CUET_320@LT-EDI-2025: A Multimodal Approach for Misogyny Meme Detection in Chinese Social Media

Madiha Ahmed Chowdhury, Lamia Tasnim Khan, MD. SHAFIQUL HASAN, Ashim Dey
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004052, u2004048, u1904083}@student.cuet.ac.bd, ashim@cuet.ac.bd

Abstract

Detecting misogyny in memes is challenging due to their complex interplay of images and text that often disguise offensive content. Current AI models struggle with these cross-modal relationships and contain inherent biases. We tested multiple approaches for the Misogyny Meme Detection task at LT-EDI@LDK 2025: ChineseBERT, mBERT, and XLM-R for text; DenseNet, ResNet, and InceptionV3 for images. Our best-performing system fused fine-tuned ChineseBERT and DenseNet features, concatenating them before final classification through a fully connected network. This multimodal approach achieved a 0.93035 macro F1-score, winning **1st** place in the competition and demonstrating the effectiveness of our strategy for analyzing the subtle ways misogyny manifests in visual-textual content.

1 Introduction

The continuous rise of social media platforms has reshaped how people communicate and share content online. However, this digital transformation has also led to the proliferation of harmful content, including gender bias and sexism, often expressed through memes. Memes—though academically debated—have become deeply embedded in everyday digital interactions. Their combination of visual elements and textual content creates significant challenges for content moderation due to their multi-modal nature and the cultural context they carry (Chakravarthi et al., 2022a; Chakravarthi, 2022; Chakravarthi et al., 2022b). The *Shared Task on Misogyny Meme Detection in Chinese Social Media* at *DravidianLangTech@LDK 2025* aims to address this issue by identifying misogynistic content within Chinese memes. Chinese memes are unique in that they often include complex characters, cultural references, and idioms, necessitating specialized approaches for accurate detection. Given the limited research on multi-modal misogyny detection in Chinese social media, this task fills a significant gap in the field (Chakravarthi, 2020).

Our contribution to this shared task includes:

- Developing a novel multi-modal framework that effectively integrates visual and textual features from Chinese memes.
- Implementing specialized pre-processing techniques tailored for Chinese text and meme images to capture both linguistic and visual nuances.
- Demonstrating state-of-the-art performance using a fusion of `Chinese-BERT` and `Vision Transformer` models (Helboukkouri, 2020).
- Analyzing misogynistic patterns in Chinese memes to better understand the cultural nuances of online misogyny and how they manifest in this language.

Our approach achieved top performance in the task, advancing the field of multi-modal content moderation for Chinese social media and contributing to efforts aimed at creating safer online spaces. Our implementation details are available online[1].

2 Related Work

Detecting misogynistic content in memes presents unique challenges due to their multimodal nature. Research has approached this through text, images, or combined modalities.

Researchers in Pamungkas et al., 2020 have focused on text-based misogyny detection using traditional machine learning, primarily employing SVMs to identify hateful language. Other text-based approaches analyzed hate comments

[1] https://github.com/lamiatasnimkhan/CUET-320-Multimodal-Misogyny-Meme-Detection

to better understand linguistic patterns in misogynistic content (Tofa et al., 2025). However, text-only methods show limitations when targeting specific groups, highlighting the need for specialized misogyny detection datasets rather than generic hate speech systems. Similarly, vision-based research using CNNs and Transformers has demonstrated promise, but visual cues alone often prove insufficient. Integrating textual and visual features has yielded superior results. Systems combining Naive Bayes classifiers with visual processing have improved detection in low-resource languages, while advanced fusion techniques with Perceiver IO, RoBERTa, and Vision Transformers effectively address both binary and multi-label tasks (Pramanick et al., 2021). Other approaches use pre-trained CLIP models to bridge the semantic gap between modalities (Aho and Ullman, 1972). Recent work has extended multimodal misogyny detection to low-resource languages like Tamil and Tulu, emphasizing the need for culturally aware systems (Mallik et al., 2025).

3 Task and Dataset Description

We used the dataset from the Shared Task on Misogyny Meme Detection - LT-EDI@LDK 2025 (Ponnusamy et al., 2024; Chakravarthi et al., 2025), which includes 1,190 training, 170 development, and 340 test samples. The data comprises code-mixed Chinese-English memes, with a notable imbalance—fewer misogynistic samples than non-misogynistic ones (as shown in Table 1 and Figure 1). The dataset contains social media-style memes with real templates, combining sarcastic, code-mixed, and abusive captions with reaction images or symbolic visuals.

Set	Misogyny	Non-misogyny	Total
Train	349	841	1,190
Dev	47	123	170
Test	104	236	340

Table 1: Class distribution across dataset splits

Each transcription was annotated with a binary label:

- **Misogynistic:** Content that conveys hate, harassment, or derogatory views targeted at women.
- **Not Misogynistic:** Content that lacks misogynistic features or targets.

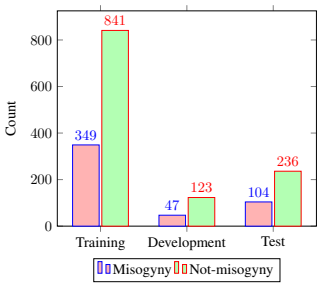

Figure 1: Dataset Distribution for Misogyny and Not-Misogyny

4 Methodology

The schematic representation of our approach is depicted in Figure 2.

4.1 Data Preprocessing

Meme samples underwent parallel preprocessing for both modalities. Text processing included regex-based URL and non-linguistic character removal, Jieba segmentation, filtering of 25 common Chinese stopwords, and truncation to 512 tokens for transformer compatibility. Images were validated, converted to RGB, resized to 224×224 pixels, and enhanced via histogram adjustment ($\alpha = 1.2$, $\beta = 20$). We implemented a dual loading strategy with OpenCV and PIL fallback, and normalized using ImageNet statistics.

4.2 Data Augmentation

Class-balanced augmentation is used exclusively on misogynistic samples (label=1) through three transformations: random brightness adjustment (factor=0.5), probabilistic grayscale conversion (p=0.2), and 4-bit color posterization. Each positive sample generated two augmented variants, effectively tripling the minority class representation. Text data remained unaugmented due to the semantic sensitivity of Chinese language transformations.

4.3 Feature Extraction

We employ separate pipelines for text and image feature extraction.

4.3.1 Text Modality

We employ three multilingual pretrained language models for text feature extraction. ChineseBERT (Cui et al., 2021) incorporates glyph and phonetic information for enhanced Chinese language processing. XLM-R (Conneau and Lample, 2019) utilizes the RoBERTa objective across 100 languages. The multilingual BERT baseline mBERT

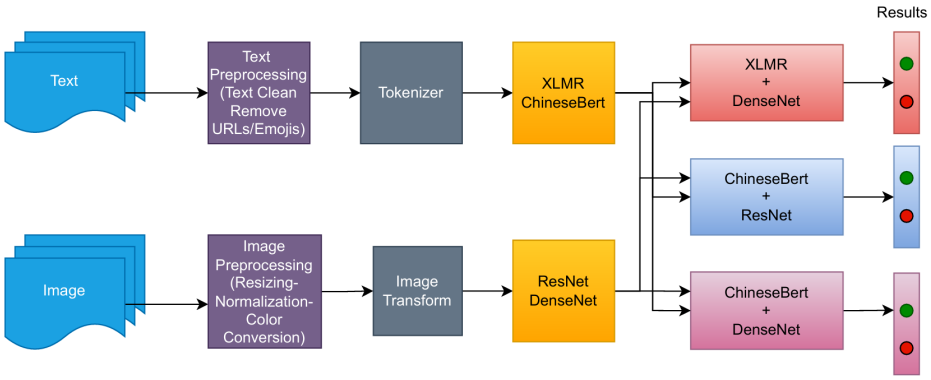

Figure 2: Abstract View of Methodology

(Devlin et al., 2019) covers 104 languages through masked language modeling. All models generate 768-dimensional embeddings via mean pooling of token outputs.

4.3.2 Image Modality

For visual feature extraction, we employ three CNN architectures: ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and Inception-V3 (Szegedy et al., 2016). Each model's classification head is removed to extract 2048-dimensional features from ResNet-50 and Inception-V3, and 1024-dimensional features from DenseNet-121.

4.4 Multimodal Fusion

We combined our best unimodal models—ChineseBERT for text and DenseNet-121 for images—for multimodal fusion. Each modality was independently processed, with text encoded through mean-pooling of transformer token embeddings and visual data via flattened CNN output. These representations were then concatenated into a unified multimodal embedding. The fused embedding was processed through a fully connected layer with softmax activation. Rather than complex attention mechanisms, our simpler embedding concatenation approach reduced model complexity while maintaining strong performance, suggesting basic fusion suffices for this task. Table 2 illustrates the hyperparameters used for the models. All models used Adam optimizer.

5 Result and Analysis

Our experimental results demonstrate that multimodal fusion of ChineseBERT and DenseNet-121 achieves superior performance ($F1 = 0.93$) for misogyny meme detection, outperforming unimodal approaches where text-based ChineseBERT ($F1 = 0.91$) surpassed image-only models. The 2.2% improvement from multimodal integration confirms the complementary value of combining linguistic and visual features, with DenseNet-121 ($F1 = 0.82$) proving more effective than other CNN architectures for visual analysis, while maintaining balanced precision-recall ratios across all models. We evaluated the performance of our models using macro-averaged precision, recall and F1 score (Macro-F1). Among these, the Macro-F1 score serves as the primary metric for assessing the overall effectiveness of the systems.

Modality	LR	WD	BS	EP
Text	2×10^{-5}	0.01	16	10
Image	1×10^{-4}	0.00	16	20
Bimodal	1×10^{-4}	0.00	16	30

Table 2: Training hyperparameters for all models. LR: Learning Rate, WD: Weight Decay, BS: Batch Size, EP: Epochs.

5.1 Quantitative Analysis

We evaluated several models using macro F1-score (MF1) to identify the best approach for misogyny detection. Detailed result is presented in Table 3. Unimodal models served as baselines, while the multimodal DenseNet + Chinese BERT model achieved the highest MF1 of 0.93 by effectively leveraging both image and text features. As shown in Figure 3, this model correctly classified 227 non-misogynistic and 93 misogynistic samples, with 9 false positives and 11 false negatives. These errors likely stem from class imbalance and limited data diversity. Multimodal models are effective when misogyny arises from interactions between

text and images (e.g., sarcasm). Still, their performance may decline if one modality dominates or if fusion introduces noise. Image-only models may outperform due to noise in visual features, whereas text modalities often provide stronger discriminative signals; suboptimal fusion can degrade these signals. Following prior work, we trained unimodal models using joint labels. We note this may introduce label noise, as one modality may lack full context, and suggest future work explore modality-specific labels or weak supervision.

Type	Model	P	R	F1
Text	**ChineseBERT**	0.92	0.90	**0.91**
	XLM-R	0.91	0.88	0.89
	mBERT	0.89	0.87	0.88
Image	ResNet-50	0.79	0.75	0.76
	DenseNet-121	0.81	0.83	**0.82**
	InceptionV3	0.79	0.81	0.80
Multi-modal	ChineseBERT+ ResNet	0.84	0.78	0.80
	ChineseBERT+ DenseNet	0.94	0.92	**0.93**
	XLM-R+ DenseNet	0.85	0.77	0.79

Table 3: Macro-averaged classification scores: Precision (P), Recall (R), and F1 across model types.

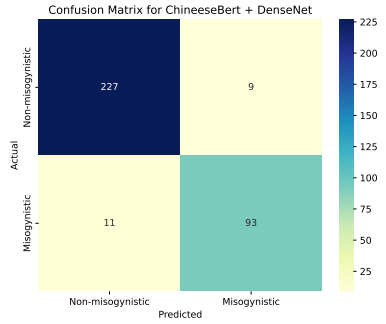

Figure 3: Confusion Matrix for Multimodal DenseNet + Chinese BERT Model.

5.2 Qualitative Analysis

Figure 4 highlights both correctly classified and misclassified cases. Among the misclassifications:

False Positives: Sample 882.jpg ("Friday Friday Reborn") was incorrectly flagged as misogynistic despite containing no gendered language, suggesting model overfitting on stylistic cues without context.

False Negatives: Sample 1562.jpg ("A handsome and smart baby boy invites you to take him home. Refuse to take him home.") went undetected despite containing objectifying language and gender stereotypes.

Image_ name	Result	Prediction	Transcriptions
1342.jpg	Misogyny	Misogyny	“你这么懒以后一定会被婆婆骂” “女生不用读太多书，反正最后都要嫁人的” “25岁还没对象我都替你丢人” “35岁之后就是嫁不出去的老姑娘了” ("You will definitely be scolded by your mother-in-law for being so lazy" "Girls don't need to study too much, they will get married in the end anyway" "I feel ashamed for you not having a partner at the age of 25" "After the age of 35, you will be an old maid who can't get married)
1582.jpg	Not-Misogyny	Not-Misogyny	光顾着上学 忘记上吊了(I was so busy with school that I forgot to hang myself.)
882.jpg	Not-Misogyny	Misogyny	周五周五 脱胎换骨 (Friday Friday Reborn)
1562.jpg	Misogyny	Not-Misogyny	帅气聪明的男宝宝 邀请你接他回家 拒绝 接他(A handsome and smart baby boy invites you to take him home. Refuse to take him home.)

Figure 4: Examples of the DenseNet + Chinese BERT model’s anticipated outputs with English translations.

These errors reveal the model’s difficulty with indirect misogyny, especially in sarcastic, metaphorical, or superficially neutral language. While it handles explicit discrimination well, detecting implicit bias and understanding cultural context requires further improvement.

6 Conclusion

In this study, we started by experimenting with a few unimodal models, focusing separately on text and image data. While these gave us a decent starting point, it was the multimodal models, which combine both visual and textual features, that really stood out. In particular, our DenseNet + Chinese BERT model achieved the best results, reaching an F1 score of 0.93. Despite working with a relatively limited dataset, these findings show that combining modalities is crucial for capturing the complex and often subtle nature of misogynistic content in memes. For future work, we plan to expand the dataset, explore better data augmentation, and fine-tune our multimodal fusion techniques to push performance even further.

Limitations

While our results are promising, model performance is limited by several constraints. The dataset size was relatively small, especially for detecting subtle or implicit misogyny. Despite data augmentation, exposure to diverse examples remains limited. Pretrained language models like Chinese-BERT and XLM-R may overlook nuances in slang, dialects, or sarcasm. Multimodal pairs—e.g., Chinese-BERT with ResNet or DenseNet—struggled with interpreting irony or misalignment between text and image, which is common in memes. Additionally, our current models do not support audio, leaving out video-based content for future work.

Ethics Statement

Our team conducted this research with a deep commitment to ethical standards. As researchers and internet users ourselves, we recognize the real harm that online misogyny causes to women and marginalized communities. By developing better methods to identify harmful content in Chinese memes, we hope our work contributes to creating digital spaces where everyone feels welcome and respected.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

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).

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, and 1 others. 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.

Alexis Conneau and Guillaume Lample. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Zheng Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *CVPR*, pages 4700–4708.

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.

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.

Soumick Pramanick, Dimitar Sharma, Dimitar Dimitrov, Prasenjit Mukherjee, Marcos Minovski, Marta Villegas Enrich, Miguel Angel García Carmona, Manuel Núñez-García, Javier Casas, Antti Ilari Vatanen, and Preslav Nakov. 2021. Detecting misogyny and xenophobia in Spanish tweets using multilingual contextual embeddings and multimodal information. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, pages 278–289. CEUR Workshop Proceedings.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826.

Farjana Alam Tofa, Khadiza Sultana Sayma, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech 2025: Abusive comment detection in Malayalam text targeting women on social media using transformer-based models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 483–488, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.