CUET_Ignite@LT-EDI-2025: A Multimodal Transformer-Based Approach for Detecting Misogynistic Memes in Chinese Social Media

MD.Mahadi Rahman , Mohammad Minhaj Uddin , Mohammad Oman
and **Mohammad Shamsul Arefin**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904094, u1904118, u1904085}@student.cuet.ac.bd, sarefin@cuet.ac.bd

Abstract

Misogynistic content in memes on social media platforms poses a significant challenge for content moderation, particularly in languages like Chinese, where cultural nuances and multimodal elements complicate detection. Addressing this issue is critical for creating safer online environments, A shared task on multimodal misogyny identification in Chinese memes, organized by LT-EDI@LDK 2025, provided a curated dataset for this purpose. Since memes mix pictures and words, we used two smart tools: ResNet-50 to understand the images and Chinese RoBERTa to make sense of the text. The data set consisted of Chinese social media memes annotated with binary labels (Misogynistic and Non-Misogynistic), capturing explicit misogyny, implicit biases, and stereotypes. Our experiments demonstrated that ResNet-50 combined with Chinese RoBERTa achieved a macro F1 score of 0.91, placing second in the competition and underscoring its effectiveness in handling the complex interplay of text and visuals in Chinese memes. This research advances multimodal misogyny detection and contributes to natural language and vision processing for low-resource languages, particularly in combating gender-based abuse online.

1 Introduction

Misogynistic content fuels hostility and discrimination, particularly targeting women, and poses a significant barrier to fostering safe and inclusive online spaces. These memes flooding Chinese social media platforms like Weibo aren't just harmless jokes—they're digital barbs that mock women, blending snarky text with images to spread hostility (Kiela et al., 2020). Detecting misogyny in these multimodal formats is complex, as the intent hinges on the interplay between visual and textual elements (Chen and Pan, 2022). Subtle misogyny can dodge automated tools, or worse, those tools might flag innocent posts by mistake (Jindal et al., 2024). This is not just a technological problem, it is a social one, as these memes shape attitudes and amplify harm. The Misogynistic Meme Detection Shared Task at LT-EDI@2025 took aim at this, challenging teams to spot harmful memes in Chinese social media with precision. Our team, CUET_Ignite, participated in the LISN 2025 CodaLab competition to wrestle with these issues. We set out to build a system that could handle the tricky interplay of images and Chinese text. Our key contributions include the following:

- Used ResNet-50 to dig into images and Chinese RoBERTa to decode Chinese text, nailing the visual and linguistic cues of misogyny.
- Ran tests on image-only and text-only models to figure out which pulls more weight, landing an F1-score of 0.91 for solid accuracy and balance.

Inspired by (Rahman et al., 2025), which tackled abusive Tamil text with transformers, we pushed their ideas into the multimodal world of Chinese memes. This is our working way of making the Internet less toxic. For more details, our code is available at https://github.com/MHD094/Chinese-Misogyny-Meme-Detection.

2 Related Work

Social media is packed with harmful content such as misogyny and hate speech. In recent years, NLP researchers have been working on spotting trolled, hostility, and abusive content on social media. The early work was mostly about text alone (Anzovino et al., 2018). (Nozza et al., 2021) showed that hate speech tools struggle with different hate types, so misogyny needs its own focus.

Now, researchers are working on memes that mix text and images, making things trickier. Recent research has investigated multimodal approaches

to boost classification performance. For example, (H et al., 2024) developed a method to label Tamil and Malayalam memes as "Misogynistic" or "Non-Misogynistic" using Multinomial Naive Bayes, merging results with weighted probabilities. (Chen et al., 2024) used a CLIP model to see how text and pictures work together in misogynistic memes. (Mahesh et al., 2024) studied Tamil and Malayalam memes, pairing mBERT or MuRIL with ResNet-50, hitting F1-scores of 0.73 and 0.87. (Attanasio et al., 2022) built a Perceiver IO system, blending ViT for images and RoBERTa for text, which performed well at catching misogyny in memes. (Jha et al., 2024) launched MultiBullyEx, a dataset for cyberbullying memes in mixed languages. A Contrastive Language-Image Pretraining (CLIP) projection based multimodal shared-private multitask approach has been proposed there for visual and textual explanation of a meme. (Ahsan et al., 2024) shared MIMOSA, with 4,848 Bengali memes, using a fusion method to sort aggression. (Zhou et al., 2024) introduced Multi3Hate, a multilingual meme dataset capturing cultural variability in hate interpretation, and evaluated several vision-language models on this task. Similarly, (Lee et al., 2022) proposed Hate-CLIPper, which achieved state-of-the-art results by modeling cross-modal interactions between CLIP-encoded image and text features. These efforts show how hard it is to catch harmful memes in different languages and cultures, especially with subtle humor or jabs. Our work at CUET_Ignite@LT-EDI-2025 (Chakravarthi et al., 2025) stands out as we used ResNet-50 and Chinese RoBERTa to nab misogynistic Chinese memes, hitting an F1-score of 0.91. We addressed Chinese slang, idioms, and cultural vibes, making our model relevant for China's social media and helping to keep online spaces safer.

3 Task and Dataset Description

The pervasive spread of harmful content on social media, especially misogynistic material, has become increasingly common often hidden within memes that combine both text and images. These memes can reinforce negative stereotypes and promote gender-based hate speech. This work focuses on building automated systems that detect misogynistic memes by jointly analyzing visual and textual information, specifically in Chinese language memes. It is a multimodal classification task, where each meme must be categorized as either:

Misogynistic: Memes that contain content demeaning, targeting, or offending women.
Non-Misogynistic: Memes without harmful or offensive intent toward women.
The dataset requires analyzing both the image and the accompanying Chinese text, making the task challenging in the fields of Natural Language Processing and Computer Vision. It also contributes toward advancing multimodal and multilingual AI systems for hate speech detection.

This dataset builds upon the MDMD (Misogyny Detection Meme Dataset) originally introduced by (Ponnusamy et al., 2024), which focused on Tamil and Malayalam memes. The present dataset extends their methodology and annotation guidelines to Chinese social media content. A detailed overview of dataset design and objectives is also provided in (Chakravarthi et al., 2024). To ensure consistency with the original task, the same annotation schema was adopted and adapted for Chinese memes.

Classes	Train	Development	Test
Misogyny	349	47	104
Non-Misogyny	841	123	236
Total	1190	170	340

Table 1: Dataset distribution.

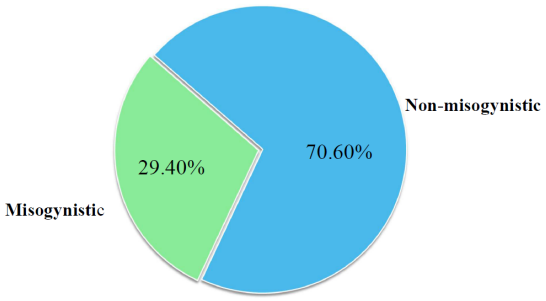

Figure 1: Percentage distribution of two different classes.

The dataset shows moderate imbalance, with 500 misogynistic and 1,200 non-misogynistic samples across all subsets. A total of 1200 memes are included: 1190 for training, 170 for development, and 340 for testing.

4 Methodology

The objective of this study is to detect misogynistic content in multimodal Chinese memes by integrating visual and textual features. Our approach begins with preprocessing the memes, followed

by feature extraction from both modalities. The features are then fused using an attention-based mechanism, and a classifier predicts whether the meme is misogynistic or non-misogynistic. Figure 2 provides a visualization of our methodology.

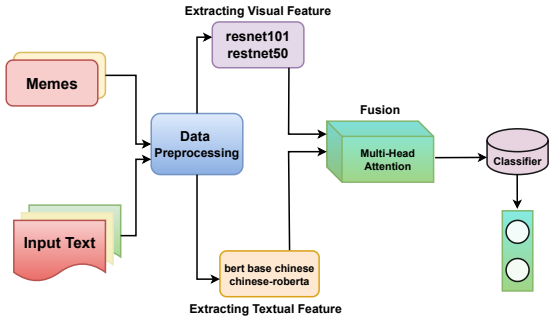

Figure 2: An abstract view of the proposed methodology

4.1 Data Preprocessing

In this step, we preprocess Chinese text and images for model compatibility. Text is tokenized using the hfl/chinese-roberta-wwm-ext tokenizer (Cui et al., 2020), transformed into 128-token numerical representations with [CLS] and [SEP] tokens, leveraging Chinese-RoBERTa's vocabulary for slang. Images are resized to 224×224 pixels, normalized with ImageNet statistics (Sayma et al., 2025), and converted to RGB.

4.2 Visual Approach

For visual feature extraction, we experiment with two pre-trained convolutional neural network (CNN) models: ResNet-50 and ResNet-101 (He et al., 2016), both pre-trained on ImageNet. The fully connected layer of each model is replaced with an identity layer to extract 2048-dimensional feature vectors. These models were chosen for their ability to capture complex visual patterns, such as culturally nuanced imagery or humorous elements in Chinese memes.

4.3 Textual Approach

The textual component of memes is processed using transformer-based models optimized for Chinese. We experiment with BERT-Base-Chinese and Chinese-RoBERTa-wwm-ext, both leveraging pre-trained weights. The [CLS] token's output from the last hidden state is extracted, generating 768-dimensional feature vectors. Chinese-RoBERTa-wwm-ext is prioritized for its whole-word masking strategy, which enhances its ability to capture contextual nuances in meme-specific language (Cui et al., 2021).

4.4 Multimodal Approach

Our multimodal approach combines the visual and textual features through a fusion strategy. The visual features from ResNet-50 or ResNet-101 (2048-dimensional) and textual features from Chinese-RoBERTa-wwm-ext or BERT-Base-Chinese (768-dimensional) are first projected to a common 512-dimensional space using linear layers, each followed by ReLU activation. These projected features are then fused using a multi-head attention mechanism (8 heads, embed dim=512) to capture cross-modal interactions, with the attention output averaged to produce a 512-dimensional representation (Wang et al., 2024). This combined representation is processed through a two-layer neural network classifier. The first layer reduces the dimensionality to 512, followed by ReLU activation and dropout (0.3) for regularization. The final layer produces binary classification outputs for misogyny detection. The training protocol uses Adam (learning rate: 1e-5, batch size: 16) for 10 epochs, with a weighted cross-entropy loss to address class imbalance, where class weights are computed as the inverse of class frequencies. This strategy aligns with recent approaches in multimodal harmful meme detection that emphasize cross-modal attention for enhanced feature fusion (Huang et al., 2024). Table 2 shows the list of tuned hyperparameters used in the experiment.

Hyperparameters	Value
Optimizer	Adam
Learning rate	1e-05
Epochs	10
Batch size	16
Dropout Rate	0.3

Table 2: Overview of optimized hyper-parameters.

5 Results & Discussion

This section presents a comparative performance analysis of various experimental approaches for classifying Chinese memes as misogynistic or non-misogynistic. The effectiveness is primarily assessed based on the weighted F1-score, while precision and recall are also considered in some cases. Table 3 presents a summary of the precision (P), recall (R), and F1 (F1) scores for each model on the test set. The results show that ResNet-50

and Chinese-RoBERTa-wwm-ext performed best among the visual and textual models, respectively, with an F1-score of 0.73 and 0.84. However, the top classification performance was observed in the multimodal models, where combining Chinese-RoBERTa-wwm-ext and ResNet-50 resulted in the highest F1-score of 0.91. These findings highlight the superiority of multimodal models in meme classification by effectively integrating text and visual features.

Approach	Classifier	P	R	F1
Visual	ResNet-101	0.70	0.72	0.71
	ResNet-50	0.73	0.74	0.73
Textual	Bert-Base-Chinese	0.76	0.77	0.77
	Chinese-Roberta	0.85	0.84	0.84
Multimodal	Bert-Base-Chinese + ResNet-101	0.83	0.86	0.84
	Chinese-Roberta + RestNet-50	**0.92**	**0.90**	**0.91**
	Bert-Base-Chinese + RestNet-50	0.88	0.83	0.85

Table 3: Evaluation of various models on the test set.

5.1 Quantitative Discussion

The results highlight the effectiveness of multimodal models in identifying misogynistic content in Chinese memes. The confusion matrix in Figure 3 shows that the multimodal model (Chinese-RoBERTa-wwm-ext + ResNet-50) outperforms unimodal approaches, correctly classifying 228 Not-Misogyny and 87 Misogyny instances, with fewer misclassifications (8 false positives and 17 false negatives). These findings affirm that leveraging both visual and textual features improves precision and recall in detecting misogynistic memes, particularly in reducing false positives.

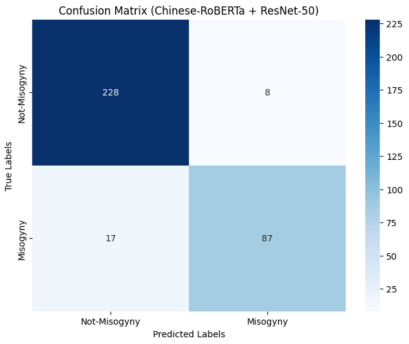

Figure 3: Confusion matrix of best performing approach.

5.2 Qualitative Discussion

Figure 4 presents sample predictions from our best-performing Chinese-RoBERTa-wwm-ext + ResNet-50 model. The first sample, incorrectly classified as non-misogynistic (label 0), contains

English Translation:
cheap person

Actual: 1
Predicted: 0

English Translation:
Baby, aren't you going to coax me to sleep?

Actual: 0
Predicted: 1

Figure 4: Examples of some misclassified samples from the top-performing model.

a derogatory term indicating misogyny (label 1), likely misclassified due to the model interpreting the cartoon character's mischievous expression as playful. Conversely, the second sample, genuinely non-misogynistic (label 0), was misclassified as misogynistic (label 1), possibly because the distressed cartoon cat's expression suggested conflict, despite the text's lighthearted tone. Cultural norms in Chinese internet memes, involving exaggerated expressions, may have influenced these errors. The multimodal model struggles with nuanced cases involving culturally specific language and ambiguous visuals, a challenge also seen in experiments with BERT-Base-Chinese and ResNet-101.

6 Conclusion

This work presented the details of the methods and performance analysis of the models for detecting misogynistic memes in Chinese, exploring visual, textual, and multimodal fusion techniques. The results revealed that the Chinese-RoBERTa-wwm-ext + ResNet-50 model achieved the highest F1-score of 0.91, demonstrating that multimodal fusion with attention mechanisms significantly enhances model performance. The attention-based fusion effectively captured cross-modal interactions, leading to improved precision and recall compared to unimodal approaches. In the future, we plan to explore advanced fusion strategies, such as cross-attention or graph-based methods, and extend the dataset to include more diverse meme content for better robustness, especially in handling Chinese internet slang and culturally specific references. Adding Chinese cultural knowledge and reducing model biases enhances adaptability, fairness, and generalization.

Limitations

A key limitation of this study stems from its dependence on pre-trained models for visual and textual feature extraction, which may not adequately address the intricacies of Chinese meme culture and context. Although the multimodal framework yields strong results, these models often lack the ability to generalize effectively to culturally specific or niche meme content. Moreover, the training dataset may not fully encompass the diverse range of Chinese memes, potentially undermining the model's robustness. The influence of cultural elements, such as humor, irony, and regional slang prevalent in Chinese online spaces, has also not been thoroughly examined. Misogynistic intent can often be conveyed indirectly through satire or cultural references, posing challenges for AI models in accurately discerning intent. Future efforts should focus on expanding the dataset, developing Chinese-specific models, and conducting in-depth analyses of humor and cultural influences to improve accuracy and adaptability. While the dataset was balanced and did not necessitate augmentation, applying data augmentation techniques in future work with larger, imbalanced datasets—through synthetic text or image transformations—could mitigate class imbalances and enhance generalization across diverse categories. This approach would lead to better performance in underrepresented scenarios, fostering a more resilient and effective model for practical deployment.

References

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A multimodal framework to detect target aware aggression in memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2487–2500, St. Julian's, Malta. Association for Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, page 57–64, Berlin, Heidelberg. Springer-Verlag.

Giuseppe Attanasio, Debora Nozza, and Federico Bianchi. 2022. MilaNLP at SemEval-2022 task 5: Using perceiver IO for detecting misogynous memes with text and image modalities. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 654–662, Seattle, United States. Association for Computational Linguistics.

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.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora Salim. 2024. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1864–1871, New York, NY, USA. Association for Computing Machinery.

Yuyang Chen and Feng Pan. 2022. Multimodal detection of hateful memes by applying a vision-language pre-training model. *PLOS ONE*, 17:e0274300.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Shaun H, Samyuktaa Sivakumar, Rohan R, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@LT-EDI 2024: A SVM-ResNet50 approach for multitask meme classification - unraveling misogynistic and trolls in online memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 221–226, St. Julian's, Malta. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 770–778, Las Vegas, NV, USA. IEEE.

Jianzhao Huang, Hongzhan Lin, Ziyan Liu, Ziyang Luo, Guang Chen, and Jing Ma. 2024. Towards low-resource harmful meme detection with lmm agents. *arXiv preprint arXiv:2411.05383*.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. 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.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jungbin Lee, Sungrae Jin, Donghyun Kim, Jihie Kim, and Jinwook Kim. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05916*.

Sidharth Mahesh, Sonith D, Gauthamraj Gauthamraj, Kavya G, Asha Hegde, and H Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring joint representation for memes classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 282–287, St. Julian's, Malta. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. 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.

MD.Mahadi Rahman, Mohammad Minhaj Uddin, and Mohammad Shamsul Arefin. 2025. CUET_Ignite@DravidianLangTech 2025: Detection of abusive comments in Tamil text using transformer models. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 392–397, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Khadiza Sultana Sayma, Farjana Alam Tofa, Md Osama, and Ashim Dey. 2025. CUET_Novice@DravidianLangTech 2025: A multimodal transformer-based approach for detecting misogynistic memes in Malayalam language. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 472–477, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Xinyi Wang, Jun Liu, Min Zhang, and Wei Chen. 2024. Toxicn mm: A multimodal benchmark for chinese harmful meme detection. *arXiv preprint arXiv:2410.02378*.

Yujie Zhou, Yutai Ge, Qian Liu, Yue Zhang, and Paul Rottger. 2024. Multi3hate: Multimodal, multilingual, and multicultural hate speech detection with vision-language models. *arXiv preprint arXiv:2411.03888*.