Overview of Shared Task on Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes

[1]Bharathi Raja Chakravarthi, [2]Saranya Rajiakodi, [3]Rahul Ponnusamy, [2]Kathiravan Pannerselvam, [4]Anand Kumar Madasamy, [5]Ramachandran Rajalakshmi, [4]Hariharan RamakrishnaIyer LekshmiAmmal, [6]Anshid Kizhakkeparambil, [7]Susminu S Kumar, [2]Bhuvaneswari Sivagnanam, [7]Charmathi Rajkumar

[1]School of Computer Science, University of Galway, Ireland
[2]Department of Computer Science, Central University of Tamil Nadu, India
[3]Data Science Institute, University of Galway, Ireland [4]NIT Karnataka, India
[5]VIT Chennai, India [6]WMO Imam Gazzali Arts and Science College, Kerala, India
[7]The American College, Madurai, Tamil Nadu, India

Abstract

This paper offers a detailed overview of the first shared task on "Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes," organized as part of the LT-EDI@EACL 2024 conference. The task was set to classify misogynistic content and troll memes within online platforms, focusing specifically on memes in Tamil and Malayalam languages. A total of 52 teams registered for the competition, with four submitting systems for the Tamil meme classification task and three for the Malayalam task. The outcomes of this shared task are significant, providing insights into the current state of misogynistic content in digital memes and highlighting the effectiveness of various computational approaches in identifying such detrimental content. The top-performing model got a macro F1 score of 0.73 in Tamil and 0.87 in Malayalam.

1 Introduction

In the ever-changing landscape of online communication (Lin et al., 2024; Priyadharshini et al., 2022), memes have emerged as a remarkable phenomenon, transcending linguistic, cultural, and geographical boundaries (Ford et al., 2023). Their ability to succinctly and often humorously convey complex ideas and emotions has made memes an integral part of digital discourse (Kostadinovska-Stojchevska and Shalevska, 2018; Priyadharshini et al., 2023). However, this rise in meme culture has also revealed the obscene side of online content, which features misogynistic stories and trolling (Rasheed et al., 2020; Suryawanshi and Chakravarthi, 2021). We initiated the "Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes" competition to understand and address these critical issues through memes. This pioneering endeavor leverages gold-standard datasets to illuminate the intricate world of online memes. Our competition aims to inspire the development of cutting-edge models for meme classification, primarily focusing on detecting misogyny and trolling in various languages. Further, it is centered around carefully selected high-quality datasets. These datasets have been precisely annotated to establish a standard for classifying memes. These datasets represent various languages and meme categories, offering a comprehensive view of the meme landscape. Let us delve into the two core tasks our competition addresses:

Task 1: Detecting Misogynistic Memes: This task revolves around identifying misogynistic memes, which perpetuate harmful stereotypes and attitudes towards women. The gold standard dataset for this task spans languages like Tamil and Malayalam, reflecting the global reach of this issue. Participants must develop models capable of analyzing textual and visual elements within memes to distinguish between misogynistic and non-misogynistic content.

Task 2: Troll Meme Classification: This task broadens our perspective to encompass the classification of troll memes characterized by provocative and disruptive behavior. The gold standard dataset for this task includes languages such as Kannada and Telugu. Participants face the challenge of categorizing memes into 'Troll' and 'Non-Troll' categories, navigating the intricate interplay of humor, satire, and harmful intent.

Our competition is structured to encourage innovation and collaboration across the global research and practitioner communities. Participants receive comprehensive training and development datasets, offering various memes for practical model training.

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 139–144

Evaluation of these models relies on the macro-F1 score, a robust metric commonly used in natural language processing. In total, 52 teams participated in our shared task: Four teams in Tamil and three teams in Malayalam submitted a system to Task 1 and achieved the top score of 0.73 in Tamil and 0.87 in Malayalam. Due to the null participation in Task 2, we stopped running the task further.

Beyond the competition, our overarching goal is to contribute to a safer and more inclusive digital ecosystem. By dissecting and understanding the dynamics of meme content, we aim to pave the way for more effective content moderation strategies. We envision this initiative as a catalyst for fostering responsible online behavior and promoting gender equality.

2 Related Works

In recent research, Singhal et al. (2022) did a comprehensive data collection of 22,435 instances of fact-checked content from social media to scrutinize the proliferation of fake news across India between 2013 and 2020. This dataset is distinguished by its coverage across 13 languages, encapsulating 14 distinct attributes. It highlights the diversity and complexity of fake news dissemination within the multilingual and multicultural Indian context, offering insights into the dynamics of misinformation across various domains and media types.

Singhal et al. (2019) presented "SpotFake," a novel framework that surpasses existing systems by avoiding dependency on sub-tasks like event discrimination, focusing instead on directly leveraging textual and visual content through advanced language and image processing models (BERT and VGG-19). This approach demonstrates superior performance on Twitter and Weibo datasets, improving detection accuracy significantly.

Ramamoorthy et al. (2022) introduced a pioneering approach to meme analysis, providing gold-standard data for sentiment analysis, emotion classification, and intensity of emotion. The study presented baseline models, including a text-only model using LSTM and a multimodal model combining ResNet-50 and BERT, demonstrating the potential of incorporating text and images for improved performance.

Suryawanshi et al. (2023) proposed a comprehensive framework for analyzing image-with-text (IWT) memes, or "troll memes," introducing a three-level taxonomy to understand trolling's impact on domain-specific opinion manipulation. They enriched the Memotion dataset to create the TrollsWithOpinion dataset, containing 8,881 IWT memes in English, revealing challenges in classifying memes on the third level of the taxonomy.

Hossain et al. (2022) introduced the multimodal dataset "MemoSen" for the Bengali language, comprising 4,368 memes annotated with sentiment labels. Experiments on the MemoSen dataset showed a significant enhancement in meme sentiment classification with multimodal information integration.

Gasparini et al. (2022) created a benchmark meme dataset for automatic misogyny detection using 800 memes collected from various online sources. The dataset, analyzed by experts and crowdsourcing, included categories such as misogynistic, hostile, and ironic, with 100% agreement on 800 memes from three experts.

Suryawanshi et al. (2020) developed a system employing an early fusion technique to combine text and image modalities, contrasting its efficacy with baseline models focusing solely on either text or image.

Koutlis et al. (2023) introduced MemeFier, a deep learning-based architecture, featuring a dual-stage modality fusion module for fine-grained Internet image meme classification. Hegde et al. (2021) presented a transformer-transformer architecture, incorporating attention as a key component for classifying memes in the Tamil language.

Potential research gaps include the need for a unified evaluation metric and benchmark dataset for consistent comparison, the exploration of cross-cultural meme classification, the investigation of interpretability in model decision-making, and the development of more robust techniques to address biases and fairness concerns in meme classification models.

3 Task Description

The competition, "Shared Task on Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes," includes a challenge that focuses specifically on spotting harmful content in memes. While the overall competition has two parts, this paper discusses the part about finding misogynistic memes in Tamil and Malayalam languages. Organized as a part of the LT-EDI@EACL 2024 event[1], this task is designed to encourage

[1] https://2024.eacl.org

experts to come up with ways to identify when a meme is offensive towards women.

In Task 1, the participants are tasked with creating a tool that can look at memes (combining pictures with text) and determine if the meme is disrespectful or harmful to women. The task deals with memes in Tamil and Malayalam, and the participants are given training and development sets. They use these sets to teach their tools to distinguish between misogynistic memes and those that are not. Then, we will provide them with the test set without the labels. With this set, the participants will make the model to predict whether the meme is misogyny and submit it as a submission. Finally, we will judge the participant's model with their prediction with the true labels of the test set.

The tools are judged by how accurately they can make these distinctions, with the macro F1 score. The goal here is to push forward the development of tools that can spot and reduce the sharing of memes that can be hurtful to women in these two languages. This challenge is meant to attract attention from people worldwide who work in language and technology-related fields, hoping to spark more research and solutions in this important area. This shared task is conducted via the Codalab competition[2].

4 Dataset description

The dataset underneath the misogyny meme classification competition offers a comprehensive look at the manifestation of misogynistic content within the digital landscape, particularly within Tamil and Malayalam languages. Misogynistic memes target women or girls, often by leveraging stereotypes, displaying bias, or promoting discrimination. They might generalize women's capabilities with statements implying inferiority, demean their achievements, or mock female-specific issues to reinforce negative stereotypes and biases. Our dataset encompasses both monolingual and bilingual memes, with some featuring a mix of Tamil-English or Malayalam-English content, presenting an open challenge for research due to the code-mixed nature of these texts.

This dataset focuses primarily on monolingual content in Tamil and Malayalam, both of which are part of the Dravidian language family. Through this dataset, we aim to understand the extent and nature of misogynistic memes in these languages, exploring the linguistic and cultural factors contributing to their creation and spread. This is vital for researchers and practitioners dedicated to combating digital misogyny, especially in the context of Dravidian languages. The dataset consists of 1,776 Tamil memes, with 1,135 employed in the training set, 285 in development, and 356 in the test set. The data statistics for Malayalam data are shown in Table 1. The Malayalam dataset consists of 1,000 memes, with 640 in the training, 160 in the development, and 200 in the test set. The data statistics for Malayalam data are shown in Table 2.

5 Participants methodology

A total number of 52 participants were enrolled in this competition. In Task 1, we got a total of 4 submissions for the Tamil language and 3 submissions for the Malayalam language. The methodologies and results of these tasks have been discussed. To get more crucial material, please consult their papers, which are listed below:

Quartet (H et al., 2024) team participated in Task 1. They employed two different approaches to obtain the classification probabilities from the image and text data. With the textual data, every word of the text was translated into English. Subsequently, the translated sentences were preprocessed by eliminating emojis, punctuations, and stopwords. Then, the TF-IDF vectorizer is employed to obtain the embeddings from the preprocessed texts. The probability of the text being misogynistic was determined using the Multinomial Naive Bayes classifier. With the Pictorial Data, they employed the ResNet50 model (He et al., 2016) for performing transfer learning to obtain the probability of images being misogynistic. Using those probabilities, the employed fusion technique calculates the resultant probability.

DLRG team participated in Task 1. They worked on only textual data to classify the memes as misogyny or not. They employed Multilingual Bert(Bidirectional Encoder Representations from Transformers) (Kenton and Toutanova, 2019), a transformer-based multilingual pretrained model. They performed a transfer learning approach with the transcriptions and the labels.

Word Wizards team participated in Task 1. They worked on only textual data to classify the memes as misogyny or not. They performed tokenization and extracted word embeddings using

[2] https://codalab.lisn.upsaclay.fr/competitions/16097

Sets	Misogyny	Not-misogyny	Total
Train	272	863	1135
Development	76	209	285
Test	100	256	356
Total	448	1,328	1,776

Table 1: Data statistics for Task1 Tamil dataset for misogyny memes classification

Sets	Misogyny	Not-misogyny	Total
Train	256	384	640
Development	64	96	160
Test	80	120	200
Total	400	600	1,000

Table 2: Data statistics for Task1 Malayalam dataset for misogyny memes classification

TF-IDF vectorizer. With the work embeddings got from TF-IDF, they trained the SVM classifier to classify the meme into misogyny or not-misogyny for Tamil and Malayalam.

MUCS (Mahesh et al., 2024) team also participated in Task 1. They work on both meme images and transcriptions. Their methodology comprises a dual-encoder approach incorporating three distinct textual feature encoders alongside a shared image feature encoder: i) bert-base-uncased + ResNet-50, ii) muril-base-cased + ResNet-50, and iii) bertbase-multilingual-cased + ResNet-50.

6 Results

This section describes the results of a misogyny meme classification competition, where participants were evaluated based on the Macro F1 score—a measure used to test the accuracy of their machine learning (ML) and deep learning (DL) algorithms. In the Tamil results described in 3, MUCS_run3 achieved the highest rank with a Macro F1 score of 0.73, followed by DLRG with 0.69, Quartet with 0.65, and WordWizards_run1 with 0.60, ranking them from first to fourth, respectively. In the Malayalam results illustrated in 4, MUCS_run2 came out on top with an impressive Macro F1 score of 0.87, Quartet followed closely with a score of 0.83, and WordWizards_run1 also showed strong performance with a score of 0.8. These rankings provide a quantitative assessment of the participants' algorithmic approaches in the classification task. The results, as presented in this paper, showcase not only the potential but also the challenges inherent in automating the detection of misogyny and trolling in memes. While the best-performing systems exhibited promising results, there remains considerable scope for improvement, especially in handling code-mixed content and subtle cultural nuances. The shared task has also highlighted the need for further research into the creation of more sophisticated algorithms that can navigate the complexities of language, context, and intent.

Team name	M_F1	Rank
MUCS_run3 (Mahesh et al., 2024)	0.73	1
DLRG	0.69	2
Quartet (H et al., 2024)	0.65	3
WordWizards_run1	0.60	4

Table 3: Tamil results for misogyny memes classification

Team name	M_F1	Rank
MUCS_run2 (Mahesh et al., 2024)	0.87	1
Quartet (H et al., 2024)	0.83	2
WordWizards_run1	0.80	3

Table 4: Malayalam results for misogyny memes classification

7 Conclusion

In conclusion, the first shared task on "Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes" has been a groundbreaking effort to address the pressing issue of online misogyny and trolling within the context of Dravidian languages. The participation of dedicated teams in the Tamil and Malayalam classification tasks demonstrates a collective commitment to understanding and combating such harmful online

content. The datasets, precisely compiled and annotated, provided a robust foundation for the teams to deploy and test a variety of machine learning and deep learning models, which were assessed based on their Macro F1 scores.

Furthermore, this paper stands as a testament to the collaborative efforts required to address the multifaceted challenges presented by online misogynistic and troll memes, and it is hoped that it will inspire continued research and action in this vital area.

8 Acknowledgement

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223, supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2).

References

Trenton W Ford, Rachel Krohn, and Tim Weninger. 2023. Competition dynamics in the meme ecosystem. *ACM Transactions on Social Computing*, 6(3-4):1–19.

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.

Shaun Allan H, Samyuktaa Sivakumar, Rohan R, Nikilesh Jayaguptha, and Durairaj Thenmozhi. 2024. Quartet@LT-EDI 2024: A SVM-ResNet50 Approach For Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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. *arXiv preprint arXiv:2104.09081*.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Memosen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Bisera Kostadinovska-Stojchevska and Elena Shalevska. 2018. Internet memes and their socio-linguistic features. *European journal of literature, language and linguistics studies*, 2(4).

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memefier: Dual-stage modality fusion for image meme classification. *arXiv preprint arXiv:2304.02906*.

Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2024. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*.

Sidharth Mahesh, Sonith D, Gauthamraj, Kavya G, Asha Hegde, and H L Shashirekha. 2024. MUCS@LT-EDI-2024: Exploring Joint Representation for Memes Classification. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, Subalalitha Cn, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. Overview of shared-task on abusive comment detection in Tamil and Telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Parth Patwa, Amitava DaS, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 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*.

AAPK Rasheed, C Maria, and A Michael. 2020. Social media and meme culture: A study on the impact of internet memes in reference with 'kudathai murder case'. *Kristu Jayanti College*.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1322–1331.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2023. Trollswithopinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes. *Multimedia Tools and Applications*, 82(6):9137–9171.