Transformers at SemEval-2022 Task 5: A Feature Extraction based Approach for Misogynous Meme Detection

Shankar Mahadevan[1], **Sean Benhur**[2], **Roshan Nayak**[3], **Malliga Subramanian**[4], **Kogilavani Shanmugavadivel**[4], **Kanchana Sivanraju**[2], **Bharathi Raja Chakravarthi**[5]
[1] Thiagarajar College of Engineering, India [2]PSG College of Arts and Science, India
[3]B.M.S College of Engineering, India [4]Kongu Engineering College, India
[5]Insight SFI Research Centre for Data Analytics,National University of Ireland, Galway
shankarmahadevan12901@gmail.com[1], seanbenhur@gmail.com[2], roshannayak610@gmail.com[3]
mallinishanth72@gmail.com[4], kogilanvani.sv@gmail.com[4], kanachana@psgcas.ac.in[2],
bharathi.raja@insight-centre.org[5]

Abstract

Social media is an idea created to make the world smaller and more connected. Recently, it has become a hub of fake news and sexist memes that target women. Social Media should ensure proper women's safety and equality. Filtering such information from social media is of paramount importance to achieving this goal. In this paper, we describe the system developed by our team for SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. We propose a multimodal training methodology that achieves good performance on both the subtasks, ranking 4th for Subtask A (0.718 macro F1-score) and 9th for Subtask B (0.695 macro F1-score) while exceeding the baseline results by good margins. The code will be available here[1]

1 Introduction

With the rising usage of social media and the Internet, it is tougher to establish an inclusive and welcoming community among users. Offensive speech, hate speech, and targeted insult have been increasing among users, disturbing everyone. With the rising utilization of the Internet in a pandemic, hate speech prevalence on the Internet is also increased. Online hate speech with targeted discrimination also creates threats and crimes offline.

Among these, online misogyny or sexist comments have been increasing among women (Salter et al., 2018), which includes name-calling, sexual threats, shaming. This emphasizes the need for specialized automatic misogyny detection in online platforms. Platforms such as Twitter [2] and Facebook already have policies for banning hateful content. However, these systems are primarily through manual methods and might not scale well for large users and multimodal content. Moreover, hate speech is also prevalent in multimodal form since most social media platforms support Images, text, audio and video content. These memes have been popular among users to express their opinions since people express information through memes, GIFs, and videos. But, unfortunately, this also causes the rise of multimodal hate, and offensive content, which is disturbing to users (Suryawanshi et al., 2020). This includes misogynistic posts, which are targeted towards women.

Previous works on Misogyny detection have been primarily focusing on one modality, which is text (Pamungkas et al., 2020). Misogyny detection in text falls under an area of text classification. Text classification methods such as BERT, LSTM, Naive Bayes have been used to detect misogynistic comments. In this work, we have used the provided data, which contains both images of the memes and the extracted OCR text from the memes.

This paper describes our submission for the task; we have used multiple concatenation-based fusion models and ensembled them for the final submission.

[1] https://github.com/shankrmahadevan/semevalmami2022

[2] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 550 - 554
July 14-15, 2022 ©2022 Association for Computational Linguistics

2 Related Work

The past works have concentrated on collecting the dataset from popular social media sites such as Facebook, Reddit, and Twitter. Recent statistics do not simply focus on hate but also on the kind of hate the meme attempts to spread. Work has also been done in detecting offensive memes using various pre-trained models. (Fersini et al., 2019) presents a novel dataset for the sexist meme classification task. Sexism could be of several forms that could be categorized based on the context of the caption and the objects on the meme. The main types of sexism against women addressed in the dataset are shaming, stereotypes, objectification and violence. The research paper largely focused on a comparison between the unimodal and multimodal classifiers. The article has attempted to answer various research concerns such as whether unimodal architectures can predict the target correctly, will merging the features of both text and picture capture the inherent complexity of the sexist memes, and which one of the two modalities dominates the other. The research discovered that unimodal classifiers have shown that textual information is an excellent indicator, whereas visual information is a poor indicator to identify sexist memes. This study between unimodal and multimodal showed that unimodal architectures performed better than multimodel architectures.

In the paper, (Zia et al., 2021) the analysis is done on the dataset that focuses beyond hateful or not hateful by annotating the hate meme dataset further by the kind of hate the meme is actually spreading. This would help in understanding the meme and the intention of the person who created the meme better rather than just labelling it as hateful or not. The paper focused on two tasks. The first task was to identify the kind of hate the meme intended to spread. The second task was to detect the type of attack the meme did on a particular group such as slurs, inferiority, and mocking. Models such as CLIP (Contrastive Language Image Pre-Training) and LASER (Language Agnostic SEntence Representations), LaBSE (Language agnostic BERT Sentence Embedding) were used to extract features from the image and text. The paper concluded that multimodel architectures outperformed unimodal architecture. The multimodal, concatenated textual features (CLIP, LASER, and LaBSE) and visual features (CLIP) was the best performing model with AUROC of 0.96 and 0.97 for task A and task B, respectively.

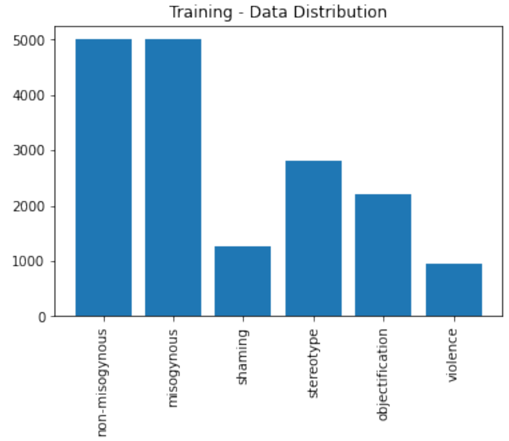

Figure 1: Training data distribution. It can be seen that the positive and negative classes are in equal proportions.

(Guest et al., 2021) introduced a taxonomical dataset of 6,383 samples from Reddit. The dataset has a three-level taxonomy which makes this dataset very different from what already exists. The first level is a binary classification between misogynistic content and non-misogynistic content. The second level corresponds to the subtypes of misogynistic and non-misogynistic content. Misogynistic content categories include misogynistic pejoratives, misogynistic treatment, misogynistic derogation and gendered personal attacks against women. Non-misogynistic content categories include counter speech against misogyny, non-misogynistic personal attacks and None of the categories. In the third level, additional flags for some of the second-level categories have been defined. BERT based models were trained on the dataset to achieve a test accuracy of 0.93.

3 Dataset

The dataset provided for the competition (Fersini et al., 2022) consisted of images of memes and OCR extracted text and labels for both subtask A and B. For Subtask A, the binary label of misogynous is given; for Subtask B, four labels were given: they are shaming, stereotype, objectification and violence, each of them containing binary values 0 or 1. The provided dataset contains 10,000 training images, 100 validation images and 1000 images for test set submission. The training dataset was randomly shuffled and split into five-folds, with each fold containing 2000 images each. The first four folds were used to train the model, and the last fold, along with the eval dataset provided, was chosen as the validation dataset to improve the model performance. The final number of samples in training, validation and test set are reported in Table 1. The Data distribution is given in Figure 1.

4 Preprocessing

Since the text was extracted from the memes using OCR tools, much noise in the text had to be cleaned manually. First, all internet links, stopwords and Twitter user handles were removed from the text. Then, the text was lemmatized using a word-net based lemmatizer. The text truncation length was set to 256. An interesting observation in the dataset was that memes that were not misogynistic in nature did belong to the other four classes. So, it was evident that a meme might not be misogynistic yet belong to any of the other subcategories.

5 Methodology

5.1 Models

Since this topic was multimodal in nature, we finetuned multiple text-based models and image-based models to handle this task. Convolutional Neural Networks (CNNs) excel in image classification challenges due to their intrinsic spatial inductive bias. CNNs have been leading the computer vision research arena for the last two decades due to their superior spatial comprehension ability. The CNN based image models chosen for this task are: InceptionV3 (Szegedy et al., 2015) and EfficientNetB7 (Tan and Le, 2020) from the TensorFlow library. The BERT (Devlin et al., 2019) model was used as the text feature extraction backbone. We also tried finetuning CLIP (Radford et al., 2021) for this task, as it was trained in a multimodal fashion. For both CLIP and other multimodal models we added a fully connected layer with softmax for classification.Another approach was to extract a set of embeddings from State-of-the-Art Text and Image models and classify the features using Support Vector Machines (SVMs). The models selected for this approach were: XLM-RoBERTa (Conneau et al., 2020), DistilBERT(Sanh et al., 2020), ResNext (Xie et al., 2017) and Data-efficient Image Transformer (Touvron et al., 2021).

5.2 Experiment Setup

We implement our multimodal training in TensorFlow using Tensor Processing Units (TPUs) offered by Google Colab for training. TPUs greatly shortened the time required to conduct numerous tests and hyper-parameter optimization. All the photos were resized to 256x256. The TensorFlow version of the BERT and CLIP models from the transformers library was used. A CUDA-accelerated implementation of SVM from the cuML library created by NVIDIA was used in the SVM training.

Split	No. of Samples
Training	8,000
Validation	2,100
Test	1,000

Table 1: Dataset Split

Image Augmentation methods Typically, CNNs are trained using millions of images to attain good accuracy. However, since the number of photos available in the dataset was less in nature, image augmentation methods were added to generate synthetic augmented images and thus boost the amount of data utilized to train the model. This ensures that the model better generalizes to the patterns present in the image modality. We employ (i) random resizing, (ii) random cropping, (iii) random horizontal flipping and (iv) random vertical flipping as the augmentation methods.

The SVMs were trained using a single K80 GPU provided by Google Colab.

5.3 Multimodal Training

The Multimodal training illustrated in this section follows the procedure shown in Figure 2. The model using InceptionV3 and BERT backbone is termed as Model A, EfficientNet B7 and BERT as Model B, CLIP Image and CLIP Text Backbone as Model C. Models A and B use Adam optimizer with a base learning rate of 1e-06 and a linear learning rate decay. Model C uses Adam optimizer with a base learning rate of 6e-05 and a linear learning rate decay. Image preprocessing is done as provided by the model authors. Text cleaning and tokenization is performed for feeding to the text model. All the models are trained for 50 epochs with an Early Stopping callback to terminate the training when the model does not learn any discriminatory features and/or overfits to the training set. A fully connected layer is added at the end to perform classification.

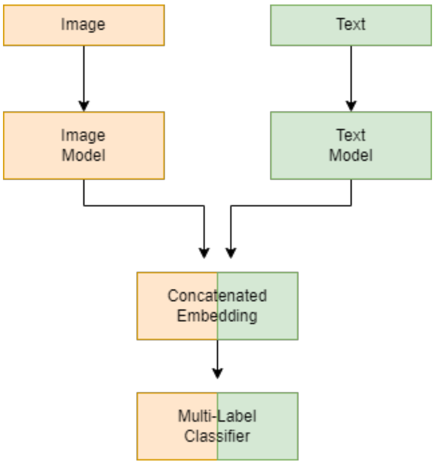

Figure 2: **An overview of the multimodal training approach.** Late fusion is adopted for effective classification.

5.4 SVM Training

In the multimodal training approach, complete models were finetuned for classification. In this approach, the pre-trained image and text models are used as feature extractors alone, and the features are supplied to SVMs for classification. Since the time complexity of training SVMs increases quadratically with respect to the available data, when the data becomes higher than tens of thousands of samples, it practically becomes impossible to train SVMs on CPUs. Since there is a significant amount of data in the training set, SVMs accelerated using CUDA from the cuML library were utilized for training the SVMs. Due to the highly parallel nature of GPU computation, the time required to train the SVM is reduced to seconds. Since SVM is a binary classifier, the multiclass classification problem is broken down into smaller binary classification problems. Thus, 5 SVMs are employed for classification.

6 Results and Discussion

The Test set results are reported in Table 2. Finetuning the CLIP model (Model C) gave results worse than the baseline findings provided by the task authors. So, Model C is not used when building the ensemble. Model A, Model B and SVM results exceed the baseline results by a good margin. This also illustrates that finetuning the models on a downstream task helped boost the accuracy, unlike the case of SVM where it was trained to classify using features extracted by a pre-trained network.

Model	**Task A**	**Task B**
Baseline	0.6500	0.6210
Model A	0.6893	0.6774
Model B	0.7005	0.6823
Model C	0.6537	0.5937
SVM	0.6760	0.6447
Ensemble	**0.7182**	**0.6949**

Table 2: Test Set Results

7 Conclusion

Thus we illustrate the system developed by us for SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. We compare multimodal finetuning vs classification of pre-trained network feature extraction. We have also discussed various methods adopted to train such models and also the data preprocessing done. We show the potential of employing such a model in real-world use cases.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and In-*

telligent Interaction Workshops and Demos (ACIIW), pages 226–231.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing Management*, 57(6):102360.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Michael Salter, Molly Dragiewicz, Jean Burgess, Ariadna Matamoros-Fernandez, Nic Suzor, Delanie Woodlock, and Bridget Harris. 2018. Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms. *Feminist Media Studies*, 18.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

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).

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers distillation through attention.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219, Online. Association for Computational Linguistics.