Team_Luminaries_0227@LT-EDI-2025: A Transformer-Based Fusion Approach to Misogyny Detection in Chinese Memes

Adnan Faisal, Shiti Chowdhury,
Momtazul Arefin Labib, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2004002, u2004027, u1904111}@student.cuet.ac.bd,
hasanmurad@cuet.ac.bd

Abstract

Memes, originally crafted for humor or cultural commentary, have evolved into powerful tools for spreading harmful content, particularly misogynistic ideologies. These memes sustain damaging gender stereotypes, further entrenching social inequality and encouraging toxic behavior across online platforms. While progress has been made in detecting harmful memes in English, identifying misogynistic content in Chinese remains challenging due to the language's complexities and cultural subtleties. The multimodal nature of memes, combining text and images, adds to the detection difficulty. In the LT-EDI@LDK 2025 Shared Task on Misogyny Meme Detection, we have focused on analyzing both text and image elements to identify misogynistic content in Chinese memes. For text-based models, we have experimented with Chinese BERT, XLM-RoBERTa and DistilBERT, with Chinese BERT yielding the highest performance, achieving an F1 score of 0.86. In terms of image models, VGG16 outperformed ResNet and ViT, also achieving an F1 score of 0.85. Among all model combinations, the integration of Chinese BERT with VGG16 emerged as the most impactful, delivering superior performance, highlighting the benefit of a multimodal approach. By exploiting these two modalities, our model has effectively captured the subtle details present in memes, improving its ability to accurately detect misogynistic content. This approach has resulted in a macro F1 score of 0.90355, securing 3rd rank in the task.

1 Introduction

The rise of social media has transformed communication but also contributed to the spread of harmful content, including misogynistic memes. These memes combine text and images to reinforce negative gender stereotypes (Gasparini et al., 2022). While research has focused on English memes (Farinango Cuervo and Parde, 2022), misogyny is increasing in Tamil and Malayalam memes (Suryawanshi et al., 2020c). Often humorous (Ponnusamy et al., 2024), they still normalize disrespect toward women (Singh et al., 2024), highlighting the need for multimodal detection models (Huang et al., 2024).

The LT-EDI@LDK 2025 Shared Task on Misogyny Meme Detection tackles the challenge of identifying misogynistic content in memes from Chinese social media, requiring models to analyze both text and images. The task's goal is to classify memes into Misogynistic and Non-misogynistic categories (Ponnusamy et al., 2024; Chakravarthi et al., 2025).

In this study, we have proposed a multimodal framework that integrates textual and visual features for detecting misogynistic content. Integrating BERT (bert-base-chinese) for text representation and VGG16 for visual feature extraction, the fusion of BERT (bert-base-chinese) + VGG16 results in remarkable advancements in performance. It achieves an impressive F1 score of 0.92 on the validation set, underscoring the power of combining both textual and visual data for more accurate hate speech detection. For text-based models, we have experimented with Chinese BERT, XLM-RoBERTa and DistilBERT, with Chinese BERT yielding the highest performance. In terms of image models, VGG16 outperformed ResNet and ViT, demonstrating superior ability in extracting crucial features. The combination of Chinese BERT and VGG16 has proven to be the most effective, yielding the best results in the task. The core contributions of our research work are as follows-

- We have implemented a novel integration of BERT (bert-base-chinese) for text embeddings and VGG16 for image features, significantly improving misogyny detection and classification performance.
- We have developed a multimodal classifier that combines text and image features, im-

proving accuracy while reducing reliance on manual feature extraction.

For a comprehensive guide on the implementation process and to access the complete codebase, please visit the GitHub repository: https://github.com/AJFaisal002/Misogyny-Meme-Detection.

2 Related Work

Misogyny detection has evolved into a critical area of research, initially concentrating on identifying misogynistic content in English memes (Farinango Cuervo and Parde, 2022), but gradually expanding to include multilingual contexts and more complex forms of content. Transformer-based models like BERT and RoBERTa have shown strong performance in understanding nuanced language, especially for multilingual tasks (Devlin et al., 2019; Liu et al., 2019). Early meme detection relied on unimodal models processing text or images separately, limiting their effectiveness. Multimodal approaches like embedding-level fusion (Suryawanshi et al., 2020a) and dual-stage fusion in MemeFier (Koutlis et al., 2023) enhanced performance. Benchmarks from Memotion (2020, 2022) and MultiOFF (Suryawanshi et al., 2020b) have driven progress in offensive content detection. The MDMD dataset by (Ponnusamy et al., 2024) focuses on misogyny in Tamil and Malayalam memes, providing detailed gender bias annotations. The Multitask Meme Classification shared task by (Chakravarthi et al., 2024) explored misogyny and troll content detection, specifically in Tamil and Malayalam, offering insights and benchmarks for current approaches. (Chakravarthi et al., 2025) provide an overview of misogyny meme detection methods and results for Chinese social media, setting benchmarks for this task.

3 Data Description

The Misogyny Meme Detection dataset, derived from Chinese social media, has combined text and image data, split into Train, Dev and Test sets. Each Train and Dev sample includes an image, label and transcription, while the Test set provides only the image and text for classification. The data distribution is shown in Table 1.

Labels	Train	Development	Test
Misogyny	349	47	93
Not-Misogyny	841	123	247
Total	1,190	170	340

Table 1: Data Distribution of Misogyny and Non-Misogyny in Train, Development and Test datasets

4 Methodology

4.1 Problem Formulation

The task is to classify memes as Misogynistic or Non-Misogynistic using multimodal data from Chinese social media. Given a meme m consisting of a text t and an image i, the task is to classify m as either Misogynistic or Non-Misogynistic. Let $t \in R^n$ represent the textual features of the meme and $i \in R^m$ represent the image features. The objective is to learn a mapping function $f(t, i) \rightarrow \{0, 1\}$, where 0 indicates Non-Misogynistic and 1 indicates Misogynistic, using multimodal fusion of both text and image features to maximize classification accuracy.

4.2 Data Preprocessing

The text data has been processed by removing URLs and special characters and Jieba has been used for tokenizing the Chinese text. Due to the moderately balanced class distribution, we have slightly avoided under- or over-sampling. We have experimented with Chinese BERT, DistilBERT and XLM-RoBERTa, but Chinese BERT has proven to be the most effective for feature extraction. This preprocessing has ensured the text is clean, well-structured and ready for model input.

For image data, images have been resized to 224x224 pixels, with random flips and rotations applied for augmentation. We have experimented with ResNet, ViT and VGG16, with VGG16 proving to be the most effective for feature extraction. Techniques like color jitter and rotations have been used to enhance the dataset and improve model performance.

4.3 Uni-modal Models

4.3.1 Text-based Model

We have used BERT (bert-base-chinese), a strong model trained on Chinese text that has understood context well. We have also experimented with XLM-RoBERTa, a multilingual model for many languages, and DistilBERT, a smaller, faster version of BERT. Among these, Chinese BERT has delivered the best results.

4.3.2 Image-based Model

We have experimented with several image feature extraction models, including VGG16 — a deep

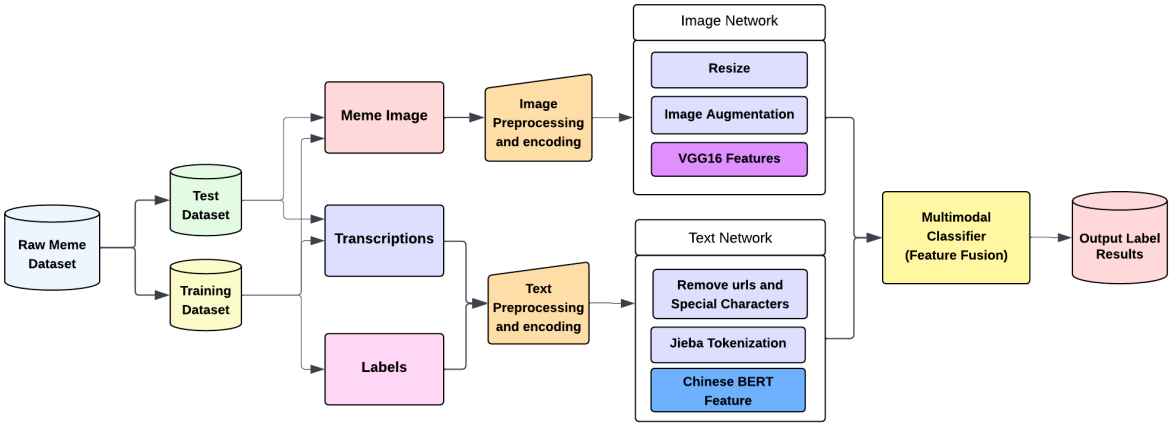

Figure 1: Multimodal Process Flow Framework for Detecting Misogynistic Content in Memes

convolutional neural network known for capturing important visual features — Vision Transformer (ViT), and ResNet, to find the best approach for misogyny meme detection. Among these, VGG16 consistently delivered better results, effectively extracting key visual features crucial for accurately identifying misogynistic memes.

4.4 Fusion Model

We have imple1ented a multimodal classifier using late fusion for simpler integration and better performance than early fusion and attention. We have employed BERT (bert-base-chinese) for text processing, which has proven to be more effective than XLM-RoBERTa and DistilBERT. For image feature extraction, we have explored VGG16, which has outperformed ViT and ResNet in detecting misogynistic memes. Figure 1 shows that combining modalities improves detection of subtle misogynistic details.

4.5 Evaluation Metrics

The models have been evaluated using macro-F1 score, precision and recall to ensure balanced performance and accurate identification of misogynistic content.

5 Results and Analysis

This task has evaluated models using text and image data to detect misogyny. While training performance was strong, test results have revealed overfitting, highlighting the need for better generalization and fusion techniques.

5.1 Task: Multimodal Detection of Misogynistic Memes in Chinese

Table 2 shows the performance of models in the task of Misogyny Meme Detection for Chinese. Chinese BERT leads with an impressive F1 score of 0.86, surpassing XLM-RoBERTa (0.83) and DistilBERT (0.79). Among image-based models, VGG16 outperforms ResNet and ViT with a strong F1 score of 0.85. The most powerful combination is the fusion of Chinese BERT and VGG16, which achieves a remarkable F1 score of 0.92, highlighting the effectiveness of combining textual and visual features for optimal detection performance.

Model	Classifier	P	R	F1
Unimodal (Text)	XLM-RoBERTa	0.82	0.85	0.83
	DistilBERT	0.78	0.81	0.79
	Chinese BERT	**0.84**	**0.88**	**0.86**
Unimodal (Image)	ViT	0.80	0.83	0.81
	ResNet	0.79	0.82	0.80
	VGG16	**0.84**	**0.87**	**0.85**
Multi-modal	(XLM-RoBERTa + ViT)	0.84	0.86	0.85
	(DistilBERT + ResNet)	0.81	0.83	0.84
	(**Chinese BERT + VGG16**)	**0.86**	**0.91**	**0.92**

Table 2: Model performance comparison for unimodal and multimodal classifiers.

Chinese BERT and VGG16 together outperform other models due to their exceptional capabilities in processing text and images. Chinese BERT effectively handles Mandarin text, while VGG16 excels in image feature extraction. Despite this, the model showed signs of overfitting, which we have addressed by implementing early stopping and applying a 0.3 dropout for regularization. This combination has enhanced classification accuracy while minimizing overfitting.

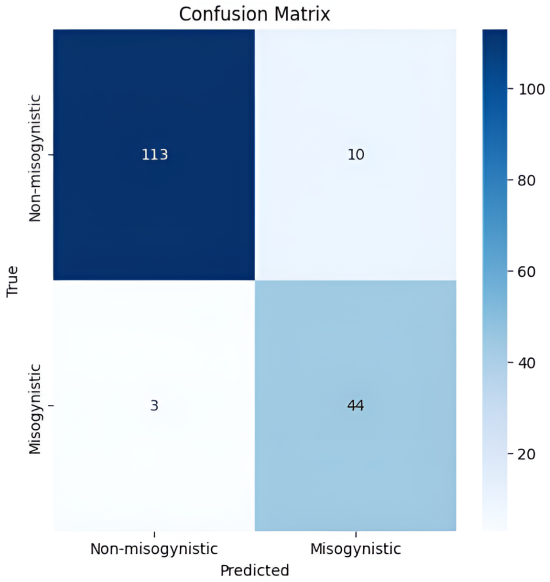

Figure 2: Confusion Matrix for Misogyny Meme Detection

Figure 2 shows the confusion matrix for the Misogyny Meme Detection model, illustrating its ability to differentiate between non-misogynistic and misogynistic texts. The model has correctly identified 113 non-misogynistic (TP) and 44 misogynistic (TN) texts, while misclassifying 10 as false positives and 3 as false negatives, indicating strong overall accuracy with few errors.

5.2 Parameter Setting

Table 3 lists the hyperparameters used for training XLM-RoBERTa + ViT, DistilBERT + ResNet, and Chinese BERT + VGG16. A learning rate of 1×10^{-5} or 2×10^{-5}, AdamW optimizer, and batch size of 8 have contributed to improved performance and reduced overfitting in multimodal misogyny detection.

Model	Learning Rate	Optimizer	Batch Size
XLM-RoBERTa + ViT	2e-5	AdamW	8
DistilBERT + ResNet	1e-5	AdamW	8
Chinese BERT + VGG16	1e-5	AdamW	8

Table 3: Key Hyperparameters for Model Training

6 Conclusion

The LT-EDI@LDK 2025 Shared Task highlighted challenges in detecting misogynistic Chinese memes, where traditional models struggled with subtle language and visuals. Chinese BERT performed well but overfitting remained an issue. Multimodal fusion of text and images improved detection by capturing nuanced patterns, helped by regularization and fine-tuning. These results stress the importance of multimodal methods and diverse datasets. Although late fusion showed promise, the model's use beyond misogyny detection is limited by domain and cultural factors. Future work should explore broader datasets for better generalization. This system can be integrated into real-time moderation to improve automated harmful content detection and intervention.

Error Analysis

The model has struggled with detecting subtle misogynistic content, as shown in the confusion matrix, misclassifying non-misogynistic memes and missing some misogynistic ones. Despite experimenting with various models, the fusion of Chinese BERT and VGG16 has proven most effective. Class imbalance has been addressed with resampling and class weight adjustments. Further improvements in multimodal fusion and handling indirect misogyny are expected to boost accuracy.

Limitations

The model has faced issues with overfitting and multimodal integration. Cultural nuances in Chinese memes have been hard to capture and manual validation has been needed for back translation. Insufficient training data has led to poor generalization and the model may struggle with other languages without further adaptation. This study offers limited insight into cultural nuance handling and multimodal fusion, highlighting key areas for future exploration.

Ethical Statement

All data processing and modeling followed ethical guidelines for handling sensitive, misogynistic content. The study aims to improve misogyny detection while protecting users' rights and privacy. The goal is to enhance moderation on online platforms and create safer spaces, free from misogyny. We have addressed any biases or limitations in the dataset to the best of our ability.

Acknowledgement

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. We also thank earlier work, including transformer models like BERT and RoBERTa for advancing multilingual tasks. We also acknowledge Suryawanshi et al. (2020) and Koutlis et al. (2023)

for their contributions to multimodal fusion, and the MDMD dataset (Ponnusamy et al., 2024) and Multitask Meme Classification task (Chakravarthi et al., 2024) for valuable benchmarks in misogyny detection.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Charic Farinango Cuervo and Natalie Parde. 2022. Exploring contrastive learning for multimodal detection of misogynistic memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 785–792, Seattle, United States. Association for Computational Linguistics.

Francesca Gasparini, Elisabetta Fersini, 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@NAACL 2022*, pages 533–549, Seattle, Washington, United States. Association for Computational Linguistics.

Anzhong Huang, Qiuxiang Bi, Luote Dai, and Yinghui Ma. 2024. Investigating the impact of financial development on the resource curse through its dual effect. *Resources Policy*, 86:104174.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memefier: Dual-stage modality fusion for image meme classification. *Preprint*, arXiv:2304.02906.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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.

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*. Just Accepted.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020a. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020b. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Pranav Verma, Mihael Arcan, John Philip McCrae, and Paul Buitelaar. 2020c. A dataset for troll classification of tamilmemes. In *Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation*, pages 7–13, Marseille, France. European Language Resources Association (ELRA).