalam language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 472–477, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Mohamed Arsath H, Ramya K, and Ragav R. 2025a. TEAM_STRIKERS@DravidianLangTech2025: Misogyny meme detection in Tamil using multimodal deep learning. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 619–623, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Kogilavani Shanmugavadivel, Malliga Subramanian, Pooja Sree M, Palanimurugan V, and Roshini Priya K. 2025b. InnovationEngineers@DravidianLangTech 2025: Enhanced CNN models for detecting misogyny in Tamil memes using image and text classification. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 162–166, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Radhika K T and Sitara K. 2025. CVF-NITT@LT-EDI-2025:MisogynyDetection. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 47–53, Naples, Italy. Unior Press.

Yeshan Wang and Ilia Markov. 2024. CLTL@multimodal hate speech event detection 2024: The winning approach to detecting multimodal hate speech and its targets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 73–78, St. Julians, Malta. Association for Computational Linguistics.