CUET's_White_Walkers@LT-EDI-2025: A Multimodal Framework for the Detection of Misogynistic Memes in Chinese Online Content

Md Mubasshir Naib[a], Md Mizanur Rahman[b], Jidan Al Abrar[c]
Md Mehedi Hasan[d], Md Siddikul Imam Kawser [e], Mohammad Shamsul Arefin [f]
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904089[a], u1904116[b], u1904080[c],u1904067[d],u1904081[e]}@student.cuet.ac.bd,
sarefin@cuet.ac.bd[f]

Abstract

Memes, combining visual and textual elements, have emerged as a prominent medium for both expression and the spread of harmful ideologies, including misogyny. To address this issue in Chinese online content, we present a multimodal framework for misogyny meme detection as part of the LT-EDI@LDK 2025 Shared Task. Our study investigates a range of machine learning (ML) methods such as Logistic Regression, Support Vector Ma- chines, and Random Forests, as well as deep learning (DL) architectures including CNNs and hybrid models like BiLSTM-CNN and CNN-GRU for extracting textual features. On the trans- former side, we explored multiple pretrained models including mBERT, MuRIL, and BERT- base-chinese to capture nuanced language rep- resentations. These textual models were fused with visual features extracted from pretrained ResNet50 and DenseNet121 architectures using both early and decision-level fusion strategies. Among all evaluated configurations, the BERT-base-chinese + ResNet50 early fusion model achieved the best overall performance, with a macro F1-score of 0.8541, ranking 4th in the shared task. These findings underscore the effectiveness of combining pretrained vision and language models for tackling multimodal hate speech detection.

1 Introduction

In recent years, meme culture has become a dominant form of expression on Chinese social media platforms. Memes are often humorous or satirical, but like any medium, they are not immune to misuse (Das et al., 2022). Increasingly, this format is being exploited to disseminate discriminatory or hateful ideas—misogyny among the most concerning. Identifying such harmful content is essential in creating safer digital communities.

While the detection of offensive content in memes has seen growing attention (Mohiuddin et al., 2025), research that targets misogynistic content specifically, particularly in the Chinese context, is still in its infancy. Prior studies have largely focused on high-resource languages such as English, Hindi, and Arabic where comprehensive datasets and pretrained models are more readily available. By contrast, the Chinese language, despite its vast user base, remains comparatively underrepresented in this domain (Chowdhury et al., 2025).

To address this shortfall, we introduce a multimodal detection framework tailored to the nuances of Chinese meme content. Central to our approach is a newly curated dataset, annotated with misogynistic and non-misogynistic labels. The dataset reflects the linguistic and cultural diversity of Chinese online spaces, ensuring the model learns from contextually relevant examples.

Our methodology combines textual analysis, leveraging powerful transformer-based models to parse captions and embedded text. To complement textual analysis, we extract visual features using pretrained convolutional networks like ResNet50 and DenseNet121. These features are then integrated with text representations using early and decision-level fusion strategies. Our best-performing configuration, which combines BERT-base-chinese and ResNet50 via early fusion, demonstrates the effectiveness of this multimodal approach in capturing the complex and often subtle nature of misogynistic content in memes (Fersini et al., 2019).

Our main contributions are as follows:

- We develop and fine-tune multimodal models that integrate both textual and visual information.
- We evaluate multiple model configurations, offering insight into effective strategies for detecting nuanced hate speech in memes.

The implementation details have been pro-

vided in the following GitHub repository:-https://github.com/MubasshirNaib/Misogyny-Meme-Detection.

2 Related Work

The rise of misogynistic(Hossan et al., 2025) and harmful content(Naib et al., 2025; Sakib et al., 2025) in online memes has sparked growing concern and has become a significant area of research. As these memes typically combine both text and images, researchers have increasingly turned to multimodal learning techniques to improve detection capabilities. These techniques aim to process and interpret both the visual and linguistic components of a meme simultaneously—a task that becomes particularly complex in the Chinese context, given its rich cultural references, nuanced language, and diverse writing systems.

Although studies directly targeting misogynistic memes in Chinese are limited, various recent works offer solid ground for building suitable approaches. For example, (Jindal et al., 2024) introduced the MISTRA model, which merges text features with image embeddings using variational autoencoders (VAEs) to compress the image data effectively. This fusion allows the system to capture deeper semantic correlations between text and visuals.

Expanding on this idea, (Srivastava, 2022) developed MOMENTA, a deep neural network that looks at both broad and detailed features within memes. By analyzing overall structure alongside fine-grained details, the model is better equipped to spot nuanced forms of hate speech or misogyny.

Addressing the multilingual nature of memes, (Singh et al., 2024) compiled a large code-mixed dataset in Hindi and English, aimed specifically at identifying misogynistic content. They showed that multimodal approaches, especially those trained on code-switched language, are better suited for the mixed-language realities of social media—an insight that is also applicable to Chinese content, which often blends Mandarin with dialects, slang, or romanized expressions.

A notable contribution by (Pramanick et al., 2021) is the SCARE framework, which emphasizes strong alignment between textual and visual data. The model works by maximizing mutual information across the two modalities, making their shared representation more cohesive and informative. At the same time, it refines how each modality is represented on its own.

Meanwhile, (Habash et al., 2022) took a different approach by combining multiple models into an ensemble. This method benefits from the strengths of each individual model, helping to offset their weaknesses and improve overall detection rates of misogynistic content.

The importance of linguistic and cultural diversity in meme detection was also highlighted in the DravidianLangTech-2022 shared task (Das et al., 2022), where teams focused on memes in languages like Tamil and Malayalam. Their findings reinforced the value of fusing image and text data, especially in low-resource languages. Supporting studies by (Ghanghor et al., 2021) and (Chakravarthi et al., 2024) echoed these results, offering evidence that multilingual, multitask frameworks can effectively capture offensive and misogynistic content across different languages and contexts.Despite these strides, Chinese memes remain an underexplored territory. The combination of symbolic imagery, character-based writing, sarcasm, and internet-specific language poses unique challenges. For any framework designed to detect misogyny in Chinese memes, it's crucial to handle visual-linguistic humor, character-level nuance, and even cultural cues that may not be obvious without context.

3 Task and Dataset Description

This study addresses the task of misogyny meme detection in Chinese social media as part of the LT-EDI@LDK 2025 Shared Task(Chakravarthi et al., 2025). The given dataset (Ponnusamy et al., 2024) is multimodal, comprising image-text meme pairs labeled as Misogyny or Not-Misogyny. It is divided into training (1190 samples), validation (170 samples), and a test set (340 samples). The training set includes 841 non-misogynistic and 349 misogynistic examples, while the validation set includes 123 and 47 respectively. The textual data consists of 4,553 total words and 3,902 unique words, reflecting rich linguistic diversity. This setup enables the development of multimodal models that capture both visual cues and nuanced language patterns essential for detecting gender-based harmful content. Table 1 shows the class-wise distribution of samples for the Chinese dataset.

Class	Train	Validation	Test	W_T	UW_T
Not-Misogyny	841	123	236	2216	1996
Misogyny	349	47	104	1373	1139
Total	1190	170	340	4553	3902

Table 1: Class distribution across training, validation, and test splits, where W_T represents total words and UW_T represents total unique words.

4 Methodology

Several ML, DL, and transformer-based models were investigated to construct a robust framework for misogyny meme detection (Figure 1).The implementation details of the models have been open-sourced to ensure reproducibility[1].Appendix A presents the system requirements.

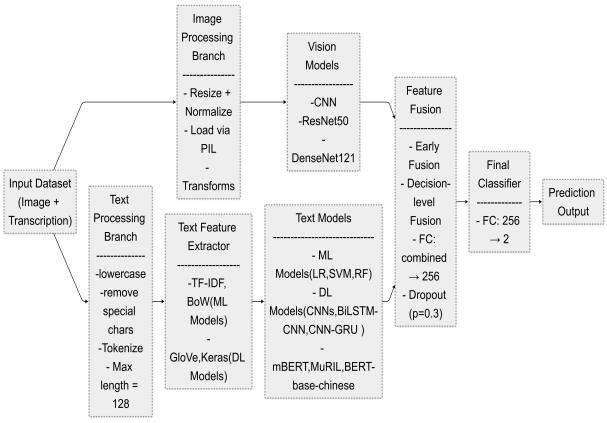

Figure 1: Schematic workflow for detecting misogynistic memes in Chinese social media content.

4.1 Data Preprocessing

The dataset comprises memes with both image and textual components. These are processed through two distinct branches to prepare them for feature extraction. In the image processing branch, all images are resized, normalized, and loaded using the Python Imaging Library (PIL). Further transformations, such as random cropping and flipping, are applied to enhance robustness and mitigate overfitting. On the textual side, transcriptions are first normalized by converting to lowercase and removing special characters. The cleaned text is then tokenized, with a maximum token length of 128, ensuring consistency in input dimensions for downstream models.

[1]https://colab.research.google.com/drive/1TYrg-vma1h46UwuhTdtX0eO2pCVPwJTn?usp=sharing

Hyperparameter	BERT-base-chinese + ResNet50 (Chinese)
Learning Rate	5e-5
Batch Size	8
Number of Epochs	5
Max Sequence Length	128
Optimizer	AdamW
Dropout Rate	0.3
Image Model	ResNet50
Text Model	BERT-base-chinese
Scheduler	ReduceLROnPlateau

Table 2: Tuned hyperparameters used in the best-performing multimodal model for Chinese misogyny meme detection.

4.2 Feature Extraction

Visual features are extracted using a range of pretrained convolutional neural networks, including generic CNNs, ResNet50, and DenseNet121. These models transform the image input into high-level visual embeddings. In parallel, textual features are extracted through two categories of approaches. Traditional machine learning pipelines utilize TF-IDF and Bag-of-Words representations. These features are suitable for models such as Logistic Regression, Support Vector Machines, and Random Forests. For deep learning models, word embeddings like GloVe and Keras embeddings are employed, supporting CNN-based architectures and hybrid models like BiLSTM-CNN and CNN-GRU. Additionally, transformer-based language models—mBERT, MuRIL, and BERT-base-chinese—are used to extract context-rich text representations. BERT-base-chinese, in particular, demonstrated superior performance in capturing linguistic nuances specific to Chinese discourse.

4.3 Baselines

To establish comparative performance, both unimodal and multimodal baselines were evaluated.

4.3.1 Unimodal Baselines

In the unimodal setup, the text and image modalities are processed independently. Text-based models include both machine learning classifiers using traditional features (TF-IDF, BoW) and deep learning architectures with embeddings. Similarly, image-based baselines rely on pretrained CNNs like ResNet50 and DenseNet121. Each modality is passed through its respective pipeline, and the final classification is performed separately to assess individual performance.

Approaches	Classifiers / Models	P	R	F1	G1
Textual Only	Logistic Regression	0.7721	0.6105	0.6453	0.6827
	SVM	0.7854	0.6282	0.6723	0.6985
	Random Forest	0.7668	0.5921	0.6302	0.6694
	CNN	0.7013	0.6450	0.6718	0.6729
	BiLSTM-CNN	0.7219	0.6523	0.6841	0.6855
	CNN-GRU	0.7352	0.6604	0.6907	0.6959
Visual Only	ResNet-50	0.7024	0.6201	0.6482	0.6595
	DenseNet-121	0.7480	0.6443	0.6827	0.6928
Multi-modal Fusion (Early Fusion)	mBERT + ResNet-50	0.8387	0.8034	0.8172	0.8209
	MuRIL + ResNet-50	0.8460	0.8127	0.8239	0.8292
	BERT-base-chinese + ResNet-50	**0.8812**	**0.8307**	**0.8541**	**0.8556**
Multi-modal Fusion (Late Fusion)	BERT-base-chinese + ResNet-50	0.8650	0.8201	0.8384	0.8421

Table 3: Comparison of various unimodal and multimodal models for misogyny meme detection in Chinese. EF: Early Fusion, LF: Late Fusion, P: Precision, R: Recall, F1: F1-score, G1: Geometric mean of P and R.

4.3.2 Multimodal Baselines

The multimodal approach fuses information from both text and image branches. Two primary fusion strategies are explored. Early fusion combines the intermediate feature representations from each modality and feeds them into a joint fully connected layer, followed by a dropout layer with a probability of 0.3 for regularization. In decision-level fusion, the output scores from unimodal classifiers are merged to generate the final prediction. Among all configurations, the early fusion model that integrates BERT-base-chinese and ResNet50 outperformed others, and the tuned hyperparameters are presented in Table 2.

5 Result Analysis

The performance evaluation of various unimodal and multimodal models on the Chinese misogyny meme detection task, which are presented in Table 3, reveals key insights into the effectiveness of different fusion strategies and model combinations. Among unimodal textual models, CNN-GRU and BiLSTM-CNN outperformed classical classifiers like Logistic Regression and SVM, demonstrating that sequential and convolutional architectures are better at capturing linguistic patterns. Visual-only models, particularly DenseNet-121, also performed reasonably well, though their standalone effectiveness remained slightly lower than that of text-based models.

Multimodal fusion approaches significantly outperformed unimodal methods. Early fusion models, especially BERT-base-chinese combined with ResNet-50, achieved the highest performance with an F1-score of 0.8541, indicating strong synergy between visual and textual features. Late fusion models also improved results but were slightly less effective than early fusion, emphasizing the value of integrating modalities early in the learning process. These findings affirm that combining vision and language models is crucial for accurately detecting misogynistic content in memes.Appendix B presents the error analysis.

6 Conclusion

In this study, we proposed a robust multimodal framework for misogyny meme detection in Chinese social media content, developed as part of the LT-EDI@LDK 2025 Shared Task. By combining pretrained transformer-based textual encoders such as BERT-base-chinese with visual feature extractors like ResNet-50 and DenseNet-121, our approach effectively captured the nuanced interplay between language and imagery. Among all tested configurations, the early fusion model of BERT-base-chinese and ResNet-50 achieved the best overall performance, demonstrating the strength of deep multimodal representation learning in tackling hate speech detection tasks. These results reinforce the importance of using culturally and linguistically aligned pretrained models for context-sensitive applications like misogyny detection.

7 Limitations

Despite promising results, our work has some limitations. First, the dataset was relatively small, particularly the test set, which may limit the generalizability of the findings. Second, the binary classification setting (misogyny vs. not-misogyny) does not capture the full spectrum or subtlety of harmful content. Third, although our fusion strategies improved performance, more sophisticated fusion

mechanisms such as attention-based or cross-modal transformers could further enhance the model's interpretability and accuracy. Finally, domain-specific biases in pretrained models and visual encoders may impact performance in culturally nuanced cases, calling for the development of more inclusive and fair AI systems.

Acknowledgments

The authors gratefully acknowledge Centro Interuniversitario di Ricerca Scienze Umane e Sociali e Intelligenza Artificiale (ELIZA) – University of Naples "L'Orientale" for its support in covering the registration costs, which enabled their participation.

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, Ramachandran Rajalakshmi, Hariharan LekshmiAmmal, Anshid Kizhakkeparambil, Susminu S Kumar, Bhuvaneswari Sivagnanam, Charmathi Rajkumar, and 1 others. 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.

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.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Hate-alert@ dravidianlangtech-acl2022: Ensembling multi-modalities for tamil trollmeme classification. *arXiv preprint arXiv:2204.12587*.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 222–229.

Mohammad Habash, Yahya Daqour, Malak Abdullah, and Mahmoud Al-Ayyoub. 2022. Ymai at semeval-2022 task 5: Detecting misogyny in memes using visualbert and mmbt multimodal pre-trained models. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 780–784.

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.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Md Mohiuddin, Md Minhazul Kabir, Kawsar Ahmed, and Mohammed Moshiul Hoque. 2025. Cuet-nlp_mp@ dravidianlangtech 2025: A transformer-based approach for bridging text and vision in misogyny meme detection in dravidian languages. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 514–521.

Md. Mubasshir Naib, Md. Saikat Hossain Shohag, Alamgir Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. cuetRaptors@DravidianLangTech 2025: Transformer-based approaches for detecting abusive Tamil text targeting women on social media. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 739–745, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Nazmus Sakib, Md. Refaj Hossan, Alamgir Hossain, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. CUET-NLP_Big_O@DravidianLangTech 2025: A BERT-based approach to detect fake news from Malayalam social media texts. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 440–447, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Harshvardhan Srivastava. 2022. Misogynistic meme detection using early fusion model with graph network. *arXiv preprint arXiv:2203.16781*.

A System Requirements

The entire framework is implemented using Python, leveraging libraries such as PyTorch and HuggingFace Transformers for deep learning, and Scikit-learn for traditional machine learning models. A GPU-enabled system with at least 16 GB of RAM is recommended to efficiently train and evaluate deep models, especially those involving transformer architectures and multimodal fusion.

B Error Analysis

We conducted both quantitative and qualitative error analyses to gain comprehensive insights into the performance of the proposed model.

B.1 Quantitative Analysis:

The confusion matrix B.1 reveals that while the model effectively identifies non-misogynistic content with a high specificity of 96.74% (119 true negatives vs. 4 false positives), it struggles to detect misogynistic content, as evidenced by a lower sensitivity of 59.57% (28 true positives vs. 19 false negatives). This indicates that the model is prone to under-detecting misogynistic content, potentially due to subtle or implicit cues that are not effectively captured by the current multimodal fusion strategy. Misclassification of misogynistic content as non-misogynistic suggests that the textual and visual features may not be sufficiently aligned, particularly in cases where misogyny is conveyed indirectly or through ambiguous visual elements. Addressing these gaps could involve refining feature extraction, implementing more targeted attention mechanisms, and expanding the training set with diverse and nuanced misogynistic examples.

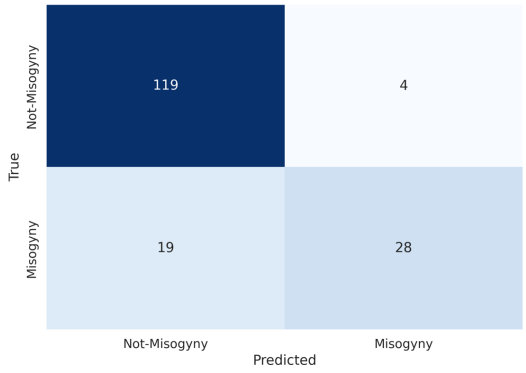

Figure B.1: Confusion matrix of the proposed model

B.2 Qualitative Analysis:

Image Id: 1582.jpg
True Label: Not-Misogyny(0)
Predicted Label: **Not-Misogyny(0)**
Sample 1

Image Id :1342.jpg
True Label: Misogyny(1)
Predicted Label: **Misogyny(1)**
Sample 2

Image Id: 933.jpg
True Label: Misogyny(1)
Predicted Label: **Not-Misogyny(0)**
Sample 3

Image Id :1305.jpg
True Label: Not-Misogyny(0)
Predicted Label: **Not-Misogyny(0)**
Sample 4

Figure B.2: Some outputs predicted by the best model.

The qualitative analysis B.2 shows that the model correctly identifies non-misogynistic content in Sample 1 and Sample 4, proving that it can

recognize harmless content even when the visuals seem aggressive, like the fire scene. In Sample 2, the model correctly detects obvious misogynistic content expressed through clear text, showing that it can identify direct misogynistic messages. However, in Sample 3, the model wrongly labels misogynistic content as non-misogynistic, suggesting that it struggles to detect more subtle or hidden forms of misogyny, especially when sarcasm or cultural references are used. To reduce such errors, the model could be improved by training it with more examples of subtle and indirect misogynistic content and by using attention mechanisms to focus on specific regions or words that convey implicit biases. Strengthening the alignment between textual and visual features could also help in capturing nuanced cues more effectively.