CVF-NITT@LT-EDI-2025: A Vision-Language Approach for Detecting Misogynistic Memes in Chinese Social Media

Radhika K T, Sitara K
National Institute of Technology, Tiruchirappalli, India
406322003@nitt.edu,sitara@nitt.edu

Abstract

Online platforms have enabled users to create and share multimodal content, fostering new forms of personal expression and cultural interaction. Among these, memes—combinations of images and text—have become a prevalent mode of digital communication, often used for humor, satire, or social commentary. However, memes can also serve as vehicles for spreading misogynistic messages, reinforcing harmful gender stereotypes, and targeting individuals based on gender. In this work, we investigate the effectiveness of various multimodal models for detecting misogynistic content in memes. We propose a BERT+CLIP+LR model that integrates BERT's deep contextual language understanding with CLIP's powerful visual encoder, followed by Logistic Regression for classification. This approach leverages complementary strengths of vision-language models for robust cross-modal representation. We compare our proposed model with several baselines, including the original CLIP+LR, and traditional early fusion methods such as BERT + ResNet50 and CNN + InceptionV3. Our focus is on accurately identifying misogynistic content in Chinese memes, with careful attention to the interplay between visual elements and textual cues. Experimental results show that the BERT+CLIP+LR model achieves a macro F1 score of 0.87, highlighting the effectiveness of vision-language models in addressing harmful content on social media platforms.

1 Introduction

Misogyny is broadly defined as the hatred of, aversion to, or prejudice against women. It manifests in various forms, including verbal abuse, stereotyping, objectification, or the dissemination of harmful content through social media. Misogyny often appears subtly or overtly in multimodal formats, combining text and imagery to demean, ridicule, or marginalize women. Online misogyny represents a pervasive and deeply rooted societal issue that perpetuates gender-based discrimination and inequality in virtual environments. This toxic behavior not only undermines efforts toward achieving gender equity but also significantly discourages women's active participation in online platforms—ranging from social media to professional and educational digital spaces—thereby silencing their voices and limiting their opportunities for expression, representation, and empowerment(Mohasseb et al., 2025). The increasing prevalence of such content on social media platforms necessitates robust detection systems to mitigate its harmful societal impact.

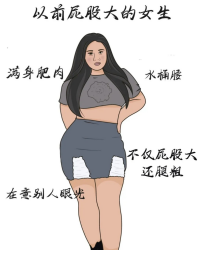

Figure 1: Misogynistic Meme

The Shared Task on Misogyny Meme Detection at LT-EDI@LDK 2025 focuses on building automatic systems to classify memes as *misogynistic* or *non-misogynistic*. This task is especially complex due to the combination of vision and language, as well as the multilingual nature of social discourse—this edition emphasizes Chinese language content. The task encourages advancements in multimodal classification and promotes responsible AI development.

Figure1 depicts a misogynistic meme, shows a cartoon-style illustration of a woman with various derogatory labels surrounding her body. The central theme of the text is body shaming, targeting women with larger body types. The top caption refers to "girls who used to have big butts", setting a critical tone from the start. Additional phrases placed around the figure describe her as "covered

Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 47–53

in fat", having a "barrel waist", and "not only a big butt, but also thick legs". Another phrase suggests that she "cares about others' opinions", implying insecurity or social pressure. Together, these captions convey a negative and mocking portrayal of women who do not conform to conventional beauty standards, specifically critiquing body size and shape. This image exemplifies misogynistic and fatphobic content, as it reinforces harmful stereotypes and societal expectations about women's appearance.

The field of image classification has witnessed considerable advancements due to deep learning models like Convolutional Neural Networks (CNN)(Kalchbrenner et al., 2014) and transformers(Kalyan et al., 2022), which have revolutionized the way we process and understand visual data. In parallel to these advancements, Large Language Models (LLMs), have transformed natural language processing by enabling more nuanced understanding and generation of text(Naveed et al., 2023). The integration of these two fields—vision and language—has led to the development of Vision-Language Models (VLM), which combine the strengths of both visual and textual data. These models often use both an image encoder and a text encoder to generate embeddings, which can be fused for various multimodal tasks(Ghosh et al., 2024).

In this work, we utilize the vision-language model Contrastive Language–Image Pre-training (CLIP), introduced by OpenAI, to detect misogynistic content in Chinese memes. CLIP is a powerful pretrained model that maps images and text into a shared embedding space, enabling effective joint understanding of visual and textual modalities. We explore CLIP's capabilities for image representation and pair it with a traditional Logistic Regression (LR) classifier, striking a balance between performance and computational efficiency. We conducted experiments with traditional multimodal baselines, including Bidirectional Encoder Representations from Transformers (BERT) (Vaswani et al., 2017) combined with Residual Networks (ResNet50) (He et al., 2016), as well as Convolutional Neural Networks (CNN) with InceptionV3 (Szegedy et al., 2016), to benchmark performance across different model architectures. BERT+CLIP+LR model performed well among the models and is made available as an open-source resource on GitHub[1].

2 Related Works

The study by (Lei et al., 2024), presents an explainable hateful meme detection model that employs uncertainty-aware dynamic fusion to improve both generalization and interpretability. By dynamically evaluating the uncertainty of visual and textual modalities, the model assigns adaptive weights for feature fusion. They report that visual features are more influential than textual ones in hateful meme detection, and the model's interpretability aids in understanding its decision-making process, although fairness remains a concern for future work.

The work by (Rizzi et al., 2024) proposes a probabilistic framework for detecting elements of disagreement in misogynistic memes by analyzing both the visual and textual components. It explores various strategies for leveraging these elements to identify instances where annotators may disagree in their interpretations. The EXIST 2024 shared task (Vetagiri et al., 2024) focuses on advancing research in detecting and countering sexism on social networks, a persistent and complex societal issue. CNN-BiLSTM for text and ResNet50-CNN-BiLSTM for memes was presented in the paper(Vetagiri et al., 2024) to better identify explicit and implicit sexist content. The task fosters the development of effective strategies through a competitive framework aimed at improving content moderation.

Study by (Ramamoorthy et al., 2022) marked a significant step forward in understanding memes by creating carefully labeled, high-quality data for analyzing sentiment, classifying emotions, and gauging their intensity. To demonstrate the value of this resource, they established initial performance benchmarks using both a text-based model and a multimodal model that integrated visual features with textual understanding. Their findings highlighted the advantage of considering both text and image content for achieving better results in various meme analysis tasks.

(Ponnusamy et al., 2024) introduced the Misogyny Detection Meme Dataset(MDMD), an annotated resource focused on online misogyny within Tamil and Malayalam-speaking communities, offering valuable insights into gender bias and supporting

[1] https://github.com/CyMa-AI/CVF-NITT-LDK2025.git

efforts to combat digital gender-based discrimination. Additionally, the literature review highlights that the Shared Task on Misogyny Meme Detection at LT-EDI@LDK 2025 marks the focused initiatives aimed at detecting misogyny in memes.

3 Methodology

Vision-Language Models can be broadly categorized into three types: (1) Vision-Language Understanding (VLU) models, which interpret and reason over visual and textual inputs; (2) Text Generation with Multimodal Input, where models generate coherent text based on both image and text inputs; and (3) Multimodal Input-Output models, capable of processing and generating across multiple modalities. CLIP belongs to the VLU category, as it learns joint image-text representations through contrastive pretraining (Li et al., 2023). We build upon CLIP by proposing a BERT+CLIP+LR model, where BERT replaces CLIP's text encoder to enhance contextual language understanding. The resulting image and text embeddings are fused and classified using Logistic Regression. To benchmark performance, we also evaluate the original CLIP+LR model and traditional early fusion baselines such as BERT + ResNet50 and CNN + InceptionV3, enabling a comparative analysis of modern VLMs versus conventional multimodal approaches for harmful content detection.

3.1 Problem Definition

Let $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$ represent a dataset of memes, where each data sample is a pair (x_i, y_i) with $x_i \in \mathcal{X}$ denoting a meme and $y_i \in \mathcal{Y}$ indicating whether the meme contains misogynistic content ($y_i = 1$) or not ($y_i = 0$).
Each meme x_i consists of two modalities: an image component v_i and a text component t_i, so that $\mathcal{X} = (\mathcal{V}, \mathcal{T})$. The task of misogynistic meme detection is formulated as a binary classification problem. The goal is to learn a predictive function:

$$f : \mathcal{V} \times \mathcal{T} \rightarrow \mathcal{Y}$$

which determines whether a given meme $x_i = (v_i, t_i)$ expresses misogynistic content.

3.2 Data preprocessing

Image part of meme is loaded and preprocessed through a pipeline that includes resizing, center cropping, normalization using predefined mean and standard deviation values, and conversion to a tensor. Faulty, corrupted, or unreadable image files are either skipped or replaced with zero or NaN vectors to maintain consistent batch dimensions and prevent downstream errors during model inference. For the textual modality, each caption or transcription is tokenized, lowercased, and then either truncated or padded to fit the model's maximum token length. Missing or malformed text inputs are handled by substituting neutral placeholders such as "[UNK]" tokens or blank vectors, ensuring input consistency across the dataset.

3.3 Feature Extraction in the Proposed Models

1. **CLIP+LR** :
   The general architecture of the CLIP-based classification model involves separate image and text encoders that generate embeddings, which are then fused into a unified feature vector. CLIP jointly embeds images and text into a shared 512-dimensional space. LR, a widely used linear classification algorithm, is employed for its simplicity, interpretability, and efficiency in high-dimensional spaces (Hosmer Jr et al., 2013). The fused vector obtained from CLIP is passed to the LR classifier to predict the final class label based on the combined visual and textual information. The overall process is illustrated in Figure 2.

2. **BERT+CLIP+LR**: The general architecture of the BERT+CLIP+LR classification model builds upon CLIP by replacing CLIP's original text encoder with BERT(Vaswani et al., 2017), a powerful transformer-based language model known for its deep contextual understanding. In this setup, image inputs are encoded using CLIP's visual encoder, while textual inputs are processed using BERT. The resulting image and text embeddings are concatenated or fused into a single feature vector representing both modalities. This fused vector, which captures complementary visual and linguistic information, is then fed into a LR classifier.

3. **BERT+ResNet50**: ResNet (He et al., 2016) is a deep convolutional neural network known for its ability to train very deep networks using residual connections, making it highly effective for image classification tasks. In

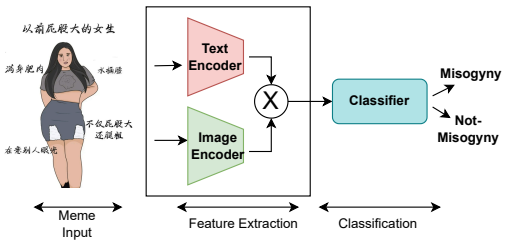

Figure 2: General work flow of proposed models

our multimodal setup, textual features are extracted using BERT, while visual features are obtained from ResNet50. These feature vectors are then concatenated and passed to a classification layer, enabling the model to jointly reason over both modalities. This fusion allows the model to detect nuanced cases of misogyny that may arise from the interaction between image and text in memes.

4. **CNN+InceptionV3**: CNNs treat text as a sequence of word embeddings and learn to capture local dependencies and hierarchical features within the text. For the image component, we employ InceptionV3 (Szegedy et al., 2016), an advanced CNN architecture that uses a multi-branch design to capture features at various levels of abstraction. The multi-branch design enables InceptionV3 to efficiently process images by learning both fine-grained details such as edges and textures and larger, high-level patterns.

3.4 Early Fusion

Instead of explicit concatenation, CLIP encodes each modality independently and enables implicit early fusion through its contrastive pretraining objective. By aligning image and text embeddings in a shared semantic space, CLIP naturally captures cross-modal relationships without requiring manual fusion strategies. In contrast, for traditional models, we implemented explicit early fusion by first extracting textual features using CNN or BERT and visual features using InceptionV3 or ResNet50. These unimodal embeddings were then concatenated to form a joint multimodal feature vector, which was fed into downstream classification. This approach, while effective in controlled settings, often lacks the semantic alignment benefits of contrastively pretrained models.

3.5 Classification

In the proposed model, BERT+CLIP+LR, we employed a logistic regression classifier trained on the aligned image and text embeddings produced by CLIP's encoders. This lightweight classifier proved effective in leveraging CLIP's pretrained semantic representations for binary meme classification, Misogyny vs. Not Misogyny. The final layer outputted probability scores used to determine the predicted class. In the traditional models,Using CNN, BERT, InceptionV3, or ResNet50 for feature extraction, the concatenated multimodal vectors were passed to a dense neural network or a fully connected classification layer.

4 Experiment setup

This section presents the dataset details and describes the experimental configuration used to train and evaluate our models.

4.1 Dataset Description

The dataset developed by (Ponnusamy et al., 2024), provided for the training and development phase, contains the file name of each meme image along with its associated transcribed text. Each meme is annotated with a binary label indicating whether it is misogynistic or non-misogynistic. This structured format allows to develop models that can analyze both visual and textual components of memes. Given the presence of Chinese-language text and complex visual cues, the dataset poses a multilingual and multimodal challenge, encouraging the use of advanced techniques in vision-language understanding for accurate classification. The dataset details are given in Table1.
The dataset is partitioned into 1190 training samples, 170 development/validation samples, and 340 test samples sets, following a roughly 70:10:20 ratio to support robust training, hyperparameter tuning, and final evaluation. The Misogyny class contains 349 training, 47 validation, and 104 test samples, while the Not-Misogyny class includes 841 training, 123 validation, and 236 test samples. The same preprocessing steps are uniformly applied to all three subsets to maintain input consistency during training and evaluation phases.

4.2 Hyperparameters and Model Configuration

In our experiments, we utilized four different models for multimodal meme classification:

Class	Train	Dev	Test
Misogyny	349	47	104
Not-Misogyny	841	123	236
Total	1190	170	340

Table 1: Dataset details

BERT+CLIP+LR, CLIP+LR, BERT+ResNet50, and CNN+InceptionV3, each with distinct hyperparameters.

For CLIP+LR, the image and text features were extracted using CLIP's pretrained Vision Transformer (ViT-B-32) for text and the corresponding image encoder. Both image and text embeddings were normalized using L2 normalization, and logistic regression was employed as the classifier, with a learning rate of 2e-5 for feature extraction and training.

In the BERT+ResNet50 model, BERT was used for text feature extraction, and ResNet50 was employed for image features. Both models were fine-tuned using the Adam optimizer with a learning rate of 2e-5, and the CrossEntropy Loss was used for classification.

In the CNN+InceptionV3 setup, images were processed using a pre-trained InceptionV3 model, without the top classification layer, with the input image size set to (299, 299, 3). The output from the InceptionV3 was passed through a GlobalAveragePooling layer and a fully connected layer with 128 units and ReLU activation. The model was compiled with the Adam optimizer, using a learning rate of 1e-4 and categorical cross-entropy loss for classification.

5 Experimental Evaluation

We evaluated our multimodal architectures for misogyny meme classification using standard metrics, including accuracy, macro precision, macro recall, and macro F1-score.

5.1 Overall Performance

The BERT+CLIP+LR architecture achieved the best overall performance in our experiments. By combining CLIP's powerful pretrained visual encoder with BERT's deep contextual text representations, the model captured cross-modal relationships more effectively than the original CLIP+LR setup or conventional fusion strategies. This approach enhanced textual understanding while retaining the efficiency and semantic alignment benefits of CLIP's visual features. Despite the architectural simplicity—without relying on complex deep fusion layers—BERT+CLIP+LR delivered superior accuracy while remaining computationally efficient and interpretable.

Model	Precision	Recall	Macro F1	Weighted F1	Accuracy
CNN+InceptionV3	0.80	0.81	0.76	0.80	0.81
BERT+ResNet50	0.85	0.84	0.82	0.85	0.84
CLIP+LR	0.86	0.86	0.83	0.86	0.86
BERT+CLIP Image Encoder+LR	**0.89**	**0.89**	**0.87**	**0.89**	**0.89**

Table 2: Evaluation of proposed models for misogynistic meme detection

Model	Total Inference Time (s)	Avg. Time per Sample (s)
CNN + InceptionV3	8.0840	0.0238
CLIP + LR	20.7715	0.0611
BERT + ResNet50	28.2900	0.0832
BERT + CLIP Image Encoder + LR	25.2456	0.0742

Table 3: Inference time comparison of proposed models.

As shown in Figure 3, the BERT+CLIP+LR model correctly classified 214 non-misogynistic and 87 misogynistic samples, with 22 false positives and 17 false negatives. These misclassifications likely stem from class imbalance and limited variability in the training data, which can hinder the model's ability to generalize to ambiguous or nuanced cases.

The early fusion approaches, which combined textual and visual features through concatenation, showed competitive results. Specifically, the combination of BERT for text and ResNET50 for images consistently outperformed CNN-based text and image representations, highlighting the effectiveness of contextual language embeddings in understanding meme text. Summary of classification performance across models are presented in Table2.

We calculated the total inference time and the average inference time per sample for the proposed methods on the test dataset. The results are presented in Table 3. While the CNN + InceptionNet model offers the fastest inference time, CLIP + LR provides a reasonable compromise between performance and inference speed, making it suitable for applications where a balance between accuracy and efficiency is desired.

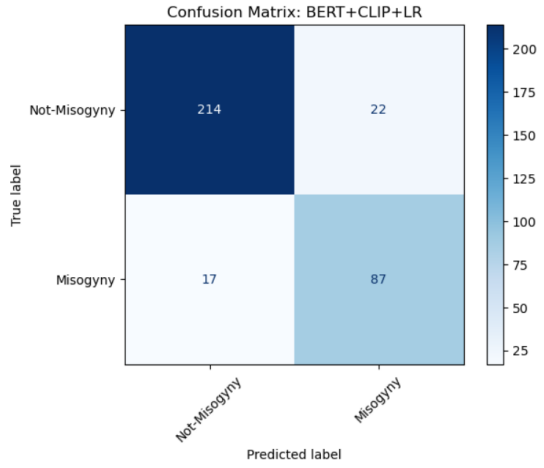

Figure 3: Confusion matrix of multimodal model BERT+CLIP+LR

5.2 Error Analysis

False Positive: Sample278.jpg contains a humorous diagram featuring the Chinese character meaning "woman" at the center of a radar chart, with all attributes of intelligence, courage, perseverance, physical strength, and lifespan. Despite being a positive or humorous depiction celebrating women's qualities, the model incorrectly flagged it as misogynistic. This misclassification likely stems from overfitting to visual or textual stylistic patterns like the presence of gender-related characters or symbols without understanding the broader context or intent. The error highlights the need for improved context-aware classification in multimodal systems.

False Negatives: Sample113.jpg, with text, "Stop talking about that trashy woman" was misclassified as not misogyny despite containing gendered slurs and hostile language toward women. The meme expresses contempt using a derogatory Chinese phrase aimed at women, reinforcing misogynistic sentiment. However, the model likely failed to detect this due to language limitations of non-English text transcription and lack of cultural context, leading to an undetected instance of harmful bias.

6 Conclusion

In this study, we investigated a vision-language multimodal approach for misogynistic meme classification. We began with the CLIP+LR model, where CLIP's contrastive pretraining effectively aligned image and text features without requiring explicit fusion layers, resulting in a model that was both accurate and computationally efficient. Building upon this, we proposed the BERT+CLIP+LR model, which replaces CLIP's text encoder with BERT to capture deeper contextual understanding of language. This enhancement led to improved cross-modal alignment and superior classification performance, while maintaining architectural simplicity. To benchmark our approach, we also implemented traditional early fusion models that combined CNNs and BERT for text with InceptionV3 and ResNet50 for image features. Overall, our findings highlight that Vision-Language Models offer a scalable, robust, and efficient solution for multimodal classification tasks, and represent a promising direction for future research in understanding harmful online content. For future work, we aim to develop misogynistic dataset to include more languages and cultural contexts, explore more advanced vision-language models, and investigate methods to detect implicit and context-dependent misogyny in memes.

References

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

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*, pages 770–778.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. John Wiley & Sons.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Xia Lei, Siqi Wang, Yongkai Fan, and Wenqian Shang. 2024. Hate-udf: Explainable hateful meme detection with uncertainty-aware dynamic fusion. *Software: Practice and Experience*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Alaa Mohasseb, Eslam Amer, Fatima Chiroma, and Alessia Tranchese. 2025. Leveraging advanced nlp techniques and data augmentation to enhance online misogyny detection. *Applied Sciences*, 15(2).

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

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*.

Giulia Rizzi, Paolo Rosso, and Elisabetta Fersini. 2024. From explanation to detection: Multimodal insights into disagreement in misogynous memes.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Advaitha Vetagiri, Prateek Mogha, and Partha Pakray. 2024. Cracking down on digital misogyny with multilate: A multimodal hate detection system. *Working Notes of CLEF*.