CUET-NLP_MP@DravidianLangTech 2025: A Transformer-Based Approach for Bridging Text and Vision in Misogyny Meme Detection in Dravidian Languages

Md. Mohiuddin, Md Minhazul Kabir
Kawsar Ahmed and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1904103, u1904040, u1804017}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Misogyny memes, a form of digital content, reflect societal prejudices by discriminating against women through shaming and stereotyping. In this study, we present a multimodal approach combining Indic-BERT and ViT-base-patch16-224 to address misogyny memes. We explored various Machine Learning, Deep Learning, and Transformer models for unimodal and multimodal classification using provided Tamil and Malayalam meme dataset. Our findings highlight the challenges traditional ML and DL models face in understanding the nuances of Dravidian languages, while emphasizing the importance of transformer models in capturing these complexities. Our multimodal method achieved F1-scores of 77.18% and 84.11% in Tamil and Malayalam, respectively, securing 6^{th} place for both languages among the participants.

1 Introduction

Memes have evolved into a prevalent form of communication in today's digital landscape. These concise pieces of content blend images and text, making them highly engaging to humorously convey complex ideas and emotions. However, the rapid dissemination of memes can also be exploited to spread harmful narratives particularly misogyny. A concerning trend is the frequent sharing of memes that objectify women, promote gender-based violence and propagate harmful stereotypes (Chen et al., 2024a). Further amplifying this issue, nearly 73% of the women surveyed out of over 900 media workers from 125 countries have faced online violence according to a 2022 UNESCO report (Collett et al., 2022). This unchecked rise of misogynistic meme presents a significant threat to social harmony by normalizing misogynistic attitudes and encouraging a toxic culture of gender-based violence. Therefore, effective mitigation strategies are crucial to counter the negative impact of these memes and ensure a more respectful and safer online environment.

Recent research has begun to explore different aspects of memes such as offensive and hate driven content. However, majority of these studies have primarily focused on high resource languages (Singh et al., 2024; Chen et al., 2024b; Fersini et al., 2021) and limited attention is given to low-resource languages like Tamil and Malayalam. In these languages, the detection of misogynistic content in memes remains underexplored, despite significant growth of such content across social media platforms in these regions. The Shared Task on "Misogyny Meme Detection" at DravidianLangTech@NAACL 2025 (Chakravarthi et al., 2024) aims to fill this gap. As part of this shared task, we developed a multimodal system capable of analyzing both textual and visual elements of memes in Tamil and Malayalam to accurately classify them as misogynistic or non-misogynistic. The key contributions of this work are:

- Explored a variety of models (SVM+TF-IDF, CNN, Bi-LSTM, XLM-R, mBERT) for text, (VGG16, VGG19, ResNet50, ViT, Swin) for image, and combinations of these for multimodal analysis to identify an effective method for misogyny meme detection in Tamil and Malayalam.
- Proposed a multimodal model to effectively detect misogynistic memes.

2 Related Work

In recent years, research has focused on improving misogynistic meme detection due to the rise of harmful content online. Addressing the challenge of misogynistic meme detection. Rehman et al. (2025) designed a novel approach combining MANM, GFRM, CFLM to enhance image-text interaction, refine unimodal features and add content

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 592–599

specific elements. This approach outperformed existing methods by 11.87% and 10.82% in F1 score on MAMI and MMHS150K datasets, respectively. Hasan et al. (2024) introduced the Bengali Meme Dataset (AMemD), created to aid in aggression detection in Bengali memes. Among the models tested, the CNN combined with VGG16 achieved the highest F1 score of 0.738. Singh et al. (2024) introduced a dataset of Hindi-English code-mixed misogyny meme detection dataset and investigated different text-only, image-only and multimodal models. For binary classification, their multimodal model, BiT combined with MuRIL BERT, achieved a Macro F1 score of 0.7319. For multilabel classification, their combined BiT and RoBERTa model achieved a Macro F1 score of 0.527. Likewise, Ahsan et al. (2024) also developed a Bengali aggressive meme dataset and utilized the MAF approach which combines the CLIP model for image encoding and Bangla-BERT for text encoding. This method achieved a weighted F1 scroe of 0.742. Zhang and Wang (2022) proposed a multimodal approach using CLIP and UNITER pretrained models, introducing an ensemble method called PBR that achieved top performance in the SemEval-2022 Task 5 with macro F1 scores of 0.834 and 0.731 for sub-task A and B, respectively. By utilizing MMVAE model that integrates text and image embeddings via a VAE for joint multi-task learning. Gu et al. (2022) proposed a method that achieved a macro F1 score of 0.723. Hakimov et al. (2022) presented a multimodal architecture using pretrained CLIP models for both text and image feature extraction, incorporating an LSTM layer for textual context and a fully connected layer for image features. They achieved a macro-F1 score of 0.834. Zhou et al. (2022) tackled the MAMI task at SemEval-2022 using ERNIE-ViL-Large with techniques like biased word masking, image captioning, ensemble learning, and Perspective API, achieving a Macro F1 score of 0.793.

3 Dataset and Task Description

The Shared Task on "Misogyny Meme Detection" (Chakravarthi et al., 2024, 2025) is a multimodal machine learning challenge that aims to classify misogynistic or non-misogynistic memes from social media platforms in Tamil and Malayalam languages. The presented dataset (Ponnusamy et al., 2024) contains images and a CSV file with "image_id", "labels", and "transcriptions" (text), covering both languages. The given dataset is divided into three sets: train, dev and test. The Malayalam dataset is nearly balanced, while the Tamil dataset is imbalanced. The statistics of the dataset are given in Table 1.

Set	Class	Tamil	Malayalam
Train	Misogynistic	285	259
	Non-misogynistic	851	381
Dev	Misogynistic	74	63
	Non-misogynistic	210	97
Test	Misogynistic	89	78
	Non-misogynistic	267	122

Table 1: Dataset distribution for misogyny meme detection

4 System Overview

This section outlines the methodologies and techniques employed to tackle the problem of misogynistic meme detection, encompassing data preprocessing, feature extraction, and the development of a multimodal classification framework integrating textual and visual modalities. The schematic representation of our proposed methodology is illustrated in Figure 1. The complete source code implementation is available on GitHub[1].

4.1 Data Preprocessing

To ensure data consistency and improve model performance, we applied rigorous preprocessing steps tailored to both textual and visual data.

Text Preprocessing: We cleaned the text data by removing emojis, HTML tags, and duplicate entries. Stopwords and punctuation were filtered out, and tokenization was performed using subword-based tokenizers for transformer-based models. Additionally, Term Frequency-Inverse Document Frequency (TF-IDF) (Takenobu, 1994) was employed to extract statistical text features for classical machine learning models.

Image Preprocessing: All images were resized to a uniform dimension and normalized to enhance model stability. Data augmentation techniques, including random flipping, rotation, and contrast adjustment, were applied to improve model generalization.

4.2 Models

To effectively classify memes in Tamil and Malayalam, we employed a combination of traditional

[1] https://github.com/MohiuddinPrantiq/CUET-NLP_MP-DravidianLangTech-NAACL2025

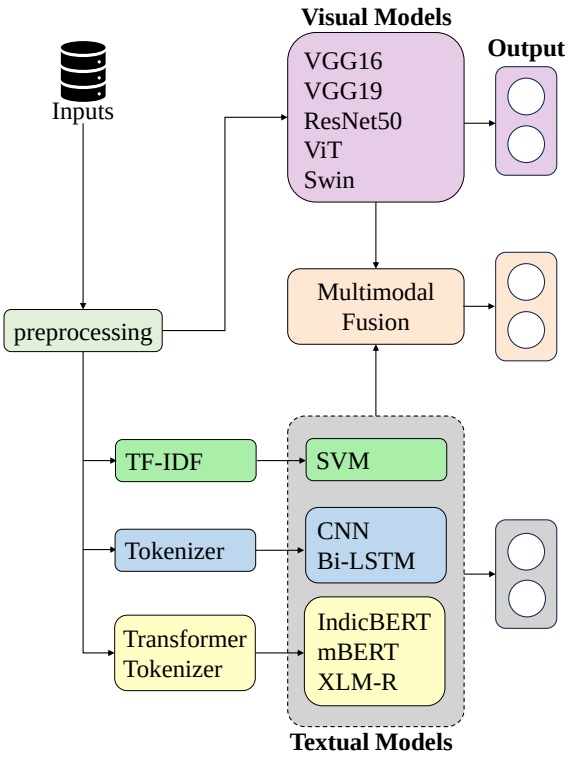

Figure 1: Schematic representation of the proposed system.

machine learning, deep learning, and transformer-based models, leveraging transfer learning to enhance performance without the need for extensive training from scratch (Ibrahim et al., 2020).

4.2.1 Unimodal Models

The provided dataset contains both image and text data, offering an opportunity to investigate individual unimodality models for text and image classification.

Text Classification: Several models were tested for text classification: SVM(Cortes and Vapnik, 1995), CNN(Kim, 2014), Bi-LSTM(Graves and Schmidhuber, 2005), mBERT (Devlin, 2018), and XLM-R (Conneau, 2019) for both Tamil and Malayalam. SVM used a Linear kernel. CNN was implemented with a vocab size of 10,000, an embedding dimension of 100, 100 filters, 0.5 dropout, and filter sizes of 3, 4, and 5. Bi-LSTM used the same vocab size, dropout, and embedding dimension as CNN, with a hidden dimension of 256. Both CNN and Bi-LSTM were trained for 5 epochs with a batch size of 32. mBERT and XLM-R, pretrained transformers from Hugging Face (Wolf, 2020), were trained for 3 epochs, with a batch size of 16 and a learning rate of 2e-5.

Image Classification: For image classification, we used pretrained models: VGG16, VGG19 (Simonyan, 2014), ResNet50 (He et al., 2016), Google's Vision Transformer (Alexey, 2020), and Microsoft's Swin Transformer (Liu et al., 2021) for both Tamil and Malayalam memes. All models were trained for 5 epochs with a learning rate of 1e-4, Adam optimizer, and a batch size of 32.

4.2.2 Multimodal Models

In this task, We need such a system that can analyze both text and images in memes (multimodal) using separate branches for each data type. In our exploration, Different models performed better for each language and modality. For Tamil, SVM and mBERT outperformed other models in text classification, while ResNet50 and Swin performed better in image classification. For Malayalam, SVM and XLM-R were more effective in text, while VGG16 and Swin were superior for images. For multimodal analysis, we selected the top two models from each modality in both languages and concatenated their features to improve prediction accuracy. This resulted in four different combinations for each language. All combinations were trained for 5 epochs with a learning rate of 2e-5.

4.2.3 Proposed Method

Our proposed multimodal fusion model integrates IndicBERT (Kakwani et al., 2020) for textual representation and ViT-Base-Patch16-224 (Wu et al., 2020) for visual feature extraction. The feature embeddings from both modalities are concatenated and passed through a dense classification layer. This fused model was trained for 5 epochs with a batch size of 16 and a learning rate of 2e-5, achieving the best overall performance.

5 Results and Analysis

Table 2 presents a comparative analysis of unimodal and multimodal approaches. In text classification, transformer models outperformed traditional ML and DL models, with mBERT achieving 58.29% in Tamil and XLM-R attaining 73.86% in Malayalam, highlighting their strong contextual understanding. SVM+TF-IDF performed better than CNN and Bi-LSTM, likely due to its efficiency in handling smaller datasets.

For image classification, ResNet50 (68.71%) and VGG16 (80.98%) surpassed vision transformers, suggesting that optimized CNN architectures

Model	Tamil	Malayalam
Text Classification		
SVM+TF-IDF	57.69	65.38
CNN	31.50	50.65
Bi-LSTM	44.84	58.82
XLM-R	23.53	**73.86**
mBERT	**58.29**	56.49
Image Classification		
VGG16	64.05	**80.98**
VGG19	21.57	77.94
ResNet50	**68.71**	75.95
ViT	66.24	73.28
Swin	67.88	80.82
Multimodal Classification		
Tamil		
SVM+ResNet50	45.53	-
SVM+Swin	41.92	-
mBERT+ResNet50	**67.76**	-
mBERT+Swin	64.10	-
Malayalam		
XLM-R+VGG16	-	79.14
XLM-R+Swin	-	**82.58**
SVM+VGG16	-	68.67
SVM+Swin	-	72.73
Proposed Model		
IndicBERT+ViT	**77.18**	**84.11**

Table 2: Comparison of Macro F1-scores across different models on the test set.

remain competitive. However, the close performance of Swin and ViT indicates vision transformers' potential when trained on larger datasets.

In multimodal classification, fusing textual and visual features improved performance. mBERT+ResNet50 reached 67.76% in Tamil, while XLM-R+Swin achieved 82.58% in Malayalam. Our proposed IndicBERT+ViT model obtained the highest F1-scores: 77.18% in Tamil and 84.11% in Malayalam, demonstrating the advantages of multimodal learning. The performance gap between Tamil and Malayalam models suggests that dataset characteristics, including class balance and linguistic complexity, play a key role.

Several key insights emerge from these results. First, text classification performance was lower in Tamil, likely due to class imbalance and the complexity of Tamil script variations. Second, multimodal models consistently outperformed unimodal models, reaffirming the importance of leveraging complementary information sources. However, the lower performance of mBERT+ResNet50 in Tamil compared to ResNet50 alone suggests that feature fusion strategies need refinement. Future work could explore advanced fusion techniques, such as attention-based feature alignment or adaptive weighting, to enhance multimodal learning further. Additionally, dataset balancing techniques may help mitigate bias and improve model generalization, particularly in underrepresented classes.

5.1 Error Analysis

To gain deeper insights into the performance of our proposed model, we conducted a comprehensive error analysis encompassing both quantitative and qualitative evaluations.

Quantitative Analysis

Figure 2 presents the confusion matrices for Tamil and Malayalam meme classification.

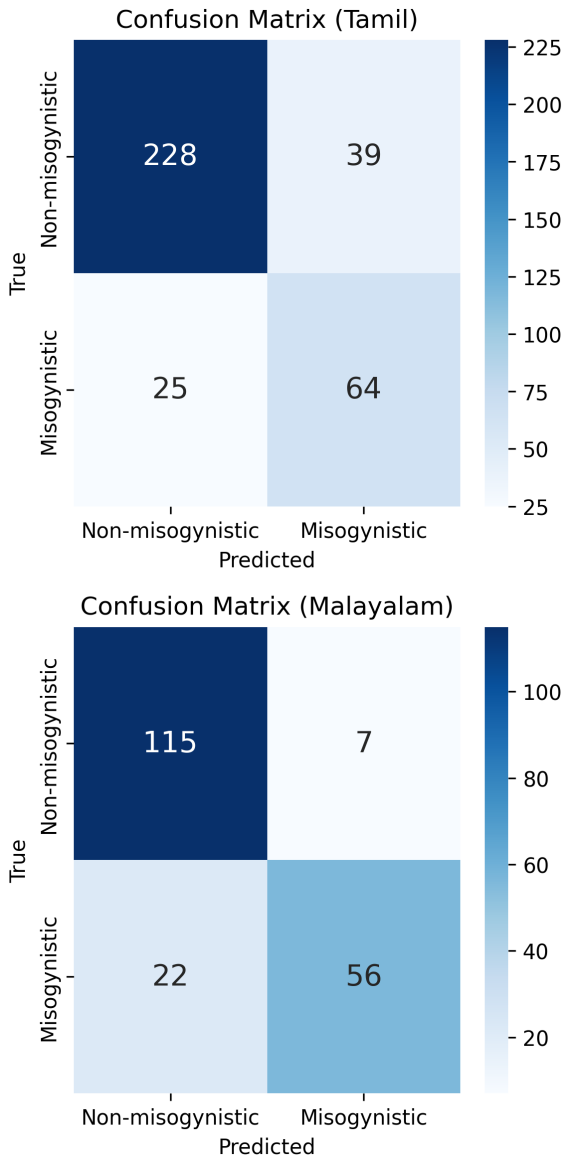

Figure 2: Confusion matrices for Tamil and Malayalam

The model achieved 82.02% accuracy in Tamil (292/356) and 85.5% in Malayalam (171/200), with higher misclassification rates for the positive class (28% vs. 14.5% in Tamil, 5.7% in Malayalam). This disparity stems from dataset imbalance, which limited the model's learning capacity. Notably, the model performed 6.93% better in Malayalam, likely due to a more balanced dataset.

Qualitative Analysis

Figures 3 and 5 highlight correct and incorrect predictions for Tamil, while figures 4 and 6 do the same for Malayalam. While the model accurately classified misogynistic memes in some cases, errors in others suggest an inability to capture nuanced distinctions, primarily due to dataset imbalance. The limited dataset size may have also restricted the model's ability to learn diverse representations, leading to misclassification in complex cases. Furthermore, the simple concatenation-based feature fusion strategy may not have effectively captured intricate relationships between textual and visual modalities, causing the model to overlook subtle cues that differentiate misogynistic from non-misogynistic content. Please see Appendix A for the visualization of the sample memes.

Text Sample	Actual	Predicted
ID_335: causally ignoring when someone complaining about me _vishal_sammy_ EVEL MEWES = = VERA LEVEL	Non-misogynistic	Non-misogynistic
ID_1576: Mano PhotoGrid IRAL Bsru பாத்துக்கிட்டே ருக்கல போல் (Mano is on PhotoGrid IRAL, just watching without participating.)	Non-misogynistic	Misogynistic
ID_1149: OUI MO $ மருமகள் மாமியார் த னக்கு டீ போட தெரியுமா $ 0 டீ போட தெரியுமாவா குக்கர்ல பால ஊத்தி மூணு விசில் விட்டு அப்டியே பொங்கவிடுவோம் (Oh! Does the daughter-in-law know how to make tea for the mother-in-law? Does she know how to make tea or not? Just pour milk into the cooker, give three whistles, and let it boil over like that.)	Misogynistic	Non-misogynistic
ID_1064: எனக்கொரு சினேகிதி சினேகிதி. தென்றல் மாதிரியா..? ல்ல.மெண்டல் மாதிரி. rr (I have a friend, a friend. Like a breeze? No, like a mental one. Haha!)	Misogynistic	Misogynistic

Figure 3: Sample predictions in Tamil.

Text Sample	Actual	Predicted
ID_945: ഇതിന് മാത്രം പിള്ളാരെ ഇവർക്കെങ്ങനെ cilşma??.... (How can these people be blamed for this alone??....)	Non-misogynistic	Non-misogynistic
ID_543: കളിതുടങ്ങി രണ്ടുമിനുട്ട് കഴിഞ്ഞപ്പോൾ ഭർത്താ ഹി ഹി മോളൂ ഭാര്യ പി പി സുനേഷ് എന്നാ വൻമരം വീണു അല്ലെ (Two minutes after the game started, husband: Ha.. ha.. dear... wife: P.P. Suresh, isn't it like a big tree fell?)	Non-misogynistic	Misogynistic
ID_545: അയലത്തെ ചേച്ചിയോട് കളി ചോദിച്ചതിന് ശേഷം നായകൻ ചേച്ചി ഒന്നും പറഞ്ഞില്ല ചേച്ചി കിടക്കു വാ ഞാൻ പോയി അടച്ചിട്ട് വരാം (After asking the neighbor's sister for a game, the hero said, 'Sister didn't say anything... Sister is resting. I'll go, close the door, and come back.)	Misogynistic	Non-misogynistic
ID_61: എന്തൊക്കെ ആകിയിട്ടും ഒരു മെന വരുന്നില്ലലോ Nikhi.. (Whatever happens, isn't there any sign of a message, Nikhi)	Misogynistic	Misogynistic

Figure 4: Sample predictions in Malayalam.

6 Conclusion

This study examined the performance of unimodal and multimodal approaches using various machine learning, deep learning, and transformer models for misogynistic meme detection in Tamil and Malayalam. The results show that the combination of Indic-BERT and ViT achieved the highest F1-scores of 77.18% and 84.11% in Tamil and Malayalam, respectively. These findings highlight the effectiveness of transformer models in both unimodal and multimodal classification for this task. To address the current constraints, future work can explore improved fine-tuning strategies, leverage large language models (LLMs) and alternative transformer architectures, and adopt more sophisticated feature fusion techniques such as attention mechanisms and cross-modal transformers. Additionally, integrating Vision-Language Models (VLMs) such as CLIP (Radford et al., 2021) and LXMERT (Tan and Bansal, 2019) could enhance multimodal understanding by jointly learning from textual and visual features in a more coherent manner. Further efforts in handling data imbalance—including data augmentation, reweighting strategies, SMOTE, class-weighted loss, and adversarial training along with expanding dataset size and diversity or incorporating a CNN-based backbone with transfer learning could also improve classification accuracy and robustness.

Limitations

Despite the strong performance of our proposed approach, several limitations remain. One key challenge is misclassification, likely influenced by dataset imbalance, which affects generalization. Additionally, our use of IndicBERT, which supports 12 Indian languages, may not extend effectively to languages outside its training scope. Another constraint arises from our straightforward feature fusion strategy, which relies on simple concatenation and may not fully capture the complex interactions between textual and visual modalities.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

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.

Dosovitskiy Alexey. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025. Findings of the Shared Task on Misogyny Meme Detection: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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. 2024a. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1864–1871.

Shijing Chen, Usman Naseem, Imran Razzak, and Flora D. Salim. 2024b. Unveiling misogyny memes: A multimodal analysis of modality effects on identification. *Companion Proceedings of the ACM on Web Conference 2024*.

Clementine Collett, Livia Gouvea Gomes, Gina Neff, et al. 2022. *The effects of AI on the working lives of women*. UNESCO Publishing.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Giuliano Rizzi, Aurora Saibene, and Francesca Gasparini. 2021. Misogynous meme recognition: A preliminary study. In *International Conference of the Italian Association for Artificial Intelligence*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Yimeng Gu, Ignacio Castro, and Gareth Tyson. 2022. Mmvae at semeval-2022 task 5: A multi-modal multi-task vae on misogynous meme detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 700–710.

Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. 2022. Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. *arXiv preprint arXiv:2204.06299*.

Md Hasan, Shawly Ahsan, Moshiul Hoque, and M. Dewan. 2024. *MuLAD: Multimodal Aggression Detection from Social Media Memes Exploiting Visual and Textual Features*, pages 107–123.

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.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2020. AlexU-BackTranslation-TL at SemEval-2020 task 12: Improving offensive language detection using data augmentation and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1881–1890, Barcelona (online). International Committee for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. 2025. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing & Management*, 62(1):103895.

Karen Simonyan. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Aakash Singh, Deepawali Sharma, and Vivek 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*.

Tokunaga Takenobu. 1994. Text categorization based on weighted inverse document frequency. *Information Processing Society of Japan, SIGNL*, 94(100):33–40.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. 2020. Visual transformers: Token-based image representation and processing for computer vision. *Preprint*, arXiv:2006.03677.

Jing Zhang and Yujin Wang. 2022. SRCB at SemEval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596, Seattle, United States. Association for Computational Linguistics.

Ziming Zhou, Han Zhao, Jingjing Dong, Ning Ding, Xiaolong Liu, and Kangli Zhang. 2022. DD-TIG at SemEval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 563–570, Seattle, United States. Association for Computational Linguistics.

A Appendix

True: Non-misogynistic | Image ID: 335 | Predicted: Non-misogynistic

True: Non-misogynistic | Image ID: 954 | Predicted: Non-misogynistic

True: Non-misogynistic | Image ID: 1576 | Predicted: Misogynistic

True: Non-misogynistic | Image ID: 543 | Predicted: Misogynistic

True: Misogynistic | Image ID: 1149 | Predicted: Non-misogynistic

True: Misogynistic | Image ID: 545 | Predicted: Non-misogynistic

True: Misogynistic | Image ID: 1064 | Predicted: Misogynistic

True: Misogynistic | Image ID: 61 | Predicted: Misogynistic

Figure 5: Sample memes in Tamil

Figure 6: Sample memes in Malayalam