Quartet@LT-EDI 2024: A SVM-ResNet50 Approach For Multitask Meme Classification - Unraveling Misogynistic and Trolls in Online Memes

Shaun Allan H
Sri Sivasubramaniya Nadar College of Engineering
shaunallan2210716@ssn.edu.in

Samyuktaa Sivakumar
Sri Sivasubramaniya Nadar College of Enginerring
samyuktaa2210189@ssn.edu.in

Rohan R
Sri Sivasubramaniya Nadar College of Enginerring
rohan2210124@ssn.edu.in

Nikilesh Jayaguptha
Sri Sivasubramaniya Nadar College of Engineering
nikilesh2210219@ssn.edu.in

Durairaj Thenmozhi
Sri Sivasubramaniya Nadar College of Engineering
theni_d@ssn.edu.in

Abstract

Meme is a very popular term prevailing among almost all social media platforms in recent days. A meme can be a combination of text and image whose sole purpose is meant to be funny and entertain people. Memes can sometimes promote misogynistic content expressing hatred, contempt or prejudice against women. The Shared Task LT–EDI 2024: Multitask Meme classification – Unraveling Misogynistic and Trolls in Online Memes Task 1 was created with the purpose to classify social media memes as "Misogynistic" and "Non - Misogynistic". The task encompassed Tamil and Malayalam datasets. We separately classified the textual data using Multinomial Naive Bayes and pictorial data using ResNet50 model. The results of from both data were combined to yield an overall result. We were ranked 2nd for both languages in this task.

1 Introduction

Social Media is a platform where millions of people connect and engage with each other. Social media has shaped the way people communicate, share ideas and information among each other colossally. With the immense number of people joining social media each day, social media platforms have become inevitable and play a pivotal role in the modern society.

With the rising usage of social media platforms in the society, they are also used as a source of entertainment. People create entertaining content and post it on social media which is then viewed by millions of people on the internet. With this trend of people posting entertaining contents on social media, a term called "Meme" has become prominent especially among youngsters in the society (Huang et al., 2022).

Memes are a ubiquitous form of internet culture whose sole purpose is to be funny and entertain people. A meme can be a text, image, video, audio or a combination of these that embodies humour, sarcasm or irony in it. Memes has the ability to transcend linguistic and cultural barriers reaching a wide and diverse range of audience. While memes can serve as a form of humorous and relatable content on social media, they also have their darker side that includes misogyny portraying very harmful stereotypes about women, objectifying them and also initiate gender-based hatred and violence.

Multimodal data analysis can be employed to analyse the memes that are available on the internet. A modality is defined as the type or the nature of representation of the data which includes text, image, video and audio. Multimodal data is a representation of data that comprises of two or more modalities of data (Lahat et al., 2015). A meme which can be a combination of different modalities such as text and image, text and video, audio and video, etc are multimodal in nature. Multimodal data analysis can be applied on these memes to classify them as "Misogynistic" or "Non-Misogynistic". Supervised learning can be used for the process for which a well labelled balanced training data is very essential. In case of unavailability of a proper training data, unsupervised leaning can also be performed to carry on multimodal data analysis.

Misogyny is something that deprives women of their rights and privileges and promotes toxic masculinity. In a society which is moulding itself towards gender equality and women empowerment, misogyny should be eliminated. With social media being an inevitable and widely used tool in recent times, having misogynistic content in them will lead to a lot of misinformation and stereotypes against women. Therefore, detecting and moderating these types of memes in social media platforms is indeed vital and assists the movement towards a better society.

The given task aims to encourage the development of models for detecting misogynistic memes in Tamil and Malayalam. The memes and the text inscribed in them were provided in the dataset for

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 221–226
March 21, 2024 ©2024 Association for Computational Linguistics

both Tamil and Malayalam.

2 Related Works

Suryawanshi et al. (2020) developed a meme classification system using an early fusion technique to combine the text and image modality and compared it with a text and an image only baseline to investigate its effectiveness.

Simple prompts were constructed and a few in-context examples were provided by Cao et al. (2023) to exploit the implicit knowledge in the pre-trained RoBERTa language model for hateful memes classification.

Koutlis et al. (2023) proposed a deep learning-based architecture for fine-grained classification of Internet image memes called MemeFier. MemeFier utilizes a dual-stage modality fusion module.

A bias estimation technique is proposed by Rizzi et al. (2023) to identify specific elements that compose a meme that could lead to unfair models, along with a bias mitigation strategy based on Bayesian Optimization. Gu et al. (2022) used a joint image and text classification technique to classify memes as either misogynistic or not.

Kumar and Nandakumar (2022) explicitly modelled the cross-modal interactions between the image and text representations contained using Contrastive Language-Image Pre-training (CLIP) encoders via a feature interaction matrix (FIM).

An ingenious model comprising of a transformer-transformer architecture was proposed Hegde et al. (2021) to classify memes in Tamil language. The proposed model tries to attain state-of-the-art by using attention as its main component.

Velioglu and Rose (2020) utilized VisualBERT that was trained multimodally on images and captions and applied Ensemble Learning to build an automatic hateful meme classification system.

Li (2021) explored a multimodal transformer for meme classification in Tamil language. According to the characteristics of the image and text, different pre-trained models were used to encode the image and text so as to get better representations of the image and text respectively.

3 Task and Data Description

The Shared Task LT-EDI 2024: Multitask Meme classification – Unraveling Misogynistic and Trolls in Online Memes[1] (Chakravarthi et al., 2024) Task 1 was created with the purpose of classifying memes as "Misogynistic" and "Non–Misogynistic". Memes on languages Tamil and Malayalam were provided to us as datasets. A sample record from the dataset encompassed the meme image, the text that is inscribed in the image and the label of whether the meme is misogynistic or not.

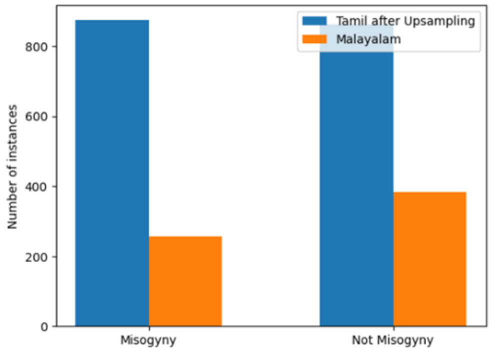


Figure 1: Data Distribution

3.1 Tamil Dataset

The training dataset for Tamil totally had 1137 records. The labels in the training dataset appeared to be case variants of the same words, “Misogyny” and “Not-Misogyny”. In the training dataset, 659 were classified as “not-misogyny”, 204 were classified as “Not-Misogyny”, 39 were classified as “misogyny” and 235 were classified as “Misogyny”. After replacing “not-misogyny” and “Not-Misogyny” with “Not Misogyny” and “misogyny” with “Misogyny”, we had 274 labelled “Misogyny” and 863 labelled “Not-Misogyny”.

The data is very imbalanced which introduces a bias towards the majority data. After up sampling the data, the number of records labelled “Misogyny” is increased to 864.

3.2 Malayalam Dataset

The training dataset for Malayalam totally had 640 records out of which 256 were classified as “Misogyny” and 384 were classified as “Not Misogyny”

4 Methodology

The approach we took was to create build two separate models for each modality. We employed Multinomial Naive Bayes for text classification and ResNet50 for image classification. The resulting probabilities from each model were considered and simple arithmetic was performed to yield the overall result.

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

4.1 Textual Data

4.1.1 Data Preprocessing

Before using the data for training the model, the data must be processed and cleansed for the model to be reliable and yield better results.

1. As emoticons and punctuations are insignificant to the classification process, these characters are removed from the texts.

2. The given text then translated to English which yields better results as most of the embedding systems available are ideally built for English.

3. Stop words are words that doesn't have any contribution in adding meaning to the text. So, these stop words are discarded from the text using the NLTK library.

4.1.2 Feature Extraction

We employed Term Frequency–Inverse Document Frequency (TF-IDF) vectorizer to covert the raw text into vectors consisting of numerical inputs. Term Frequency (TF) refers to the number of times a term appears in a particular document. Inverse Document Frequency (IDF) is a measure of how common a term is across the entire corpus of documents. TF-IDF value of a term in a document is the product of its TF and IDF.

4.1.3 Classification using ML Models

Our main focus was on obtaining the probability of the text being misogynistic rather than obtaining binary outcomes. Traditional models such as Logistic Regression and several types of Naive Bayes models that yield probabilities were experimented on the extracted features. After assessing the metrics of all the models, Multinomial Naive Bayes produced the best numbers of all.

4.2 Pictorial Data

4.2.1 Data Preprocessing

Before using the data for training the model, the data must be processed and cleansed for the model be reliable and yield better results.

1. The given pictorial data is in JPG format is converted into a M-by-N by 3 array representing the RGB values at each pixel of the image.

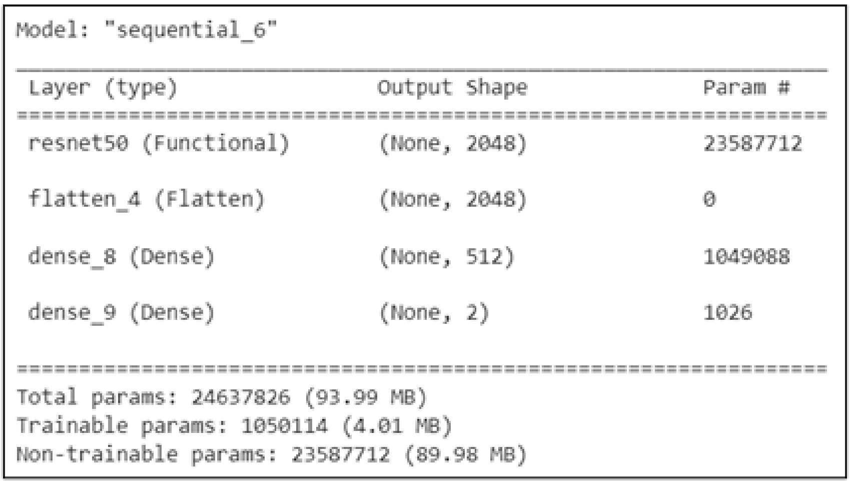
2. The size of the matrix obtained varies from image to image based on its resolution. As the model only takes inputs of a fixed size for which it is to be trained, the images are uniformly resized to 200 X 200 pixels.

3. The resultant matrix ranging from 0 to 255 is normalized to the range of 0 to 1 making computations much faster and easier.

4.2.2 Classification using ML Models

We employed Transfer Learning on ResNet50 (He et al., 2016) model to classify images. Transfer Learning is a technique in machine learning which refers to the reuse of a pre-trained model on a similar task exploiting the knowledge of the pre-trained model. ResNet50, based on Convolutional Neural Network (CNN) architecture is a very powerful pre-trained model used for image classification which is 50 layers deep and has over 23 million trainable parameters.

In our case, we constructed a sequential neural network having ResNet50 model as the first hidden layer and the output layer was activated with softmax function. The model was compiled using Adam optimizer and Categorical Crossentropy.



```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resnet50 (Functional)       (None, 2048)              23587712

 flatten_4 (Flatten)         (None, 2048)              0

 dense_8 (Dense)             (None, 512)               1049088

 dense_9 (Dense)             (None, 2)                 1026

=================================================================
Total params: 24637826 (93.99 MB)
Trainable params: 1050114 (4.01 MB)
Non-trainable params: 23587712 (89.98 MB)
```

Figure 2: Summarization of Neural Network constructed for Image classification

4.3 Fusion

A simple arithmetic formula is applied to obtain the resultant probability.

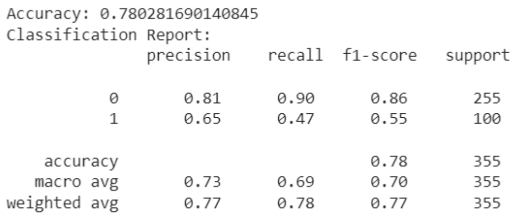
$$
\begin{aligned}
ResulatantProbability = \\
0.7 * ProbabilityFromText \\
+ 0.3 * ProbabilityFromImage
\end{aligned}
$$

The numbers 0.7 and 0.3 are numbers obtained from trial and error for which the model yielded better metrics. The data is classified as Misogyny if the resultant probability is greater than or equal to 0.5, otherwise it is classified as Not Misogyny.

5 Results

5.1 Tamil

Our model yielded an accuracy of 0.97 on training data and 0.77 on development data provided by the organizer. The macro average f1 score was 0.98 on training data and 0.69 on development data.



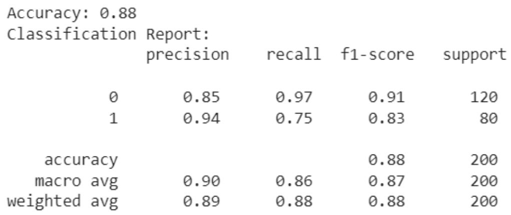
```
Accuracy: 0.780281690140845
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.90      0.86       255
           1       0.65      0.47      0.55       100

    accuracy                           0.78       355
   macro avg       0.73      0.69      0.70       355
weighted avg       0.77      0.78      0.77       355
```

Figure 3: Classification Report on Testing Data - Tamil

For the test data provided by the organizers, the model produced a macro average score of 0.70 and we were ranked 2nd in the rank list released by the organizers.

5.2 Malayalam

Our model yielded an accuracy of 0.94 on training data and 0.84 on development data. The macro average f1 score was 0.94 on training data and 0.82 on development data.



```
Accuracy: 0.88
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.97      0.91       120
           1       0.94      0.75      0.83        80

    accuracy                           0.88       200
   macro avg       0.90      0.86      0.87       200
weighted avg       0.89      0.88      0.88       200
```

Figure 4: Classification Report on Testing Data - Malayalam

For the test data provided by the organizers, the model produced a macro average score of 0.87 and we were ranked 2nd in the rank list released by the organizers.

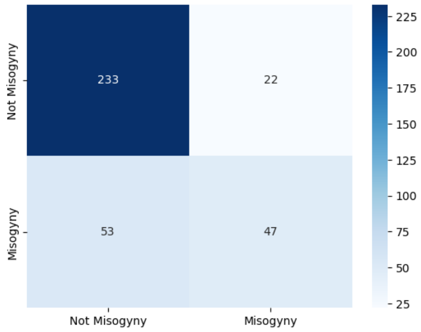


Figure 5: Confusion Matrix on Testing Data - Tamil

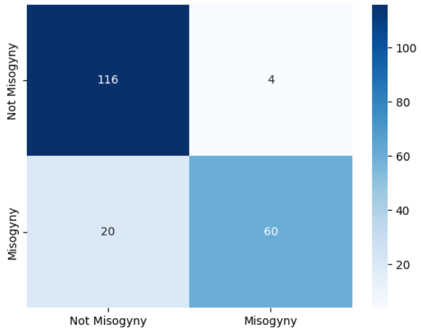


Figure 6: Confusion Matrix on Testing Data - Malayalam

Dataset	Textual Data			Pictorial Data	Overall Model (Most Optimum)
	Logistic Regression	Gaussian Naive Bayes	Multinomial Naive Bayes	ResNet50	Multinomial NB + ResNet50
Tamil	0.64	0.63	0.66	0.71	0.69
Malayalam	0.63	0.65	0.66	0.79	0.82

Table 1: Comparison of macro average f1 score on Development Data

6 Limitations

As two separate models were used in the proposed solution, one for text and the other for image, the training process takes place for both the models, thereby increasing the training time of the overall procedure. Masking the text in the images emanates a significant increase in the performance of the model which could not be implemented as the operation requires large GPU resources.

Due to unavailability of a balanced dataset, even after up sampling the Misogynistic instances, a small amount of bias towards the Non-Misogynistic category is still present over the Misogynistic category. The TF-IDF vectorizer which is used to extract features from textual data computes document similarity directly in the word-count space, which may be slow for large vocabularies. Also, the semantic relations between words are not considered during feature extraction.

7 Ethics Statement

The ACL Code of Ethics[2] has been followed and practiced throughout the process of working on the shared task. The classification system is built with the notion to eliminate misogyny from the society resulting in a safe and inclusive social environment for all community of people to participate in. All the authors whose existing ideas, invention, work or artifact has been referenced or utilized is given credit providing a link to the original work in the References section. Our solution prioritizes data privacy by not providing any access to random entities ensuring no leak of information to any other individual or organization. The given task was used as an opportunity to upgrade and enhance our skills while practicing the principles for professional competence. The proposed solution abides by the local, regional, national and international laws and regulations.

References

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification.

Bharathi Raja Chakravarthi, Saranya Rajiakodi, Rahul Ponnusamy, Kathiravan Pannerselvam, Anand Kumar Madasamy, Ramachandran Rajalakshimi, Hariharan RamakrishnaIyer 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 and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Qin Gu, Nino Meisinger, and Anna-Katharina Dick. 2022. QiNiAn at SemEval-2022 task 5: Multi-modal misogyny detection and classification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 736–741, Seattle, United States. 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.

Victor Huang, Yifan Hu, and Yaohua Li. 2022. A systematic literature review of new trends in self-expression caused by emojis and memes. In *Proceedings of the 2021 International Conference on Social Development and Media Communication (SDMC 2021)*, pages 75–79. Atlantis Press.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 586–591, New York, NY, USA. Association for Computing Machinery.

Gokul Karthik Kumar and Karthik Nandakumar. 2022. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features.

Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

Zichao Li. 2021. Codewithzichao@DravidianLangTech-EACL2021: Exploring multimodal transformers for meme classification in Tamil language. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 352–356, Kyiv. 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.

[2] https://www.aclweb.org/portal/content/acl-code-ethics

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, Marseille, France. European Language Resources Association (ELRA).

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge.