MoEzhi-Guru: Multi-Perspective Misogynistic Meme Detection via Multimodal Mixture of Experts

Sekar M[*] [†] **Pranava Mohan**[*] **A. Santhanavijayan**
Department of Computer Science and Engineering
National Institute of Technology Tiruchirappalli, India
{406124006,106125084}@nitt.edu, vijayana@nitt.edu

Abstract

Multimodal content moderation in online social spaces is a critical safety challenge, yet existing systems are heavily dominated by monolithic cultural biases and perform poorly on underrepresented languages. In this paper, we study culturally grounded misogynistic meme detection under a novel Multi-Perspective evaluation framework, which simultaneously assesses models across Indian, Chinese, and Western cultural lenses. We propose **MoEzhi-Guru** (drawing on the Dravidian term *mozhi* for language and the Indic term *guru* for expert/guide), an adaptable, resource-efficient Multimodal Mixture of Experts (MoE) model designed to handle complex cross-cultural and cross-lingual shifts. The system encodes input using mDeBERTa-v3 and SigLIP to extract rich multilingual text and global visual semantics, then fuses and routes these features through a custom MoE layer utilizing dense soft routing. This gating mechanism dynamically weights specialized expert subnetworks, capturing localized socio-cultural nuances before predicting labels through culture-specific classification heads. We evaluate our framework in four typologically diverse language subsets: Tamil and Malayalam (underrepresented Dravidian scripts), Chinese (containing colloquial and code-switched text), and English. Our system demonstrates robust cross-lingual generalization, yielding an overall average score of 0.8515 Macro-F1 and consistently exceeding a Gemma3-4b baseline by more than 0.42 Macro-F1 points. This work provides a scalable, inclusive foundation for culturally aware safety systems in low-resource and multilingual environments.

1 Introduction

Misogyny online is increasingly spread through multimodal formats like memes, which layer text over images to convey humor, sarcasm, or ideological beliefs. Detecting this kind of harmful content is difficult because the misogynistic signal is often implicit and grounded in specific cultural norms. Many current multimodal models struggle to generalize across cross-lingual and cross-cultural distribution shifts.

Cross-cultural misogynistic meme detection directly addresses this gap, benchmarking models across Indian, Chinese, and Western (English) cultural contexts. In this paper, we outline our approach to the Cross-Culture Evaluation task through **MoEzhi-Guru**. Since different cultures require different interpretive lenses, we designed a Mixture of Experts (MoE) architecture. Rather than relying on a single dense layer to process the fused multimodal features, our MoE routing mechanism lets the model dynamically allocate specialized capacity depending on the input, with the explicit goal of improving cross-cultural robustness.

2 Related Work

2.1 Multimodal Hate Speech and Misogyny Meme Detection

Early computational approaches to online abuse detection relied heavily on text-only architectures, utilizing classical machine learning or early transformer-based text encoders. However, internet memes violate text-only assumptions, as their meaning is deeply tied to multimodal interactions. To encourage the development of models capable of true multimodal reasoning, Kiela et al. introduced the *Hateful Memes Challenge* (Kiela et al., 2020). This dataset features "benign confounders", i.e., samples where either the image or the text is modified to flip the label, ensuring that unimodal shortcuts (e.g., detecting specific offensive words or visual triggers) are ineffective.

Following this, the SemEval-2022 Task 5 on *Multimedia Automatic Misogyny Identification*

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 55–63
November 9, 2026 ©2026 Association for Computational Linguistics

(MAMI) (Fersini et al., 2022) focused specifically on gender-targeted online abuse. The competition challenged models to perform binary misogyny detection (Sub-task A) and fine-grained categorization (Sub-task B), including shaming, stereotyping, objectification, and violence. Traditional SOTA approaches for MAMI leveraged large-scale pre-trained models such as VisualBERT, Oscar, UNITER, and ERNIE-Vil (Fersini et al., 2022). However, these frameworks exhibit significant performance gaps when evaluated on non-Western and low-resource languages, due to the lack of representative training corpora.

2.2 Social Media Safety in Low-Resource Dravidian Languages

Low-resource Dravidian languages such as Tamil and Malayalam present severe challenges for standard multilingual models. These languages are typologically diverse, morphologically rich, and characterized by extensive code-mixing (e.g., blending Dravidian scripts with English text) and informal phonetic spelling. To establish digital safety benchmarks in these contexts, Ponnusamy et al. introduced the *Misogyny Detection Meme Dataset* (MDMD) (Ponnusamy et al., 2024).

Recent shared tasks organized under the *DravidianLangTech* and *LT-EDI* workshops (Chakravarthi et al., 2024, 2025b) have accelerated research in Dravidian social media safety. Models evaluated on Tamil and Malayalam misogyny tasks often combine multilingual textual embeddings from IndicBERT, mBERT, or LaBSE with spatial visual features from ResNet or EfficientNet backbones (Ratul et al., 2025). More recently, Rajkumar et al. proposed a dual-stream architecture pairing XLM-RoBERTa with the Swin Transformer via cross-attention fusion (Rajkumar et al., 2026), showing consistent improvements in capturing implicit and contextual harmful signals. Furthermore, Dravidian political meme tasks highlight the necessity of capturing complex socio-political context, sarcasm, and regional visual symbolism to mitigate negative transfer across distinct domains (Rajiakodi et al., 2026).

2.3 Mixture of Experts and Gating Paradigms

The Mixture of Experts (MoE) paradigm, originally introduced by Shazeer et al. (Shazeer et al., 2017), scales model capacity by routing inputs dynamically to a set of specialized sub-networks (experts) via a trainable gating network. By executing only a handful of experts per forward pass (conditional computation), MoE architectures drastically scale model parameters without a proportional increase in FLOPS. Fedus et al. further simplified this framework with the *Switch Transformer* (Fedus et al., 2022), which utilizes a top-1 routing strategy to reduce communication overhead, simplify routing mechanics, and stabilize low-precision training.

While sparse MoE models are highly efficient for massive language modeling tasks, they are prone to training instabilities and expert collapse when applied to smaller, highly imbalanced multimodal safety datasets. In this paper, we address this by adopting dense soft routing. Rather than gating out experts, our router computes a soft probability distribution over all available subnetworks on every forward pass. This configuration maintains stable gradient flow, allows cooperative specialization across experts, and is intended to reduce gradient interference among divergent socio-cultural classification heads.

2.4 Multimodal Representation Learning

Modern multimodal architectures rely heavily on high-quality representation encoders. For multilingual natural language understanding, mDeBERTa-v3 (He et al., 2021) introduces replaced token detection (RTD) combined with gradient-disentangled embedding sharing (GDES) to decouple generator and discriminator gradients, achieving superior semantic and cross-lingual representation efficiency over XLM-RoBERTa. Its multilingual variant is pre-trained on the CC100 corpus (He et al., 2021), which includes Tamil and Malayalam, and it has been applied to abusive-comment detection in code-mixed Tamil, where it outperformed mBERT, MuRIL, and XLM-RoBERTa (Subramanian et al., 2025).

For the visual modality, contrastive vision-language pre-training (CLIP) has historically been the dominant approach. However, SigLIP (Zhai et al., 2023) proposes a simpler, pairwise sigmoid loss that replaces standard softmax normalization. SigLIP decouples batch size from the learning objective, showing exceptional zero-shot retrieval efficiency and robust multilingual visual-textual alignment on small hardware budgets. By fusing the representation capacity of mDeBERTa-v3 and SigLIP within our soft-routed MoE framework, our system captures subtle visual-textual dynamics across highly diverse cultural contexts.

3 Methodology

Our proposed architecture, **MoEzhi-Guru**, independently encodes text and images, fuses the two modalities, then passes the representation through an MoE layer before routing it to culture-specific classification heads. Figure 1 illustrates the high-level workflow of our system.

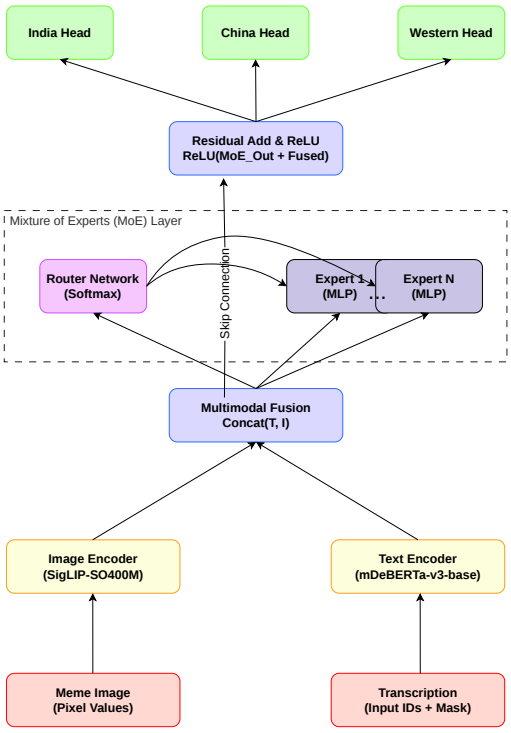


Figure 1: Overview of our MoEzhi-Guru Multimodal Mixture of Experts (MoE) architecture for cross-cultural meme detection.

3.1 Multimodal Encoders

Text Encoder: We use `microsoft/mdeberta-v3-base` (He et al., 2021) to extract rich, multilingual contextual text embeddings from meme transcriptions. Text features T are taken as the CLS-token representation (the first token of the final hidden layer, `last_hidden_state[:, 0, :]`).

Image Encoder: For the visual modality, we use `google/siglip-so400m-patch14-384` (Zhai et al., 2023), taking advantage of its strong zero-shot image representation capabilities. Image features I are obtained from the model's global pooled output (`pooler_output`). We enable gradient checkpointing on both encoders to keep memory usage manageable during training.

3.2 Feature Fusion and Mixture of Experts

We concatenate the text encoder's CLS-token representation T and the image encoder's pooled output I to form a unified multimodal representation:

$$F = \text{Concat}(T, I) \tag{1}$$

To capture divergent cultural interpretations, this fused representation F is passed through an MoE layer. Our MoE consists of a routing network and N expert networks. Each expert is a two-layer MLP with architecture $d \to d/2 \to d$, where $d = \dim(F)$ is the concatenated feature dimension. Specifically, the first projection maps down to $d/2$, followed by a GELU activation, a Dropout(0.2) layer, and a second projection back to d. The router uses *dense (soft) routing*: all N experts run on every forward pass, and the router produces a probability distribution over them via softmax, optimized end-to-end via the classification loss. The final output is the weighted sum of all experts' outputs:

$$M = \sum_{i=1}^{N} \text{softmax}(W_r F)_i \, E_i(F) \tag{2}$$

Here W_r denotes the learned weights of the routing linear layer, and E_i denotes the output of the i-th expert. Unlike sparse top-k routing, no expert is ever fully gated out; every expert contributes to every prediction, proportionally to the router's confidence. A residual connection followed by a ReLU activation ensures stable gradient flow:

$$H = \text{ReLU}(M + F) \tag{3}$$

3.3 Classification Heads

The final representation H feeds into three independent linear classification heads, one for each target cultural perspective: India, China, and Western. The model is trained using Binary Cross-Entropy with Logits Loss. Because the three heads are linear maps over the same representation H, they are equivalent to a single linear layer with three outputs and add no perspective-specific capacity; any perspective-dependent behaviour must therefore arise in the shared MoE layer.

4 Experimental Setup

4.1 Data and Preprocessing

We evaluate our approach on a consolidated benchmark combining the English MAMI dataset (Fersini et al., 2022), the Tamil

and Malayalam subsets from MDMD (Ponnusamy et al., 2024), and the Chinese CMMD dataset (Chakravarthi et al., 2025a). Images are preprocessed using the standard SigLIP image processor, and texts are tokenized via the mDeBERTa tokenizer with a maximum sequence length of 128 tokens. Missing perspective labels are masked with -1.0 and excluded from the loss calculation.

Throughout this paper, a cultural *perspective* (or *lens*) denotes the label assigned by an annotator group of a given national or regional background (Indian, Chinese, or Western). This is a coarse, national-level operationalization of culture that does not capture sub-national, linguistic, or individual variation. The perspective labels and their annotation protocol are those released by the CC-MMD 2026 organizers; we do not re-annotate any data.

Table 1: Dataset statistics per language and split. The percentage columns report the proportion of samples labeled as misogynistic under each cultural perspective. The perspective corresponding to a subset's native annotators is denoted with †.

Language	Split	#Samples	Indian (%)	Chinese (%)	Western (%)
Tamil	Train	1,137	24.0†	21.3	23.6
	Dev	284	26.4†	24.6	24.3
Malayalam	Train	640	40.0†	35.2	34.4
	Dev	160	40.0†	34.4	33.1
English	Train	9,395	30.2	52.0	50.0†
	Dev	600	29.8	50.7	49.8†
Chinese	Train	1,190	19.2	29.3†	25.3
	Dev	170	18.8	27.6†	18.2

As detailed in Table 1, the combined corpus comprises 13,576 samples across both splits. The distribution is dominated by the English (MAMI-derived) subset, which accounts for 9,995 samples ($\sim$74% of the corpus), while Tamil, Malayalam, and Chinese each contribute between a few hundred and just over a thousand instances per split. This substantial cross-lingual size disparity motivates the differential learning-rate strategy outlined in §4.2, enabling the randomly initialized Mixture-of-Experts (MoE) routing layers and task-specific classification heads to adapt rapidly without overfitting the smaller language subsets.

The perspective-level label distributions reveal pronounced cross-cultural divergence that directly justifies our multi-perspective task formulation. For the English subset, native Western annotators identify 50.0% of the training samples as misogynistic closely aligned with Chinese cross-cultural annotators at 52.0%. In contrast, Indian annotators label only 30.2% of the identical content as misogynistic, reflecting a notable $\sim$20 percentage-point gap. A similar divergence appears in the Chinese subset: native annotators assign misogynistic labels to 29.3% of samples, whereas Indian annotators flag only 19.2%. In contrast, the Tamil and Malayalam subsets exhibit narrower cross-cultural variance (within 2-6 percentage points), demonstrating that annotator disagreement varies based on cultural context rather than following a uniform offset. Our MoE-based architecture is explicitly designed to model this contextual variability.

Finally, the evaluation protocol accommodates differences in perspective annotations across subsets: Tamil, Malayalam, and Chinese subsets are evaluated across all three cultural perspectives (Native, Indian/Chinese/Western where applicable), whereas the English (MAMI) subset is evaluated on the two non-native cross-cultural perspectives (Indian and Chinese) since native Western perspectives are already captured at source. Consequently, overall English performance is computed as a 2-perspective Macro-F1 average.

4.2 Training and Hyperparameter Tuning

We train for 6 epochs using AdamW with a cosine learning rate schedule and a 10% warmup phase. We use a *differential learning rate* strategy: the pretrained backbones (mDeBERTa-v3 and SigLIP) receive base_lr $\times$ 0.1, while the randomly initialized MoE layer and classification heads receive base_lr $\times$ multiplier, letting the new components adapt more quickly than the pretrained encoders. We train with a per-device batch size of 2 (effective batch size of 4 via 2-step gradient accumulation) on one NVIDIA V100 GPU, using mixed-precision training (FP16) through PyTorch's (Paszke et al., 2019) Autocast and GradScaler. Gradient norms are clipped to 1.0 at each optimizer step via `clip_grad_norm_`.

One notable aspect of our loss formulation is that summed per-task losses are divided by the gradient accumulation step count, rather than by the number of active cultural perspectives in a batch. Thus, samples with only one valid cultural label still have their effective loss contribution halved. This behavior was retained across all trials as part of the final tuned configuration.

We performed hyperparameter optimization using Optuna (Akiba et al., 2019) over 25 trials within a 48-hour budget. The search space in-

cluded the base learning rate (log-uniform over $[10^{-6}, 5\times10^{-5}]$), the MoE learning rate multiplier (uniform over $[1.0, 10.0]$), and the number of experts ($\{4, 6, 8\}$). Underperforming trials were pruned using Optuna's `MedianPruner`, and the study was checkpointed to an SQLite database for seamless resumption. The optimal setup used 4 experts, a base learning rate of 4.2841×10^{-5}, and an MoE learning-rate multiplier of 3.2128, producing an encoder-to-MoE learning rate ratio of approximately $1:32.1$.

4.3 Evaluation Metrics

For binary misogyny classification, let $\mathcal{C} = \{0, 1\}$ be the label set and let TP_c, FP_c, FN_c denote the true positives, false positives, and false negatives for class c.

Per-class Precision (P_c), Recall (R_c), and $F1$ ($F1_c$) are defined as:

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \quad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}$$
$$F1_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{4}$$

The *Macro-F1* score averages $F1_c$ equally across classes, removing class-frequency bias:

$$\text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F1_c. \tag{5}$$

The overall score for a language subset ℓ with applicable cultural perspectives $\mathcal{P}_\ell$ is the unweighted mean Macro-F1 across all perspectives:

$$\text{Score}_\ell = \frac{1}{|\mathcal{P}_\ell|} \sum_{p \in \mathcal{P}_\ell} \text{Macro-F1}_p. \tag{6}$$

For Tamil, Malayalam, and Chinese subsets $|\mathcal{P}_\ell| = 3$; for English $|\mathcal{P}_\ell| = 2$ (see §4.1). Mean accuracy across perspectives serves as a secondary tie-breaker metric.

5 Results

5.1 Comparison Against Baseline

Table 2 compares our system against a Gemma3-4b baseline. Baseline scores[1] are computed as the unweighted average of the per-culture Macro-F1 values reported for each language. Our architecture improves consistently over the baseline by 0.42 to 0.64 Macro-F1 points across all four language subsets. As these baseline figures were not reproduced under our training setup, this comparison serves as a reference point rather than a controlled comparison; the controlled architectural comparison is the ablation in §5.3.

Table 2: Comparison against a Gemma3-4b baseline (unweighted average Macro-F1). The English baseline uses the 2-perspective average (Indian + Chinese).

Language	Baseline	Ours	Δ
Tamil	0.3247	0.8427	**+0.518**
Malayalam	0.3251	0.9133	**+0.588**
Chinese	0.2283	0.8686	**+0.640**
English	0.3574	0.7815	**+0.424**

5.2 Per-Culture Results

Table 3 reports [2] the per-culture Macro-F1 and accuracy scores, providing the full breakdown across all cultural perspectives for each language subset.

Table 3: Per-culture Macro-F1 results.

Language	Perspective (Macro-F1)			**Overall**	
	Indian	Western	Chinese	**Macro-F1**	**Accuracy**
Tamil	0.820	0.851	0.857	**0.843**	0.783
Malayalam	0.925	0.925	0.890	**0.913**	0.908
Chinese	0.874	0.856	0.876	**0.869**	0.808
English*	0.796	-	0.767	**0.782**	0.760

*English evaluated on 2 perspectives only.

5.3 Ablation Study

To evaluate the contribution of each input modality and assess the incremental effect of the proposed Mixture of Experts (MoE) fusion architecture, we conduct a controlled ablation study. Specifically, our evaluation investigates two core research dimensions:

1. **Modality Contributions**: Dissecting the relative impact of individual modalities by comparing unimodal baselines (*Text-Only* utilizing `mDeBERTa-v3-base` and *Image-Only* utilizing `SigLIP-so400m`) against joint multimodal representations ($F = [\mathbf{e}_{\text{text}} \,\|\, \mathbf{e}_{\text{img}}]$).

2. **Fusion Architecture Efficiency**: Evaluating fusion mechanisms by comparing a conventional *Single Dense Fusion Layer* (a 2-layer

[1] Baseline results sourced from the official CC-MMD 2026 Challenge Tasks benchmark: https://sites.google.com/view/cc-mmd2026/challenge-tasks?authuser=0

[2] All reported benchmark metrics correspond to official evaluation results released and publicly available through the CC-MMD Challenge evaluation platform.

MLP without dynamic routing) against our proposed *MoE Fusion Layer* configured with soft-router gating across $N = 4$ expert sub-networks.

All ablation variants are trained under strictly identical experimental setups, optimization hyper-parameters (differential base learning rate base_lr $= 2 \times 10^{-5}$, AdamW optimizer, cosine decay schedule), and random initialization seeds ($S = 42$) to isolate architectural effects. For computational uniformity across ablation variants, we fix $N = 4$ experts. Experimental results across all target cultural contexts are reported in Table 4.

Necessity of Multimodal Integration As demonstrated in Table 4, unimodal models exhibit noticeable performance degradation when evaluated across diverse cultural perspectives. The *Text-Only* variant achieves an Overall Macro-F1 of 0.6984, effectively capturing textual transcripts but remaining blind to visual meme templates and non-verbal humor. Conversely, the *Image-Only* model achieves 0.7330 Overall Macro-F1. The image-only variant likely outperforms the text-only one because meme text is rendered inside the image and was not removed: SigLIP, pre-trained on large-scale image-text data, can read at least part of this overlaid text, so the image input carries both visual and partial textual signal. Fusing both modalities in *Dense Fusion* increases the overall Macro-F1 to 0.7472 (+0.0488 absolute / +6.99% relative over Text-Only and +0.0142 absolute / +1.94% relative over Image-Only), demonstrating that cross-modal synergy is indispensable for cross-cultural meme analysis.

Superiority of MoE Expert Routing over Dense MLP A critical comparison lies between the *Dense Fusion* baseline (0.7472 Overall F1) and our proposed *MoE Fusion* architecture (0.8491 Overall F1). Conventional dense fusion forces all cultural context heads to share a rigid, single set of transformation weights, inducing negative cross-cultural transfer. In contrast, the router-gated MoE layer dynamically allocates capacity across $N = 4$ specialized expert sub-networks. This dynamic gating yields substantial performance gains across all target cultural contexts, boosting the Overall Macro-F1 by $+0.1019$ absolute points (+13.64% relative over Dense Fusion). The largest gain is observed in the China context (0.8700 Macro-F1, $+0.1378$ over Dense Fusion), followed by the Western context (0.8653 Macro-F1, $+0.1094$ over Dense Fusion) and the India context (0.8120 Macro-F1, $+0.0585$ over Dense Fusion). We hypothesize that this gain stems from the router allocating capacity differently across cultural contexts, mitigating negative transfer between perspective heads; directly verifying expert-level specialization via routing analysis is left to future work.

5.4 Discussion

Our experiments reveal strong cross-cultural performance for all four language subsets. In **Tamil**, the model is less proficient in understanding from the native Indian viewpoint (0.820) compared to the Western (0.851) and Chinese (0.857) viewpoints due to higher variance among annotators in source-culture annotations. Our model has similar performance metrics for **Malayalam** and **Chinese** (within 0.04 Macro-F1 for all viewpoints), which means MoE routing works fine even though there are different class distributions. Lower performance of **English** can be attributed to the difficulty of predicting the meaning of memes for Indian and Chinese audiences.

Although the evaluation metric is calculated per-language separately, one of the strongest features of the MoE architecture we implemented is its domain-independent consistency. In order to test this assumption, we calculated the unweighted overall average Macro-F1 for all four languages. Our architecture obtained an average overall Macro-F1 score of 0.8515, indicating that the model successfully avoids performance degradation across disparate languages, generalizing well across scripts rather than specializing narrowly in a few. Because the English subset is scored on only two perspectives (§4.1), this average is a summary statistic rather than a uniform cross-lingual measure.

As demonstrated in our ablation study (§5.3), the Mixture of Experts layer outperforms a conventional dense layer, while our hyperparameter optimization phase (detailed in §4.2) provides further empirical insight into how the Mixture of Experts layer scales with capacity. During our Optuna search space evaluation of 4, 6, and 8 expert configurations, we observed that performance remains notably stable. The top development-set Macro-F1 scores for the 4, 6, and 8 expert setups were 0.8491, 0.8383, and 0.8425, respectively. This tight spread (less than 0.011) suggests that dense soft routing is robust to the number of experts, avoiding the

Table 4: Ablation study comparing input modalities (Text-Only, Image-Only, Joint Multimodal) and fusion architectures (Single Dense MLP vs. Proposed MoE Layer) evaluated by Macro F1 score across target cultural contexts.

Model Variant	Modalities	Fusion Mechanism	India F1	China F1	Western F1	Overall F1
Text-Only	Text Only (`mDeBERTa-v3`)	N/A	0.6898	0.7113	0.6941	0.6984
Image-Only	Image Only (`SigLIP-so400m`)	N/A	0.7413	0.7400	0.7176	0.7330
Dense Fusion	Text + Image	Single Dense MLP	0.7535	0.7322	0.7559	0.7472
MoE Fusion (Ours)	**Text + Image**	**MoE Layer ($N = 4$ Experts)**	**0.8120**	**0.8700**	**0.8653**	**0.8491**

instability or expert collapse often seen in sparsely-gated models on small datasets.

Error Analysis: While our model demonstrates strong overall performance, per-perspective metrics reveal specific cross-cultural failure modes. Interestingly, although our system performs competitively on the English subset, its absolute Macro-F1 score (0.782) is the lowest across all languages. This suggests the English subset (MAMI) poses a universally high difficulty, likely because the task requires predicting Indian and Chinese perspectives on Western-origin memes, a highly challenging cross-cultural semantic transfer. Furthermore, on typologically distinct datasets like Tamil, the model struggles more with the native Indian perspective (0.820) compared to the Western (0.851) and Chinese (0.857) lenses. One possible explanation is that when a meme's misogynistic cue relies on native, non-Western visual context or localized pop-culture tropes, the model leans towards interpretations learned from the English-dominated training data, resulting in false negatives for the native perspective. The pretraining distributions of our backbones may also contribute, as SigLIP's global visual representations may not adequately capture deeply localized regional contexts. A qualitative, example-level analysis is needed to confirm this pattern.

5.5 Limitations

Despite strong overall results, several limitations of our system are worth acknowledging. First, our MoE uses dense soft routing, meaning all N experts run on every forward pass (Eq. 2). While this avoids expert collapse on a small corpus, it gives up the computational savings of sparse top-k routing and does not enforce strict expert specialization; we have not yet analyzed routing distributions to verify whether experts specialize by culture or language. Second, the benchmark covers only three cultural perspectives (Indian, Chinese, Western), whereas meme misogyny is a global phenomenon. Communities in South-East Asia, Latin America, and sub-Saharan Africa are entirely absent, which limits how broadly the generalization claims can be applied. Third, meme transcriptions are truncated to 128 tokens; longer memes in which the sarcastic or misogynistic cue appears late in the text may be silently cut off, degrading the text signal. Finally, dividing the loss by the gradient-accumulation step count rather than by the number of active cultural perspectives per batch halves the effective learning signal for single-perspective samples and may under-weight rare cross-cultural labels during training. In addition, all reported results come from a single training seed, so run-to-run variance is not quantified; the Gemma3-4b baseline was taken from the challenge rather than reproduced under our setup; and the training mixture is dominated by English (MAMI) data ($\sim$74%), so low-resource performance may depend on this mixture, which we did not vary.

6 Conclusion

We presented MOEZHI-GURU, a multimodal Mixture of Experts system for cross-cultural misogynistic meme detection. By combining mDeBERTa-v3 and SigLIP with a specialized MoE routing layer and culture-specific classification heads, the model effectively captures the implicit cultural nuances that make misogynistic memes so difficult to detect. Results show improvements exceeding 0.42 Macro-F1 points over a Gemma3-4b baseline. Our model's cross-dataset consistency yields an average Macro-F1 of 0.8515 across all four language subsets combined, confirming that the MoE approach is a strong foundation for unified, globally robust multimodal moderation systems.

Ethics and Privacy Statement

This research addresses automated detection of harmful misogynistic content, with the broader aim of contributing to safer online environments. We do, however, acknowledge the potential for dual use: automated moderation tools can mistakenly

flag benign cultural expressions when training data encodes cultural biases. Our use of cross-cultural datasets helps mitigate monolithic cultural bias, but further evaluation is needed to ensure fairness across all marginalized sub-groups. No personally identifiable information was used beyond the publicly released datasets.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Ping Du, Xiaojian Zhuang, Saranya Rajiakodi, Paul Buitelaar, Premjith B, Bhuvaneswari Sivagnanam, Anshid Kizhakkeparambil, and Lavanya S.K. 2025a. 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, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025b. Findings of the shared task on misogyny meme detection: Dravidianlangtech@ naacl 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 721–731.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, Charmathi Rajkumar, et al. 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.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research (JMLR)*, 23(120):1–43.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Subalalitha Cn, Rahul Ponnusamy, Anshid Kizhakkeparambil, Bhuvaneswari Sivagnanam, Bharathi Raja Chakravarthi, et al. 2026. Overview of the shared task on multilevel political meme classification in tamil and malayalam. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 101–113.

Charmathi Rajkumar, Malliga Subramanian, and Bharathi Raja Chakravarthi. 2026. Multimodal transformer framework for multilingual harmful meme classification. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 99–107.

MD Musa Kalimullah Ratul, Sabik Aftahee, Tofayel Ahmmed Babu, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. Cuet_networksociety@ dravidianlangtech 2025: A multimodal framework to detect misogyny meme in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 529–535.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff

Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*.

Malliga Subramanian, Sathishkumar Veerappampalayam Easwaramoorthy, and Nandhini Subbarayan. 2025. Empowering sentiment analysis in social media: a comprehensive approach to enhance the classification of abusive Tamil comments using transformer models. *Journal of Big Data*, 12:208.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11941–11952.