SSNCSE@LT-EDI-2025:Detecting Misogyny Memes using Pretrained Deep Learning models

Sreeja K, Bharathi B
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
sreeja2350625@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Misogyny meme detection is identifying memes that are harmful or offensive to women. These memes can hide hate behind jokes or images, making them difficult to identify. It's important to detect them for a safer and respectful internet for everyone. Our model proposed a multimodal method for misogyny meme detection in Chinese social media by combining both textual and visual aspects of memes. The training and evaluation data were part of a shared task on detecting misogynistic content. We used a pretrained ResNet-50 architecture to extract visual representations of the memes and processed the meme transcriptions with BERT. The model fused modality-specific representations with a feed-forward neural net for classification. The selected pretrained models were frozen to avoid overfitting and to enhance generalization across all classes, and only the final classifier was fine-tuned on labelled meme recollection. The model was trained and evaluated using test data to achieve a macro F1-score of 0.70345. As a result, we have validated lightweight combining approaches for multimodal fusion techniques on noisy social media and how they can be validated in the context of hostile meme detection tasks.

1 Introduction

The rise of misogynistic content on social media contexts is increasingly problematic, particularly as social media platforms adopt more multimodal formats such as memes that combine text and images to propagate problematic, exclusionary, or unfair messages. In the context of multilingual and multicultural online spaces that include languages such as Chinese, Tamil, Malayalam, and Hindi-English code-mixed communities, identifying misuse introduces additional challenges associated with language, culture, and multimodal resources.

Recently, researchers have reported the difficulties in identifying misogynistic memes. The authors, (Lu et al., 2024) introduced the ToxiCN-MM dataset, the first large-scale collection of harmful memes in Chinese, and proposed a Multimodal Knowledge Enhanced (MKE) model tailored for culture-specific meme detection. Their findings demonstrate the complicated nature of identification models and consider the value of contextual and cultural knowledge in detection models. Similarly, shared tasks and resources such as the (Chakravarthi et al., 2024) and (Pattanaik et al., 2025) have taken this research into low-resource Dravidian languages such as Tamil and Malayalam, collecting and collaborating on annotated datasets such as MDMD, which advance the development of systems that can identify code-mixed, as well as monolingual, data.

The MIMIC project also released a large set of human-annotated examples that will detect misogyny in Hindi-English code-mixed memes and can further multimodal hate speech research in minority languages. These two projects show not only the need for effective image-text fusion-based models but also the need to consider cultural aspects of misogyny and multimodal aspects of the data when attempting to detect misogyny in online discourse.

The rest of the paper is organized as follows: Section 2 analyzes the related works done in the previous research, and Section 3 discusses the misogyny meme corpus in the current work. Section 4 contains a detailed discussion of the proposed models used in the current work. Section 5 explains the experimental results. Section 6 discusses the limitations. In Section 7 concludes the paper.

2 Related work

The arrival of social media has brought awareness to the issue of misogyny in multimodal content, including memes.(Chakravarthi et al., 2025)provides the overview of the Misogyny Meme Detection Shared Task for Chinese Social Media. The

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 1–5

task of detecting misogyny in multimodal content with the MAMI dataset was handled (Singh et al., 2023), who found that multimodal models combining BERT for text and Vision Transformer (ViT) for images, when pretrained on hate speech data, substantially outperformed unimodal models. As noted here, (Liu et al., 2019) introduced RoBERTa and specifically focused on building on BERT by reviewing the positive aspect of pretraining on a significant amount of large-scale data and longer in both iterations and time. In that evaluation, RoBERTa beat BERT in a variety of language tasks, reinforcing the trend of using pretrained models, frozen as feature extractors. In a similar vein, (Deng et al., 2022) proposed the COLD benchmark to identify offensive language in Chinese and established the claim that Singh et al. (2023)BERT-based models could identify nuance, including biases related to race, gender, and region. As multimodal approaches matured, the use of pretrained models for both textual and visual understanding became increasingly common. Memotion 2.0 was introduced in a recent study (Ramamoorthy et al., 2022), and prior studies showed improved meme sentiment and emotion classification by combining ResNet-50 for visual features and BERT for textual features. A recent contribution to the field (Gasparini et al., 2022) developed a benchmark dataset for meme-based misogyny detection (both direct and indirect) and emphasized the importance of expert annotations in overcoming sociocultural barriers and interpreting hostile or ironic content. Another study (Chakravarthi et al., 2024) illustrated that lightweight classifiers such as MLPs are effective when used with frozen feature extractors in multilingual meme classification.

While much of the research has focused on English and Indian languages, there is limited work on misogyny detection in multimodal Chinese-language content. This study addresses that gap by establishing a pipeline that combines ResNet-50 and BERT, both used in frozen mode, with an MLP classifier—achieving a macro-averaged F1-score of 0.70345. This demonstrates the feasibility of developing a robust misogyny detection model for Chinese social media using pretrained models and lightweight classification.

This study expands on previous work by applying multimodal detection of misogyny to Chinese, a linguistically and culturally important domain that has rarely been addressed in existing literature. Many previous studies were primarily based on fine-tuning of pretrained models. In this study, we employed frozen pretrained models (BERT and ResNet-50) along with a minimal MLP classifier, to demonstrate that there's an effective and efficient usage of pretrained models in this, and many other, domains. Additionally, our strong performance in identifying misogyny in Chinese memes indicates that multimodal pretrained are generalizable to domains beyond English and Indian languages. This provides more breadth and scalability to online multilingual safety prevention research.

3 Dataset Description

The task aims to develop a model for Misogyny Meme detection. The dataset for the Shared Task on Misogyny Meme Detection found at LT-EDI@LDK 2025(Ponnusamy et al., 2024), (Chakravarthi et al., 2024) has been designed to support multimodal and multilingual research on Chinese social media data. The dataset is comprised of memes, which consist of both an image and an accompanying textual component that are typically captured from an overlay caption or a comment historically associated with the image. The image contains a variety of meme formats shown on online forums. Each of the memes is assisted by a binary label, "Misogynistic" or "Non-misogynistic," which conveys whether the meme generates or expresses any form of gender hatred or bias. The dataset consists of three (3) datasets for training, development, and testing. The label distribution for training and development data is mentioned in Fig1 and Fig2.

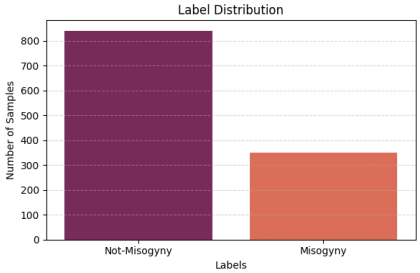

Figure 1: Training Data distribution

4 Proposed Methodology

This study presents a structured multimodal meme content classification approach into misogynistic and non-misogynistic classes. The study involves five basic steps: data preprocessing, feature extraction, model building, training, and prediction. Each step is well designed to address the unique chal-

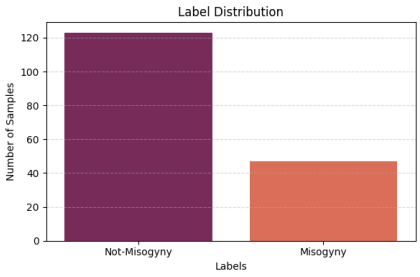

Figure 2: Development data distribution

lenges of integrating visual and textual information on low-context social media images.

4.1 Data Preprocessing

It ensures consistency and tidiness on both image and text modalities. Images are resized to a standard dimension of 224×224 and normalized via ImageNet mean and standard deviation values. Text data in the form of meme transcriptions is cleaned and tokenized with the pretrained BERT tokenizer. A PyTorch dataset class is employed to refine this process, ensuring alignment between text and image inputs and efficiently handling both training and test data formats.

4.2 Feature Extraction

Pretrained models are leveraged based on their stability to produce rich, high-level representations. The ResNet-50 model handles visual modality. It is pretrained on the ImageNet dataset. This model is frozen to prevent overfitting and preserve common visual features. At the same time, transcriptions are passed through a pretrained BERT-base model, and only the [CLS] token representation (pooler output) as the sentence-level embedding is used. Both ResNet-50 and BERT remain untrained to reduce computational costs and utilize the power of large-scale pretraining.

4.3 Model Architecture

It combines both modalities by concatenating the 2048-dimensional feature vector for images with 768-dimensional text embeddings to create a 2816-dimensional feature vector for multimodal representation. This vector is categorized using a lightweight feedforward classifier made of a fully connected layer, activation of ReLU, dropout regularization, followed by a classification layer that outputs logits for the binary classification as misogynist or not-misogynist.

4.3.1 Architecture Workflow

The architecture works in the following steps:

1. **Input Processing**
   Meme images and their corresponding transcriptions are taken as input.

2. **Visual Feature Extraction**
   The image is passed through a frozen ResNet-50 model, which gives a high-level visual feature representation of size 2048.

3. **Text Feature Extraction**
   The text is tokenized and passed through a frozen BERT-base model. The [CLS] token embedding is used as a 768-dimensional text representation.

4. **Feature Fusion**
   The image and text features are combined by simple concatenation to form a single 2816-dimensional feature vector.

5. **Classification**
   This fused feature is passed through a small feedforward neural network (MLP) that includes:

   - A fully connected layer with ReLU activation
   - Dropout for regularization
   - A final layer that predicts if the meme is misogynistic or not

4.4 Training Phase

The classifier parameters are trained while the pretrained backbone models are frozen. The model is trained using the Adam optimizer and a cross-entropy loss function. We use a batch size of 8 and a learning rate 1e-4 for five training epochs. The goal of this targeted training is to limit the risk of overfitting and also to allow the model to converge faster.

4.5 Prediction

The trained model will be used to predict unseen test data. Since the test samples now have contextual textual information associated with them, we will use both the images and their text in the prediction process. The model will interpret this multi-input to predict the probability of misogyny of any given sample. The predictions will be saved in a CSV file, with the predicted labels.

In summary, this methodology can provide an efficient and scalable pipeline for multimodal classification through the integration of a pretrained visual and language model with a selection of effective training and modularity, Multimodal features can be successfully leveraged while computationally efficient.

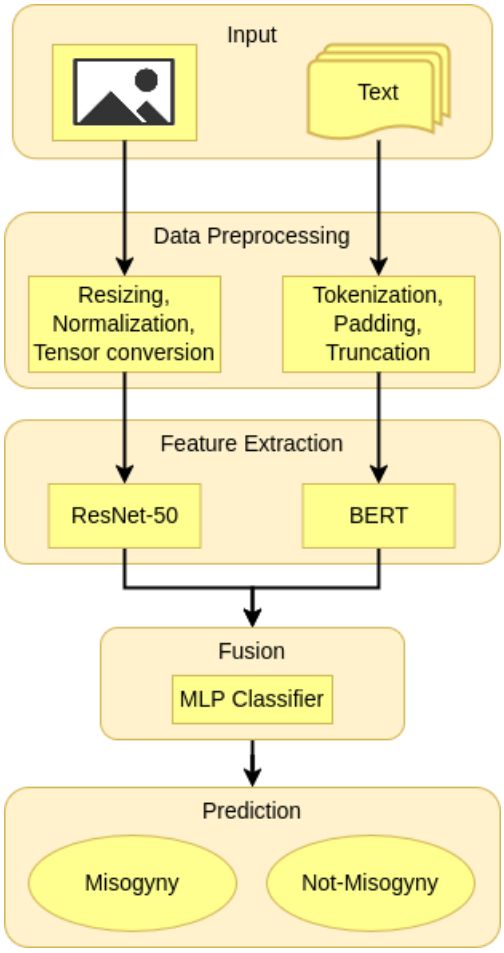

Figure 3: Architecture Diagram of Proposed Work

5 Experimental result

The proposed multimodal misogyny classification model, using both visual and text features, using ResNet50 and BERT respectively, shows that the model is learning sufficiently over the training periods of interest. The loss steadily decreases over each of the epochs, which indicates sufficient learning or convergence; for example, the training loss drops from approximately 0.55 in the first epoch to approximately 0.37 in the last epoch.

To assess the generalizability of the model, performance was evaluated on a development dataset, as shown in Table 1. The evaluation used familiar classification metrics of accuracy, precision, recall, and F1-score. The classifier achieved a macro-averaged F1-score of 0.70345, which means the performance was balanced across both "Misogyny" and "Not-Misogyny" classes. The source code for the proposed approach and found here [1].

Category	Precision	Recall	F1-score
Misogyny	0.74	0.62	0.67
Not-misogyny	0.86	0.92	0.89
Accuracy		0.84	

Table 1: Classification report for Development data

6 Limitations

The proposed method is efficient, but it has the following limitations:

- The ResNet-50 and BERT architectures are implemented without fine-tuning, which limits the model's ability to learn task-relevant features fundamental to the interpretation of misogynistic content.
- Feature fusion happens by concatenation and shallow multilayer perceptron, and does not account for complex interactions between image and text modalities.
- The classification framework provides binary output only, which does not provide the capacity to distinguish between different types of misogynistic expression or to discriminate intensity, severity, or type among misogynistic expressions.
- The model relies on provided transcriptions and does not extract text from embedded text-images to assess memes when the meme text is part of the image.

7 Conclusion

This paper described a multimodal deep learning approach to the detection of misogyny in memes by utilizing visual and textual modalities together. We took visual features from ResNet-50 and textual features from BERT. Both models used were frozen to save computational cost and training time, and we used a light-weight multilayer perceptron (MLP) to fuse the features and perform binary classification. Overall, the results exhibited that, even

[1] https://github.com/SreejaKumaravel/Misogyny-Meme-Detection

with minimal fine-tuning, the architecture was capable of capturing the implicit and explicit clues of misogyny that exist within memes. The method also proved to be very robust and simple to use for real-world deployments for harmful content moderation. Future work could include end-to-end training, different data augmentations or using cross-modal attention for a more complex fusion.

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.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44:108526.

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.

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.

Sarbajeet Pattanaik, Ashok Yadav, and Vrijendra Singh. 2025. Dll5143@DravidianLangTech 2025: Majority voting-based framework for misogyny meme detection in Tamil and Malayalam. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 191–199, 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.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and 1 others. 2022. Memotion 2: Dataset on sentiment and emotion analysis of memes. In *Proceedings of De-Factify: workshop on multimodal fact checking and hate speech detection, CEUR*.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. "female astronaut: Because sandwiches won't make themselves up there": Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada. Association for Computational Linguistics.