IIITH at SemEval-2022 Task 5: A comparative study of deep learning models for identifying misogynous memes

Tathagata Raha, Sagar Joshi, Vasudeva Varma
IIIT Hyderabad, India
{tathagata.raha, sagar.joshi}@research.iiit.ac.in
vv@iiit.ac.in

Abstract

This paper provides a comparison of different deep learning methods for identifying misogynous memes for SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In this task, we experiment with architectures in the identification of misogynous content in memes by making use of text and image-based information. The different deep learning methods compared in this paper are: (i) unimodal image or text models (ii) fusion of unimodal models (iii) multimodal transformers models and (iv) transformers further pretrained on a multimodal task. From our experiments, we found pretrained multimodal transformer architectures to strongly outperform the models involving fusion of representation from both the modalities.

1 Introduction

With the social media turning out to be a medium for propagation of hate speech and other perils, misogyny and sexism is incident upon women in explicit and implicit ways. Although memes have turned out to be a potent mechanism for exchanging humorous messages, they have been turning out to also be bearers of such malicious content. With this motivation, the task of Multimedia Automatic Misogyny Identification (MAMI) (Fersini et al., 2022) was proposed with two subtasks: (1) determining whether a meme is misogynous as a binary classification problem (2) finegrained misogyny classification into categories of stereotype, shaming, objectification and violence as a multilabel problem. In our work, we have compared different deep learning approaches for identifying misogyny in memes and also further classifying them into different kinds of misogyny.

We base our experiments on unimodal architectures making use of only either the textual or the image content in memes. The unimodal architectures were naturally superseded by their multimodal counterparts, since they made use of both the modalities in misogyny identification. Among the multimodal architectures, we initially experimented with simple fusion-based approaches which involved combining the image and text representations. These experiments were followed by trying out multimodal transformer architectures in which we made use of MMBT (Kiela et al., 2019), ViLBERT (Lu et al., 2019) and VisualBERT (Li et al., 2019). We used these architectures pretrained on unimodal as well multimodal objectives. We found VisualBERT and ViLBERT trained using multimodal objectives to perform competitively on the task. In order to further improve the capability of the models for misogynous content identification, we tried out further pretraining the models on a dataset for classifying hateful memes. This strategy involving a task-adaptive further pretraining stage turned out to further boost the performance of the models showing the benefit obtained from larger datasets for adapting models to a finegrained downstream task. Figure 1 shows the training stages of such an architecture.

Our best model achieved a macro-F1 score of 0.712 for Subtask 1, while the best performing model for Subtask 2 gave a weighted F1 of 0.706.

2 Related Work

Misogyny detection. Sexism and misogyny has been a long-studied problem, with (Barreto and Ellemers, 2005) and (Dardenne et al., 2007) bringing out the differences in explicit (hostile) and vieled (ambivalent) sexism, with the latter being observed to be subtly undermining and perilous to women. With misogynist remarks - a category of hate speech - being prevalent on social media, the dataset introduced by (Waseem and Hovy, 2016) for hate speech detection on tweets includes sexism as one of the categories in a multiclass problem. In a dataset introduced specifically for misogyny identification on tweets, (Anzovino et al., 2018) also design a taxonomy identifying different manifes-

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 673 - 678
July 14-15, 2022 ©2022 Association for Computational Linguistics

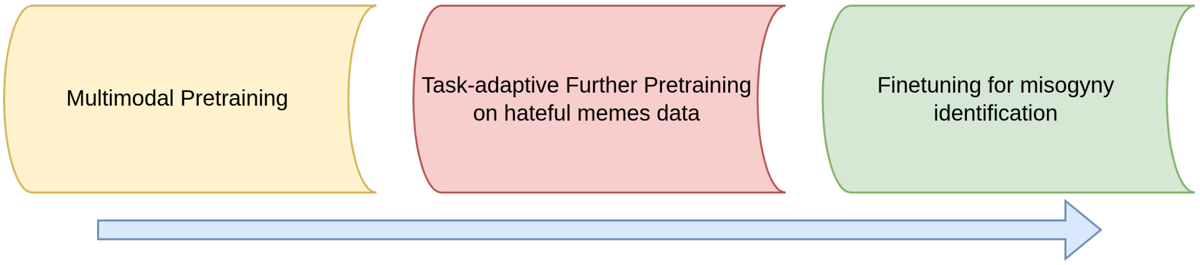
Multimodal Pretraining
Task-adaptive Further Pretraining on hateful memes data
Finetuning for misogyny identification

Figure 1: Stages of training for the model with best performance on misogyny identification

tations of sexism. Focusing on the differences in hostile and benevolent sexism, (Jha and Mamidi, 2017) curated a dataset for classifying the misogyny content in a tweet between the two categories, if the tweet is sexist in nature. Apart from detection of directed hateful content from tweets, there has been work on identifying categories of sexism from personal accounts such as (Karlekar and Bansal, 2018). (Parikh et al., 2019) created a dataset having 23 labeled categories of sexism from sexism accounts without maintaining mutual exclusivity in the categories and proposed a multi-task approach involving three auxiliary tasks for the multilabel classification in (Abburi et al., 2020).

Meme classification. The ubiquity of memes on internet, presence of malicious / hateful content in memes and the challenges involved in meme understanding were discussed in (Sharma et al., 2020). The work presented a new dataset and a challenge for understanding emotions in memes which involved subtasks for identifying the sentiment, humour category and scale (or intensity) of the detected class. (Suryawanshi et al., 2020) introduced a dataset for detection of offensive content in memes. A larger, challenging dataset was introduced in (Kiela et al., 2020) by involving 'benign confounders' to force the multimodal architectures to learn robust representations using both the modalities. The work also introduces formidable baselines with multimodally pretrained transformer encoders. Among the top performing models on this dataset, (Velioglu and Rose, 2020) perform an ensemble of multiple trained VisualBERT models, while ensembling was done in (Muennighoff, 2020) on a set of five predictions from different trained models, with the predictions for each model averaged from 3-5 different runs.

3 Task and dataset overview

The task consists of two subtasks:

- **Course-grained misogyny identification:** For this task, given a meme, we have to predict if a meme is misogynous in nature or not.
- **Fine-grained misogyny identification:** Given that a meme is misogynous, this task further identifies the kinds of misogyny among potential overlapping categories such as stereotype, shaming, objectification, and violence.

The dataset for the task was provided by the workshop organizers. The training set consisted of 10000 memes, whereas the hidden test consisted of 1000 memes. Each row in the dataset contained a unique identifier, the path to the image file for a corresponding meme, the text in the meme, and the label values of misogyny, stereotype, shaming, objectification, and violence.

4 Methodology

In the following section, we discuss our approaches for misogyny detection. We discuss our models in detail and provide a comparison between the models. We have explored unimodal models which use just the text or image as the input. The unimodal fusion models take the representation of the image part of the textual part separately and combine them to give the output. We have also exploited different pretrained multimodal transformers models. We have also experimented with how to further pretraining of these multimodal transformers models affect the quality of the predictions.

4.1 Unimodal models

For unimodal models, we experimented with the following models:

- **Image-Grid:** This is a unimodal image-based classifier that uses convolutional features with average pooling from ResNet-152 (He et al., 2016) architecture.

Setting	Subtask 1 (Macro F1-Score)	Subtask 2 (Weighted F1-Score)
Unimodal-Image-Grid	0.601	0.557
Unimodal-Image-Region	0.606	0.582
Unimodal-Text-BERT	**0.621**	**0.590**
Unimodal-Text-RoBERTa	0.619	0.585
Concat-BERT	**0.648**	**0.611**
Late-Fusion	0.626	0.608
MMBT-Grid	0.651	0.625
MMBT-Region	0.657	0.642
VilBERT	**0.687**	0.671
VisualBERT	0.684	**0.679**
VilBERT CC	0.693	0.683
VisualBERT COCO	0.685	0.689
VilBERT HM	**0.712**	0.698
Visual BERT HM	0.706	**0.702**

Table 1: Results on the testing split for each subtask. Task 1 refers to course-grained identification of misogyny and task 2 refers to the fine-grained identification ofthe types of misogyny.

- **Image-Region** In this unimodal image-based classifier, features from Faster-RCNN (Ren et al., 2015) with ResNeXt-152 (Xie et al., 2016) are used as the backbone network and are pretrained on the Visual Genome dataset (Krishna et al., 2017).

- **Text BERT:** This unimodal text-based approach uses BERT embeddings (Devlin et al., 2018) on the text given as part of the dataset.

- **Text RoBERTa:** Similar to the previous model but it used RoBERTa embeddings (Liu et al., 2019) instead of BERT.

4.2 Unimodal fusions

After taking unimodal representations, we have used the following techniques to fuse the representations to get the final representations before passing to the classifier:

- **Concat-BERT:** In this multimodal approach, an earlier fusion of the output of the unimodal ResNet-152 and BERT embeddings is performed by concatenation, and an MLP is trained for classification.

- **Late Fusion:** This is a simple multimodal approach where the output of ResNet-152 as in Image-Grid and BERT-based models is taken unimodally, and their mean is taken as the final model representation.

4.3 Multimodal transformers

For more advanced models, we have used the following multimodal transformers models:

- **MMBT-Grid:** MMBT (Kiela et al., 2019) is a multimodal supervised bitransformer architecture consisting of individual unimodally pretrained components trained to map multimodal image embeddings to text token space. MMBT-Grid uses features from ResNet-152 for image embeddings.

- **MMBT-Region:** In this approach, the MMBT transformer uses features from Faster-RCNN as in Image-Region for image embeddings.

- **ViLBERT:** ViLBERT (Lu et al., 2019) is a dual-stream multimodal transformer architecture. Here, the VilBERT model without any multimodal pretraining is used. It has BERT initializations for the text stream and uses Faster-RCNN pretrained on Visual Genome dataset to extract image region features.

- **Visual BERT:** Visual BERT (Li et al., 2019) is a multimodal single stream transformer architecture in which the text and image inputs are jointly processed by a stack of BERT-based transformer layers. It uses Faster RCNN for extracting image features.

4.4 Further pretrained models

From the models mentioned in the previous subsection, we have seen VilBERT and VisualBERT perform the best. We move forward with these models to further pretrain them on relevant datasets in a multimodal setting.

- **ViLBERT CC:** ViLBERT architecture used here is pretrained multimodally on the Conceptual Captions (CC) dataset (Sharma et al., 2018) using two pretraining tasks - masked multi-modal modelling (masking 15% of text and image region inputs and reconstructing them with unmasked inputs) and multi-modal alignment prediction (given a pair of image and text, determine if the text describes the image).

- **Visual BERT COCO:** Visual BERT architecture is pretrained multimodally on the Common Objects in Context (COCO) dataset (Lin et al., 2014). The two tasks the model is pretrained on are masked language modeling with an image (some part of the text is masked and is to be predicted using image regions and unmasked text) and sentence-image prediction (given two captions for an image, while one of them is the proper caption for the image, determine if the same holds for the remaining caption as well).

- **VilBERT HM:** For this architecture, we have pretrained the VilBERT architecture on the Hateful Memes dataset (Kiela et al., 2021) with the hypothesis that it will provide better representations given that it has been trained on memes that are hateful in nature. It has been pretrained on masked multi-modal modeling and a new task of meme-caption prediction(where given the image of the meme, the task is to choose the correct text from a given set of options).

- **Visual BERT HM:** Similar to the previous architecture, we have pretrained the Visual BERT model on the Hateful Memes dataset on masked multi-modal modeling task and meme caption prediction.

4.5 Final Setup

After the representation is obtained from any models above, it is passed through a multilayer perceptron classifier to predict the final label. For the second subtask, we used a hierarchical level modeling where the model would predict at first if a meme is misogynous or not, and if it is misogynous, it will perform further fine-grained classification.

4.6 Experimental details

We have used the pretrained models from the MMF framework (Singh et al., 2020) by Facebook AI Research for all of our experiments. We used an 80-20 split to split the dataset into training and validation datasets with a random seed of 42 using sklearn's (Pedregosa et al., 2011) train_test_split functionality. For the hyperparameters, we have used the default hyperparameters of MMF. For subtask 1, we have reported macro F1 score and for subtask 2, we have reported weighted F1.

5 Results

Table 1 contains all the results of our experiment. It can be noted that we mostly used the default hyperparameters from the MMF framework and did not perform rigorous hyperparameter tuning for our experiments, so the model performances still be improved with the search for optimal hyperparameters using cross-validation. We analyze the results of the approaches tried for each subtask. Among the baselines, we saw the unimodal-text models perform better for both the subtasks than the unimodal-image models. Among the unimodal-text models, BERT performed better than RoBERTa. The fusion models performed a bit better than the unimodal models for both the subtasks, with BERT Concatenation performing significantly better than late fusion in the first subtask. The multimodal transformers models give a performance increase over the fusion models, with VilBERT performing the best for Subtask 1 and VisualBERT giving the best performance for the second subtask. Among the multimodal transformer architectures, the ones pretrained with multimodal objectives turned out to be better in performance than those trained using unimodal objectives. Given that VisualBERT and VilBERT performed the best, we further pretrained them in a multimodal task-adaptive setting. VilBERT pretrained on the HatefulMemes dataset gave the best results for the first subtask, whereas VisualBERT pretrained on the HatefulMemes dataset was our best model for the second subtask.

6 Conclusion

In our worked, we have provided a comparative analysis of architectures for solving the task of misogyny identification. Although our best-performing model did achieve a substantial improvement over the baselines, the scores still indicate scope for further improvement. This also brings forth the challenging nature of the task in itself. Furthermore, using ensemble models like Vilio (Muennighoff, 2020) can result in better-performing models which can be tried out in future scope.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2020. Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5810–5820, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mary E. Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *NLDB*.

Manuela Barreto and Naomi Ellemers. 2005. The perils of political correctness: Men's and women's responses to old-fashioned and modern sexist views. *Social Psychology Quarterly*, 68(1):75–88.

Benoît Dardenne, Muriel Dumont, and Thierry. Bollier. 2007. Insidious dangers of benevolent sexism: consequences for women's performance. *Journal of personality and social psychology*, 93 5:764–79.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Sweta Karlekar and Mohit Bansal. 2018. SafeCity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811, Brussels, Belgium. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv*, abs/2005.04790.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *ArXiv*, abs/2012.07788.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in

Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2020. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

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.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2016. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*.