CUET_12033@LT-EDI-2025: Misogyny Detection

Mehreen Rahman, Faozia Fariha, Nabilah Tabassum, Samia Rahman, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004033, u2004012, u2004020, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Misogynistic memes spread harmful stereotypes and toxic content across social media platforms, often combining sarcastic text and offensive visuals that make them difficult to detect using traditional methods. Our research has been part of the the Shared Task on Misogyny Meme Detection - LT- EDI@LDK 2025, identifying misogynistic memes using deep learning-based multimodal approach that leverages both textual and visual information for accurate classification of such memes. We experiment with various models including CharBERT, BiLSTM, and CLIP for text and image encoding, and explore fusion strategies like early and gated fusion. Our best-performing model, CharBERT + BiLSTM + CLIP with gated fusion, achieves strong results, showing the effectiveness of combining features from both modalities. To address challenges like language mixing and class imbalance, we apply preprocessing techniques (e.g., Romanizing Chinese text) and data augmentation (e.g., image transformations, text back-translation). The results demonstrate significant improvements over unimodal baselines, highlighting the value of multimodal learning in detecting subtle and harmful content online.

1 Introduction

Memes are a popular way to share jokes, opinions, and emotions online. However, they can also spread harmful ideas like misogynyhatred or dislike toward women (Pacilli and Mannarini, 2019; Chakravarthi, 2020; Paciello et al., 2021; Priyadharshini et al., 2022). Unlike plain text, memes combine images and captions, making it harder to detect if they are offensive. A meme might look funny at first, but it can be offensive when the image and text are viewed together (Chakravarthi et al., 2022a; Chakravarthi, 2022; Chakravarthi et al., 2022b).

Misogynistic memes are harder to detect than regular hate speech because they mix images and text and often rely on subtle cultural context that simple methods can miss. These challenges were the focus of the Misogyny Meme Detection Shared Task, part of the LT-EDI@LDK 2025 workshop, in which we participated (Chakravarthi et al., 2025). To better detect misogynistic memes, we apply three main strategies:

1. **Multimodal Feature Fusion:** We use CharBERT (Helboukkouri, 2020) and BiLSTM to process the text, and CLIP to extract image features. These are combined using gated fusion, helping the system detect subtle harmful content.

2. **Handling Language Variability:** Memes often include both English and Chinese. We convert Chinese to its Romanized form and use a Chinese CharBERT model to better handle mixed and informal language.

3. **Robust Training Strategy with Optimization Techniques:** We use AMP, gradient clipping, dropout, learning rate scheduling, and early stopping to improve convergence and model stability across training stages.

Detailed implementation information is available in the linked GitHub repository below- https://github.com/bountyhunter12/Misogyny-Meme-Detection.

2 Related Work

Existing works of misogyny meme detection can be categorized into text-based, image-based, and multimodal approaches, using ML, DL, or transformer-based methods.

In the past years, text-focused misogyny detection has been done in Pamungkas et al., 2020

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 127–132

using statistical classification models, including variations of Support Vector Machines (SVM) (Pamungkas et al., 2020). Recently , multimodal learning has gained considerable attention in this field. Starting with general harmful content detection (Kiela et al., 2021), specific detection of misogynistic meme gained popularity later. Cuervo and Parde, 2022 utilized (CLIP)-like architectures, though their approach faced challenges due to OCR noise, dataset bias and unfair influence from certain words (e.g., "woman"), leading to false positives.

Recent advancements have explored more robust architectures for misogynistic meme detection. In Chinivar et al., 2024 ,V-LTCS (Vision-Language Transformer Combination Search) is one such framework that systematically combines various state-of-the-art vision (Swin, ConvNeXt, ViT) and language (CharBERT, ALCharBERT, XLM-R) transformer models to find the best-performing multimodal pairs. Among all these English contents, Mallik et al., 2025 shifts the focus to Tamil language. They have used mCharBERT and IndicCharBERT for text data, and ViT, ResNet, and EfficientNet for image data. For multimodal detection, these are combined using concatenation. While research on English datasets has progressed, Chinese harmful meme detection still needs further attention. The study Lu et al., 2024 addresses this by constructing the TOXICN-MM dataset with fine-grained harmful type annotations. It proposes Multimodal Knowledge Enhancement (MKE), a baseline detector that integrates LLM-generated context to improve classification.

3 Data

We have utilized the dataset provided under the Shared Task on Misogyny Meme Detection - LT-EDI@LDK 2025 (Ponnusamy et al., 2024; Chakravarthi et al., 2024). The dataset has been segmented into training, development, and test sets containing 1,190, 170, and 340 samples, respectively. It primarily consists of code-mixed Chinese-English memes, the type of language commonly observed in online communication. The dataset consists of a significantly lower number of misogynistic memes compared to non-misogynistic ones, as shown in Table 1.

Table 1: Dataset Distribution for Misogyny Meme Detection.

Sets	Misogyny	Non-misogyny	Total
Train	349	841	1190
Dev	47	123	170
Test	104	236	340
Total	**500**	**1200**	**1,700**

4 Methodolgy

4.1 Data Preprocessing

As this is a multimodal task, we have preprocessed both image and text. For texts, URLs, emojis, punctuation, and numbers have been removed. Traditional Chinese has been converted to simplified Chinese for better consistency. The cleaned text is tokenized using the jieba tokenizer[1]. The filtered tokens were transliterated to Romanized Chinese using the pypinyin library[2]. Images were converted to RGB and resized to 224x224 pixels for consistency in dimension. Contrast and brightness have been enhanced to improve visual quality.

4.2 Data Augmentation

For better accuracy, we have reduced the class imbalance by applying augmentation specifically on the Misogyny class. Image augmentation has been implemented using **torchvision** library[3] library by, applying techniques such as brightness adjustment, grayscale conversion, and posterization. Text augmentation has been applied using **deep-translator** library[4], followed by back-translation through intermediate languages (such as French, German, and Spanish) to produce paraphrased versions while maintaining the original semantic context. This augmentation generates a balanced class of Misogynous and Not-Misogynous dataset, as shown in Table 2.

Table 2: Dataset Distribution Before and After Augmentation

Class	Before	After
Misogyny	349	841
Non-misogyny	841	841
Total	**1190**	**1682**

[1] https://pypi.org/project/jieba/
[2] https://pypi.org/project/pypinyin/
[3] https://pytorch.org/vision/
[4] https://github.com/nidhaloff/deep-translator

4.3 Overview of the Adopted Model

4.3.1 Unimodal Models

For the unimodal text classification task, we fine-tuned CharBERT-base-Chinese and CharBERT, leveraging their strong contextual understanding of the Chinese language. In the enhanced setup, the CharBERT outputs were further passed through a 2-layer BiLSTM module with a hidden size of 128, resulting in 256-dimensional embeddings that effectively capture sequential dependencies. Text sequences were tokenized with a maximum length of 128, and training was conducted using the Adam optimizer for 20 epochs, with a learning rate of 1×10^{-4} and a batch size of 16.

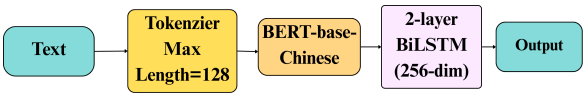

Figure 1: Unimodal Architecture for Text Classification using CharBERT-base-Chinese, followed by a 2-layer BiLSTM.

For the image-only models, we experimented with CLIP (visual encoder), Vision Transformer (ViT), ResNet-50, and EfficientNet-B0. All input images were resized to 224×224, normalized using standard ImageNet statistics, and converted to tensors. The image encoders extracted feature vectors of varying dimensions: 512-dimensional for CLIP, 768-dimensional for ViT, 2048-dimensional for ResNet-50, and 1280-dimensional for EfficientNet-B0. A final classification head was appended to map these features to the binary misogyny detection task. These models were also trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 16.

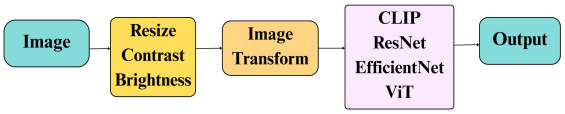

Figure 2: Unimodal Architecture for Image Data Processing and Classification.

4.3.2 Multimodal Models

Building on the unimodal baselines, we developed several multimodal architectures that integrate both text and image modalities. These include CharBERT combined with CLIP, CharBERT with ViT, CharBERT with ResNet-50 using a gated multimodal unit (GMU)-style fusion, and CharBERT with EfficientNet-B0 using concatenation followed by a multilayer perceptron (MLP). These architectures are designed to learn fine-grained cross-modal interactions between visual content and textual cues

The best-performing model consisted of CharBERT-base-Chinese combined with a 2-layer BiLSTM and ViT as the image encoder. The 256-dimensional text embeddings and 768-dimensional image features were fused using a GMU-style fusion layer. A fully connected layer followed by a softmax activation function produced the final prediction.

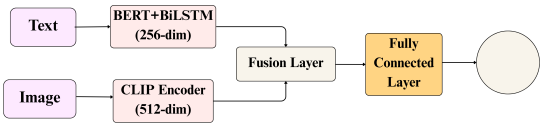

Figure 3: Unimodal Architecture for Image Data Processing and Classification.

In the CharBERT and ResNet-50 configuration, 768-dimensional textual features were fused with 2048-dimensional image embeddings using the same gated mechanism. Similarly, in the CharBERT and CLIP architecture, 256-dimensional BiLSTM outputs were combined with 512-dimensional CLIP features. In the CharBERT and EfficientNet-B0 variant, textual and visual features were concatenated and passed through an MLP for late fusion.

All multimodal models were trained for 20 epochs with a batch size of 16 and a learning rate of 1×10^{-4}. To address label imbalance, we employed class-weighted cross-entropy loss. The training process incorporated automatic mixed precision (AMP), learning rate scheduling, dropout, early stopping, and gradient clipping to ensure stable and efficient convergence.

5 Results and Analysis

This section presents the outcomes of our misogyny meme classification task, comparing unimodal and multimodal approaches to assess their effectiveness in detecting image-text-based hate content. Performance is evaluated using weighted precision (P), recall (R), and F1-score (F1), with Macro-F1 serving as the primary metric for classification efficacy.

5.1 Comparative Analysis

Among the unimodal text classifiers, CharBERT + BiLSTM performed better than other text-based approaches with a higher F1-score of 0.81. For unimodal image classifiers, ViT achieved a better score than CLIP (0.65 vs 0.42). When we combined CharBERT + BiLSTM with CLIP in a multimodal setup using gated fusion, the model achieved the highest F1-score of 0.82 with a precision of 0.81 and recall of 0.84. This shows the strength of combining textual and visual modalities for improved hate speech detection.

We have also evaluated the performances of ViT + CharBERT + BiLSTM (gated fusion) and CharBERT + BiLSTM + CLIP (early fusion) among other multimodal combinations. However, our analysis focuses primarily on the best performing approaches, with CharBERT + BiLSTM (unimodal text) and CharBERT + BiLSTM + CLIP (gated fusion) emerging as the top performers in their respective categories.

Classifier	P	R	F1
Unimodal (Text)			
CharBERT + BiLSTM	0.80	0.82	0.81
BERT+BiLSTM	0.77	0.75	0.76
Unimodal (Image)			
ViT	0.71	0.63	0.65
CLIP (Best Epoch)	0.36	0.50	0.42
Multimodal			
CharBERT + BiLSTM + CLIP (Gated Fusion)	0.81	0.84	0.82
ViT + CharBERT + BiLSTM (Gated Fusion)	0.71	0.75	0.71
CharBERT + BiLSTM + CLIP (Early Fusion)	0.71	0.77	0.70
BERT + BiLSTM + CLIP (Early Fusion)	0.69	0.68	0.68

Table 3: Performance of unimodal and multimodal systems on the test dataset.

5.2 Error Analysis

To better understand the performance and limitations of our multimodal model, we analyze the confusion matrix of the CharBERT + BiLSTM + CLIP (Early Fusion) model (Figure 4). Table 4 summarizes the classification outcomes.

	Predicted Not-Misogyny	Predicted Misogyny
Actual Not-Misogyny	161 (True Negative)	75 (False Positive)
Actual Misogyny	19 (False Negative)	85 (True Positive)

Table 4: Confusion matrix results for the BERT + BiLSTM + CLIP (Early Fusion) model on the test set.

The model successfully identified 161 non-misogynistic and 85 misogynistic instances. However, 75 non-misogynistic samples were misclassified as misogynistic (false positives), and 19 misogynistic samples were misclassified as non-misogynistic (false negatives). This indicates a moderate imbalance in the model's error pattern, with a higher false positive rate.

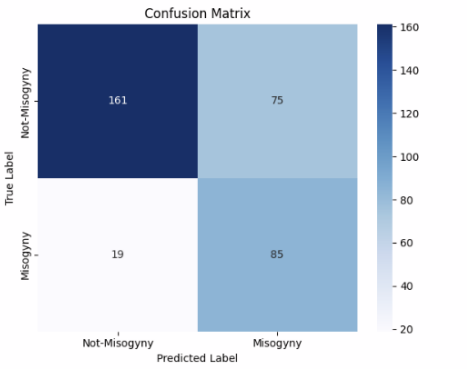

Figure 4: Confusion Matrix of the Multimodal CharBERT-BiLSTM-CLIP (Early Fusion) Model.

Quantitative Analysis. The confusion matrix suggests a relatively balanced misclassifications pattern, though the number of false positives is slightly higher. These errors may arise due to the model focusing on surface-level cues (e.g., certain words or facial expressions) rather than fully understanding context or intent. Future work could involve attention analysis or feature attribution methods to better interpret these decisions and mitigate such biases.

Qualitative Examples. To further investigate the model's decision patterns, we sampled representative examples from each confusion matrix category (Table 5).

Category	Example (Image)
True Positive (TP)	1342.jpg
True Negative (TN)	1582.jpg
False Positive (FP)	788.jpg
False Negative (FN)	1203.jpg

Table 5: Example predictions illustrating each category of the confusion matrix.

Discussion. The false positive example (788.jpg) shows that the model may misinterpret general aggression as misogyny, likely due to reliance on shallow linguistic or visual cues. The false negative example (1203.jpg) suggests a limitation in recognizing subtle or contextually

implied gender bias. This reveals that while the model performs reasonably well on overtly misogynistic content, it struggles with nuanced language and implicit bias.

Future Work. To reduce such errors, future research should explore:

- Incorporating multimodal attention maps or interpretability tools (e.g., LIME, SHAP).
- Enhancing the models semantic reasoning using external knowledge or stereotype databases.
- Fine-tuning with curated, context-rich, adversarial examples for better generalization.

6 Conclusion

In this study, we explored different ways to detect misogynistic memes using text, images, and a combination of both. By experimenting with various deep learning models like CharBERT for text and CLIP or ViT for images, we found that combining both text and image features gives the best results. In particular, the model that used CharBERT + BiLSTM + CLIP with early fusion stood out by accurately detecting harmful content, achieving a Macro F1 score of 0.70. We also used data augmentation techniques to handle the imbalance in the dataset, and tried to reduce errors where models misclassified subtle or sarcastic memes. Despite these improvements, challenges still remain, particularly with overlapping categories and class imbalance. Future work will focus on better data labeling and smarter model strategies to improve fairness and accuracy.

Limitations

Although our approach shows strong performance in detecting misogynistic memes, it comes with certain limitations. First, the dataset used in our experiments is relatively imbalanced, which may affect the model's ability to generalize to unseen, real-world data. Additionally, memes often carry nuanced cultural or sarcastic references that are difficult for automated systems to interpret correctly. The use of CLIP and other image encoders also means that only static visual features are captured, potentially overlooking deeper symbolic or evolving visual cues. Lastly, although data augmentation helps improve class balance, it may introduce artificial patterns that do not fully reflect the complexity of naturally occurring misogynistic content, which could impact the models real-world robustness.

Ethics Statement

we have been committed to maintaining the ethical practices during our work to build a system that helps detect harmful and offensive memes. Our goal is to support safer online spaces by reducing toxic content, while making sure our methods are fair and respectful to all.

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 Fifth Workshop on Language Technology for Equality, Diversity and Inclusion*, Italy. Fifth Conference on Language, Data and Knowledge (LDK2025).

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.

B.R. Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

B.R. Chakravarthi. 2022. Hope speech detection in youtube comments. *Social Network Analysis and Mining*, 12(1):75.

B.R. Chakravarthi, R. Ponnusamy, and R. Priyadharshini. 2022a. Multilingual hope speech detection in english and dravidian languages. *International Journal of Data Science and Analysis*, 14(4):389–406.

B.R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, et al. 2022b. Overview of the shared task on homophobia and transphobia detection in social media comments. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity, and Inclusion*, pages 369–377.

Sneha Chinivar, Roopa M.S., Arunalatha J.S., and Venugopal K.R. 2024. V-ltcs: Backbone exploration for multimodal misogynous meme detection. *Natural Language Processing Journal*, 9:100109.

Charic Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. pages 785–792.

Houssam Helboukkouri. 2020. Characterbert: A pretrained language model using character-level inputs. `https://huggingface.co/helboukkouri/character-bert`. Accessed: 2025-05-13.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. *Preprint*, arXiv:2410.02378.

Arpita Mallik, Ratnajit Dhar, Udoy Das, Momtazul Arefin Labib, Samia Rahman, and Hasan Murad. 2025. CUET-823@DravidianLangTech 2025: Shared task on multimodal misogyny meme detection in Tamil language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 325–329, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

M. Paciello, F. DErrico, G. Saleri, and E. Lamponi. 2021. Online sexist meme and its effects on moral and emotional processes in social media. *Computers in Human Behavior*, 116:106655.

M.G. Pacilli and T. Mannarini. 2019. Are women welcome on facebook? a study of facebook profiles of italian female and male public figures. *TPM: Test. Psychom. Methodol. Appl. Psychol.*, 26(2).

Endang Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management*, 57:102360.

Rahul Ponnusamy, Kathiravan Pannerselvam, Saranya R, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, Bhuvaneswari S, 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.

R. Priyadharshini, R. Ponnusamy, B.R. Chakravarthi, et al. 2022. Misogyny speech detection using long short-term memory and bert embeddings. *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, pages 155–159.