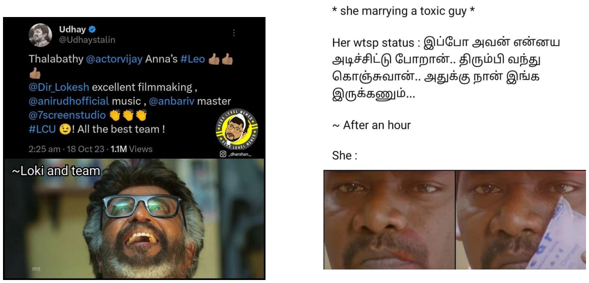
CUET_NetworkSociety@DravidianLangTech 2025: A Multimodal Framework to Detect Misogyny Meme in Dravidian Languages

MD Musa Kalimullah Ratul*, Sabik Aftahee*, Tofayel Ahmmed Babu*
Jawad Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904071, u1904024, u1904005, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

Memes are widely used to communicate on social networks. Some memes can be misogynistic, conveying harmful messages to individuals. Detecting misogynistic memes has become a broader challenge, particularly for low-resource languages like Tamil and Malayalam, due to their linguistic morphology. This paper presents a multi-modal deep learning approach for detecting misogynistic memes in Tamil and Malayalam. The proposed model integrates fine-tuned ResNet18 for extracting visual features and `ai4bharat/indic-bert` for analyzing textual content. The fusion model was then applied to make predictions on the test dataset. The model achieved a macro F1 score of 76.32% for Tamil and 80.35% for Malayalam. Our technique helped secure 7th and 12th positions for Tamil and Malayalam, respectively.


Udhay
@Udhaystalin
Thalabathy @actorvijay Anna's #Leo 👍👍👍
@Dir_Lokesh excellent filmmaking ,
@anirudhofficial music , @anbariv master
@7screenstudio 👏👏👏
#LCU 😉! All the best team !
2:25 am · 18 Oct 23 · 1.1M Views
VERA LEVEL MEMES
dharshan.
~Loki and team
mt
* she marrying a toxic guy *
Her wtsp status : இப்போ அவன் என்னய அடிச்சிட்டு போறான்.. திரும்பி வந்து கொஞ்சுவான்.. அதுக்கு நான் இங்க இருக்கணும்...
~ After an hour
She :

A misogynistic meme

A non-misogynistic meme

Figure 1: Examples of memes: (a) Misogynistic and (b) Non-misogynistic.

1 Introduction

In the past few years, the popularity of the social media platform has gained a huge response from individuals, where a multi-modal phenomenon called meme has been introduced to us. The meme is generally an image with some contextual texts of that image. Apart from humorous contents, memes also can carry harmful messages, such as misogyny, which is the hatred of, contempt for, or prejudice against women or girls. The satirical nature of such content makes identifying misogynistic memes particularly challenging, as they often employ nuanced, context-dependent signals that can evade straightforward detection (Rizzi et al., 2023). To illustrate this distinction, Figure 1 compares a misogynistic meme and a non-misogynistic meme, highlighting the subtle yet significant differences in their messaging.

Misogynistic memes pose a threat to society that basically try to normalize hatred against women and gender bias. Existing research has explored multimodal approaches to identifying misogyny contents in memes with both textual and visual features to achieve higher accuracy. For instance, Rizzi et al. (Rizzi et al., 2023) investigated 4 unimodal and 3 multimodal approaches to determine which source of information contributes more to the detection of misogynous memes. Similarly, Singh et al. (Singh et al., 2024) investigated misogyny detection in Hindi-English code-mixed memes, showcasing the complexities of handling low-resource languages and multimodal content. The shared task on multitask meme classification further emphasized the importance of distinguishing misogynistic and trolling behaviors, providing valuable datasets and benchmarks for advancing the field (Chakravarthi et al., 2024).

Even with this much advancement, research on misogynistic memes in low-resource languages such as Tamil and Malayalam remains very much challenging. Our paper introduces a multi-modal deep learning framework for detecting misogynistic memes in Tamil and Malayalam.
The key contributions of this work include:

- Proposed a multimodal framework designed to detect misogynistic memes while capturing the linguistic and cultural nuances of Tamil and Malayalam.

- Investigated the performance of several ML, DL, and transformer-based models for misogynistic meme detection, highlighting challenges in low-resource languages through quantitative and qualitative error analysis.

2 Related Work

Multimodal approaches, which leverage the interplay between text and images, have shown promising advancements. Gasparini et al. (Gasparini et al., 2022) tackled the automatic detection of misogynistic con-

[*] Authors contributed equally to this work.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 607–613
May 3, 2025 ©2025 Association for Computational Linguistics

tent in memes using multimodal data, evaluating a dataset of 800 memes (400 misogynistic and 400 non-misogynistic). Singh et al. (Singh et al., 2024) introduced *MIMIC*, a dataset containing 5,054 Hindi-English code-mixed memes for misogyny detection, demonstrating the effectiveness of multimodal fusion, where ViT+RoBERTa achieved a macro f1 score of 0.7532 for Multi-label Misogyny Classification. Rizzi et al. (Rizzi et al., 2023) addressed biases in misogynistic meme detection by proposing a debiasing framework on an 800-meme dataset, achieving up to a 61.43% improvement in prediction probabilities. Here Visual-BERT achieved a macro f1 score of 0.84. Chakravarthi et al. (Chakravarthi et al., 2024) organized a shared task on multitask meme classification for misogyny and trolling in Tamil and Malayalam memes, achieving macro F1 scores of 0.73 for Tamil and 0.87 for Malayalam. Ponnusamy et al. (Ponnusamy et al., 2024) introduced the *MDMD* dataset for misogyny detection in Tamil and Malayalam memes. Hegde et al. (Hegde et al., 2021) focused on Tamil troll meme classification, demonstrating the effectiveness of attention-based transformers. They achieve an overall F1-score of 0.96 by using images for classification using ViT.

Hossain et al. (Hossain et al., 2024) proposed an *Align before Attend* strategy for multimodal hateful content detection on the MUTE (Bengali code-mixed) and MultiOFF (English) datasets, achieving F1-scores of 69.7% and 70.3%, respectively. Ahsan et al. (Ahsan et al., 2024) developed the *MIMOSA* dataset for target-aware aggression detection in Bengali memes, introducing a multimodal aggression fusion (MAF) model that outperformed state-of-the-art approaches. Here the ViT achieved the highest weighted F1-score of 0.582 and for the textual-only Bangla-BERT model surpassed all unimodal models with a weighted F1-score of 0.641. Rahman et al. (Rahman et al., 2024) also proposed a comprehensive multimodal approach for abusive language detection in Tamil, incorporating textual, acoustic, and visual features. The study utilized a dataset annotated for abusive and non-abusive classes, employing models such as ConvLSTM, 3D-CNN, and BiLSTM. Their weighted late fusion model, ConvLSTM+BiLSTM+MNB, achieved the highest macro F1 score of 71.43%. Conneau et al. (Conneau et al., 2020) developed XLM-R, a cross-lingual representation learning model trained on CommonCrawl data in 100 languages, achieving 85.0% accuracy on XNLI. Feng et al. (Feng et al., 2022) introduced LaBSE, a language-agnostic BERT model for multilingual sentence embeddings, enhancing cross-lingual understanding. Here LaBSE achieved a highest accuracy of 95.3%

Tan and Le (Tan and Le, 2020) proposed EfficientNet for image recognition, achieving 84.4% top-1 accuracy on ImageNet, making it a strong candidate for feature extraction in multimodal tasks. He et al. (He et al., 2016) introduced deep residual learning with ResNet, achieving a 3.57% top-1 error rate on ImageNet, widely adopted for image feature extraction in multimodal studies. Zhou et al. (Zhou et al., 2015) proposed C-LSTM for text classification, leveraging CNN-LSTM hybrid models on various text classification datasets, though specific accuracy results were not mentioned. Their implementation C-LSTM got the highest accuracy of 94.6%. Arevalo et al. (Arevalo et al., 2017) proposed gated multimodal units for information fusion, focusing on multimodal data representation. Here GMU achieved a macro-f1 score of 0.541.

These works collectively underscore the potential of multimodal strategies in detecting misogyny and other forms of online abuse. They also highlight the challenges, including handling implicit content, managing noisy or low-resource data, and ensuring model fairness and generalizability across diverse cultural contexts.

3 Task and Dataset Description

The task focuses on developing models for detecting misogynistic content in memes from social media. These models analyze both textual and visual components to classify memes as *Misogynistic* or *Non-Misogynistic*. Misogynistic content includes text or visuals targeting women with harmful, offensive, or derogatory intent, while non-misogynistic memes align with respectful communication standards. This task is part of the DravidianLangTech@NAACL 2025 Shared Task on Misogyny Meme Detection (Chakravarthi et al., 2025a,b). Additionally, prior shared tasks have explored related issues, such as the LT-EDI@EACL 2024 shared task on multitask meme classification, which examined misogynistic and troll memes in Tamil and Malayalam (Chakravarthi et al., 2024). The dataset consists of Tamil and Malayalam memes, including text extracted from images, corresponding visual data, and labels. This dataset is based on the work by (Ponnusamy et al., 2024), who introduced the MDMD (Misogyny Detection Meme Dataset) to address the propagation of misogyny, gender-based bias, and harmful stereotypes in online memes. Tables 1 and 2 show the dataset distribution.

Dataset	Misogynistic	Non-Misogynistic	Total
Train	259	381	640
Dev	63	97	160
Test	78	122	200

Table 1: Dataset distribution for Malayalam memes

Dataset	Misogynistic	Non-Misogynistic	Total
Train	285	851	1136
Dev	74	210	284
Test	89	267	356

Table 2: Dataset distribution for Tamil memes

The Malayalam dataset contains 640 training, 160 development, and 200 test samples (Table 1). Similarly, the Tamil dataset comprises 1136 training, 284 development, and 356 test samples (Table 2). Both datasets include text extracted from memes, visual data, and classification labels, supporting multimodal approaches for effective classification. The GitHub link[1] contains both the source code and datasets.

4 Methodology

The proposed approach for misogyny meme detection leverages Multimodal deep learning by integrating textual and visual information extracted from memes. The process extracts of image features using Convolutional neural networks (CNNs) and the transformation of text data using transformer-based language models. The classification models are trained independently for each modality before employing a fusion strategy to improve the overall classification performance. Figure 2 presents the schematic framework for the Multimodal misogyny meme detection system.

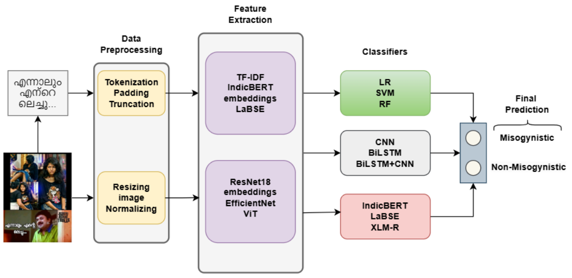


Figure 2: Schematic framework for Misogyny Meme Detection

4.1 Visual Approach

For the image modality, feature extraction was performed using two state-of-the-art pre-trained convolutional neural networks (CNNs): ResNet18 (He et al., 2016) and EfficientNet-B4 (Tan and Le, 2020). These architectures were selected due to their high performance in visual recognition tasks and their ability to capture hierarchical spatial features. Each meme image was resized to a standard resolution of 224 × 224 pixels and normalized to maintain consistent pixel value distributions. Random horizontal flipping, rotation, and brightness adjustments were applied to improve model generalization. The extracted deep features from the CNN architectures were passed through fully connected layers before being used for classification. The ResNet18 model, with its residual connections, helps mitigate vanishing gradient issues and captures robust spatial information. On the other hand, EfficientNet-B4 leverages compound scaling to balance network depth, width, and resolution, optimizing performance while maintaining efficiency. These extracted features were further combined in multimodal fusion strategies for improved classification performance.

4.2 Textual Approach

The textual component of the meme data was processed using advanced transformer-based models, which have demonstrated state-of-the-art performance in various natural language processing tasks (Vaswani et al., 2017). The textual features were extracted using IndicBERT (Kunchukuttan et al., 2020), LaBSE (Feng et al., 2022), and XLM-RoBERTa (Conneau et al., 2020). IndicBERT, specifically designed for low-resource Indian languages, was employed for its ability to handle complex linguistic variations in Tamil and Malayalam. LaBSE, a multilingual model, captured robust sentence-level embeddings, making it suitable for cross-lingual tasks. XLM-RoBERTa, a cross-lingual transformer model, provided strong contextual embeddings for diverse languages, including Tamil and Malayalam. The preprocessing pipeline included tokenization, punctuation removal, and stop-word filtering. The extracted embeddings from these models were passed to classification layers where classical machine learning classifiers such as Logistic Regression, Support Vector Machines (SVM), and Random Forests were explored.

4.3 DL Approach

Beyond traditional classifiers, deep learning architectures were employed to enhance textual classification. CNN, BiLSTM, and a hybrid BiLSTM+CNN model were tested for their ability to capture and model complex textual patterns. The CNN model captured local semantic patterns using convolutional filters (Kim, 2014), while BiLSTM processed sequential dependencies in textual data using bidirectional long short-term memory networks. For the BiLSTM model, contextual embeddings extracted from transformer-based models IndicBERT, LaBSE, and XLM-RoBERTa were input representations. These embeddings provided rich semantic text representations, enabling the BiLSTM to model temporal dependencies while effectively preserving contextual meaning. The hybrid BiLSTM+CNN model combined the local feature extraction capability of CNNs with the sequential modeling power of BiLSTMs, further enhancing feature extraction for textual data (Zhou et al., 2015). All deep learning models were trained with categorical cross-entropy loss and optimized using the Adam optimizer with a learning rate of 0.001 and a batch size of 32.

4.4 Multimodal Fusion Approach

To integrate textual and visual features, two fusion techniques were explored. In the late fusion approach, independent models for text and image classification were trained separately, and their predictions were combined using weighted averaging. Various weight allocations were tested, with uniform weighting yielding the best results. The concatenation-based fusion approach merged the text and image features at an intermediate layer before the final classification. This allowed the model to learn joint representations of multimodal data, enhanc-

[1] https://github.com/MusaRatul/Misogyny-Meme-Detection-in-Dravidian-Languages

ing discriminatory power. Late fusion was preferred over early fusion to preserve modality-specific feature representations, prevent interference between text and image embeddings, and leverage the strengths of pretrained models. The findings align with previous studies, such as Arevalo et al. (Arevalo et al., 2017) and Kiela et al. (Kiela et al., 2020), which highlight the benefits of combining image and text modalities for robust multimodal learning. The proposed multimodal approach demonstrated significant improvements in misogyny meme detection, leveraging both the hierarchical spatial representations from CNN models and the contextual embeddings from transformer-based language models.

5 Results Analysis

The classification performance is summarized in Tables 3 and 4.

Model	P	R	F1
LR	0.8000	0.7000	0.7250
SVM	0.7400	0.6900	0.7050
RF	**0.8550**	0.6150	0.6300
CNN	0.3750	0.5000	0.4286
BiLSTM	0.7837	0.7416	0.7582
BiLSTM+CNN	0.7415	0.7397	0.7406
LaBSE+EfficientNet-B4	0.7687	0.7360	0.7494
ViT+XLM-R	0.8240	0.7453	0.7620
IndicBERT+ResNet18	0.8095	**0.7772**	**0.7632**

Table 3: Performance Comparison on Tamil Dataset

Model	P	R	F1
LR	0.7276	0.7314	0.7292
SVM	0.7373	0.7373	0.7373
RF	0.6843	0.6686	0.6725
CNN	0.3050	0.5000	0.3789
BiLSTM	0.8077	0.7998	0.8032
BiLSTM+CNN	0.8030	0.7934	0.7973
LaBSE+EfficientNet-B4	0.7844	0.7857	0.7851
ViT+XLM-R	0.8115	0.8085	0.7999
IndicBERT+ResNet18	**0.8329**	**0.8162**	**0.8035**

Table 4: Performance Comparison on Malayalam Dataset

For the Tamil dataset (Table 3), the IndicBERT+ResNet18 model achieved the highest performance, with a Precision (P) of 0.8095, Recall (R) of 0.7772, and an F1-score of 0.7632. Among other models, ViT+XLM-R also performed well with an F1-score of 0.7620, followed by BiLSTM (0.7582) and LaBSE+EfficientNet-B4 (0.7494). Traditional machine learning models such as Logistic Regression (LR) and Support Vector Machines (SVM) lagged behind, with F1-scores of 0.7250 and 0.7050, respectively. CNN exhibited the lowest performance, with an F1-score of 0.4286, highlighting its limitations in capturing complex text-visual relationships. The superior performance of IndicBERT+ResNet18 suggests that ResNet18 extracted spatially rich visual features that complemented the textual representations from IndicBERT better than ViT, which may not have captured fine-grained spatial details as effectively.

For the Malayalam dataset (Table 4), IndicBERT+ResNet18 again outperformed all models, achieving a Precision of 0.8329, Recall of 0.8162, and an F1-score of 0.8035. Close contenders included BiLSTM (0.8032), ViT+XLM-R (0.7999), and BiLSTM+CNN (0.7973), showing that deep learning models performed consistently well. Traditional machine learning approaches like LR and SVM had moderate performance, with F1-scores of 0.7292 and 0.7373, respectively. CNN showed the weakest performance, with an F1-score of 0.3789, reinforcing its inefficacy in handling the multimodal nature of the task. This outcome suggests that ResNet18 provided more structured and discriminative visual embeddings than ViT in this context, leading to a more potent synergy with IndicBERT for multimodal learning.

6 Error Analysis

Both quantitative and qualitative error analyses were conducted to gain deeper insights into the performance of the proposed model.

6.1 Quantitative Analysis

To further understand the performance of the models, a quantitative analysis was performed using confusion matrices for Tamil and Malayalam datasets. Figures 3 and 4 illustrate the confusion matrices for both languages.

The confusion matrices reveal that the model performs well in identifying explicit misogynistic content, correctly classifying a high proportion of such memes. However, there are notable errors in detecting non-misogynistic memes and implicit misogyny, leading to false positives and false negatives. This indicates that while the model effectively captures explicit cues, it struggles with subtler textual or contextual features.

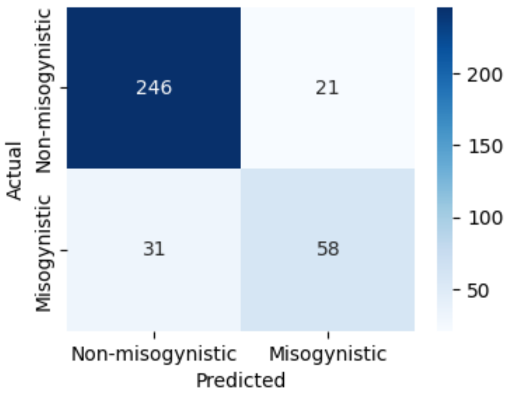


Figure 3: Confusion matrix for the top-performing model (IndicBERT+ResNet18) in Tamil misogynistic meme detection

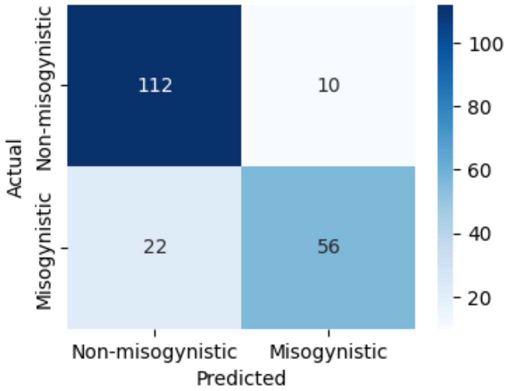


Figure 4: Confusion matrix for the top-performing model (IndicBERT+ResNet18) in Malayalam misogynistic meme detection

6.2 Qualitative Analysis

To complement the quantitative analysis, a qualitative examination of the misclassified examples was conducted. Tables 5 and 6 present representative examples of predicted outputs by the IndicBERT+ResNet18 model for Tamil and Malayalam datasets.

Image	Actual	Predicted
	Misogyny	Misogyny
	Misogyny	Non-Misogyny

Table 5: Examples of predicted outputs from the IndicBERT+ResNet18 model for Tamil misogynistic meme detection

The qualitative analysis reveals that several misclassifications occurred due to linguistic subtleties, sarcasm, and neutral visual elements. Some Tamil memes contained subtle expressions or indirect language that the model failed to interpret correctly. Malayalam memes often employed sarcasm or implicit misogynistic references, making it challenging for the model to capture the intended meaning. In cases where visual elements were neutral or did not reinforce textual cues, the model struggled to combine features effectively, leading to misclassification. These findings suggest that future work should focus on improving the fusion strategy and incorporating external knowledge to better capture contextual and implicit cues, thereby reducing such misclassifications.

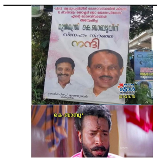
Image	Actual	Predicted
	Non-Misogyny	Misogyny
	Non-Misogyny	Non-Misogyny

Table 6: Examples of predicted outputs from the IndicBERT+ResNet18 model for Malayalam misogynistic meme detection

7 Conclusion

The proposed multimodal approach effectively combines textual and visual features for misogyny detection in Tamil and Malayalam memes, achieving competitive results. By integrating TF-IDF, IndicBERT embeddings, ResNet18, and EfficientNet with machine learning, deep learning, and transformer-based models, this paper demonstrates the potential of multimodal fusion in tackling complex classification tasks. Future work will focus on improving sarcasm detection, as sarcasm often overlaps with misogynistic content and remains challenging to identify accurately. Additionally, there is significant potential to enhance performance by leveraging larger multilingual datasets that include more diverse and representative samples across different languages and cultural contexts.

Limitations

- The model struggles with detecting implicit cues in meme text and visuals, leading to occasional misclassification of sarcastic misogynistic content.
- IndicBERT's limitations affect performance for Tamil and Malayalam, particularly in handling nuanced language structures, impacting overall classification accuracy.
- Memes with complex multimodal sarcasm require improved fusion strategies, necessitating future research on enhanced dataset augmentation and fine-tuned multilingual transformers for better linguistic and contextual understanding.

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, St. Julian's, Malta. Association for Computational Linguistics.

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. *Preprint*, arXiv:1702.01992.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025a. 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, Rahul Ponnusamy, Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Bhuvaneswari Sivagnanam, and Anshid Kizhakkeparambil. 2025b. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Siddhanth U Hegde, Adeep Hande, Ruba Priyadharshini, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021. Uvce-iiitt@dravidianlangtech-eacl2021: Tamil troll meme classification: You need to pay more attention. *Preprint*, arXiv:2104.09081.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. Align before attend: Aligning visual and textual features for multimodal hateful content detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 162–174, St. Julian's, Malta. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020. Supervised multimodal bitransformers for classifying images and text. *Preprint*, arXiv:1909.02950.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N. C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *Preprint*, arXiv:2005.00085.

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. Rahman, Abu Raihan, Tanzim Rahman, Shawly Ahsan, Jawad Hossain, Avishek Das, and Mohammed Moshiul Hoque. 2024. Binary_Beasts@DravidianLangTech-EACL 2024: Multimodal abusive language detection in Tamil based on integrated approach of machine learning and deep learning techniques. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 212–217, St. Julian's, Malta. Association for Computational Linguistics.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing Management*, 60(5):103474.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. Mimic: Misogyny identification in multimodal internet content in hindi-english code-mixed language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks. *Preprint*, arXiv:1905.11946.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A c-lstm neural network for text classification. *Preprint*, arXiv:1511.08630.