LastResort at SemEval-2022 Task 5: Towards Misogyny Identification using Visual Linguistic Model Ensembles And Task-Specific Pretraining

Samyak Agrawal **Radhika Mamidi**
International Institute of Information Technology Hyderabad
samyak.agrawal@research.iiit.ac.in,
radhika.mamidi@iiit.ac.in

Abstract

In current times, memes have become one of the most popular mediums to share jokes and information with the masses over the internet. Memes can also be used as tools to spread hatred and target women through degrading content disguised as humour. The task, Multimedia Automatic Misogyny Identification (MAMI), is to detect misogyny in these memes. This task is further divided into two sub-tasks: (A) Misogynous meme identification, where a meme should be categorized either as misogynous or not misogynous and (B) Categorizing these misogynous memes into potential overlapping subcategories. In this paper, we propose models leveraging task-specific pretraining with transfer learning on Visual Linguistic models. With our best performing models, we were able to achieve rank 5th and 10th on sub-tasks A and B respectively.

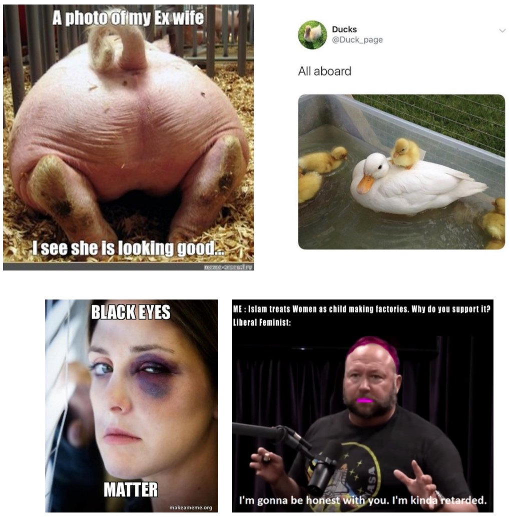

A photo of my Ex wife
I see she is looking good..
Ducks
@Duck_page
All aboard
BLACK EYES
MATTER
makeameme.org
ME : Islam treats Women as child making factories. Why do you support it?
Liberal Feminist:
I'm gonna be honest with you. I'm kinda retarded.

Figure 1: Example memes from the dataset showing the multimodal nature

1 Introduction

The term "misogyny" means hatred towards women. Misogyny can be interpreted through multiple forms such as male privilege, sexual harassment, violence against women, objectification. Memes that targeted women focus on appearance, intellect, their traditional gender roles and capabilities of women (Siddiqi et al., 2018).

For this, SemEval 2022 Task 5 (Fersini et al., 2022) focuses on identifying such behaviour in a multimodal setting (text + image). The textual cues to this task are given in the English language. The task is divided into two sub-tasks. The first subtask is modelled as a binary classification problem. The second sub-task focuses on identifying type of misogyny from a set of overlapping categories, making it a multi-label classification problem.

A meme contains text superimposed on an image. The image's aim in a meme is generally to reinforce a technique in the text, thus making its classification a multimodal problem. Both the modes of information are crucial to establishing the message conveyed by the meme, which can be very different from when the two modalities are evaluated separately.

We experiment with Visual Linguistic (VL) Models like OSCAR (Li et al., 2020) and UNITER (Chen et al., 2020) to understand the memes through both modalities. We employ transfer learning to use a model trained on another similar dataset and then finetune it on our dataset.

As task-specific pretraining has shown to improve results on several NLP tasks (Gururangan et al. (2020)), We experiment with task-specific pretraining our VL models before finetuning it and also finetune it on models task-specifically pretrained for other similar task like hateful memes detection (Kiela et al. (2021)).

We also train BERT (Devlin et al., 2019) based models on only the textual data, thus comparing the performances of multimodal setting vs unimodal settings. This comparison helps us understand how vital each modality is and how much using both together makes a difference.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 575 - 580
July 14-15, 2022 ©2022 Association for Computational Linguistics

We discover that even though detecting misogyny in memes can be modelled as a multi-modal task, it can, to a very good extent be done through working with just the textual cues but when it comes to detecting more subtle forms of misogyny, the visual cues play an important role as well. Our system ranked 38th and 19th for sub-taskA and sub-taskB respectively.

The paper is structured as follows: Section 2 describes the dataset along with related work. Section 3 describes our system and model architecture. Section 4 has information regarding the dataset size and splits with libraries used to implement our system. Section 5 has the discussion about the findings from our experiments and section 6 concludes our paper.

2 Background

Nowadays, the internet and various social media platforms have become an intrinsic part of more and more people's lives. With its growth, the problems associated with it have also increased exceedingly, like the increase in hate speech against certain groups including women.

Detecting misogyny and sexist slurs in general over social media can be challenging as its overall meaning can depend on its context and the user it is shown. (Fasoli et al., 2015). For this, look at the few examples in Figure 1 to exhibit the importance of visual and textual cues.

Memes can be defined as an image, video, or text, typically humorous in nature, that is copied and spread rapidly by internet users, often with slight variations. Memes in online culture have been seen to push potential instances of misogyny as a form of "joke" and "irony" while disguising itself as a harmless form of humour. (Drakett et al., 2018).

There has been previous work done to detect hate speech and misogyny. (Pamungkas et al., 2018) Employed Support Vector Machine(SVM) based architectures with a novel lexicon of abusive words to detect misogyny in English and Spanish tweets. (Gasparini et al., 2018) compared unimodal textual classifiers to multimodal classifiers trained with both visual and textual features using early fusion on a dataset of advertisements consisting of image and text marked for being sexist.

The meme classification task is primarily a Visual-linguistic(VL) task where we are trying to classify data where the image can be semantically correlated with the text. Traditional VL approaches are based on primary fusion techniques like early or late fusion, where each modality is learned separately. However, a multimodal pretrained model might perform better at memes classification (Afridi et al., 2020).

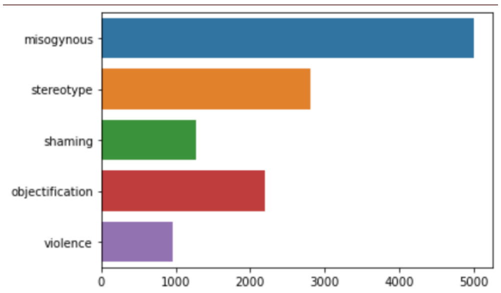


Figure 2: Data Distribution of labels in the training set.

Dataset Description

The dataset (Fersini et al., 2022) contains 10,000 memes.It is furthere divided into train and dev splits. Both tasks require the same dataset, but each task's final labels are different. Half of the 10,000 data points are marked positive and half negative. Of these half marked positive, the data is further annotated for potential overlapping categories of misogyny, namely: stereotype, shaming, objectification and violence.

3 System Overview

We use transformer based models for both the tasks with task specific modifications.

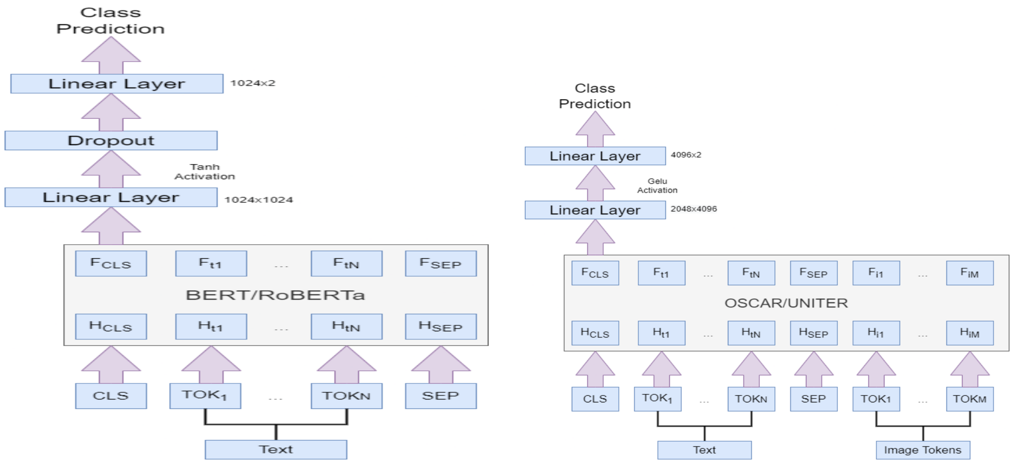
3.1 Pre Processing

Our text is tokenized into subwords to lookup the embedding. For our images, features were extracted using Faster-RCNN (Ren et al., 2016) pretrained on the VisualGenome dataset(Krishna et al., 2017) trained with and & without object attributes (Anderson et al., 2018).We extract features with object attributes of fixed box sizes 36(OSCAR36) and 50(OSCAR50) and features without object attributes of fixed box size 50 (OSCARV50).

The final input embeddings is a concatenation of both textual and image features represented as

$$h_{[CLS]}, h_{t_1}, \cdots, h_{t_n}, h_{[SEP]}, h_{i_1}, \cdots, h_{i_m}$$

Here $h_{[CLS]}$ and $h_{[SEP]}$ are the vector representations of the special [CLS] and [SEP] tokens respectively. $h_{t_1}, \cdots, h_{t_n}$ represents the text embeddings and $h_{i_1}, \cdots, h_{i_m}$ represents the vision embeddings.



(a) Text Only Models

(b) VL Models

Figure 3: Proposed architectures

3.2 Task-specific Pretraining

For both the tasks, We experiment with task-specific pretraining. Every task-specifically pre-trained models were pretrained on two pretraining objectives, namely Masked language Modelling (MLM) and Image Text Matching (ITM). We also make use of models trained on the hateful-memes dataset (Kiela et al., 2021). We use checkpoints from models that were:

1. Task-specifically pretrained on our dataset.
2. Task-specifically pretrained on hateful memes dataset.
3. Task-specifically pretrained and finetuned on hateful memes dataset.

The checkpoints for models pre-trained, fine-tuned on hateful memes dataset were taken from the vilio repository. [1]

3.3 Sub-task A

We used OSCAR as our primary VL model, we also experiemnt with another VL model named UNITER.The UNITER and OSCAR pre-trained weights are based on the BERT transformer. We used Binary Cross Entropy as our loss function to train our models. We trained 3 separate models on the 3 different visual features extracted but use the same textual features. We also experimented with ensembling these models using simple average as our ensembling technique.

We also train transformer-based models like BERT and RoBERTa (Liu et al., 2019) using just the textual cues. We use Binary Cross entropy as our loss function to train our models.

We use the CLS token embeddings from our transformer models and apply classification on top of it. The complete architecture for both text only and VL models can be seen in Figure 3.

3.4 Sub-task B

Here, instead of treating this problem as a multilabel classification problem, we treat it as a binary classification problem just like sub-taskA. We train VL models separately for each of the four labels, namely stereotype, shaming, objectification and violence, rather than training a single model for all labels. We also use an ensemble of models trained on different visual features like we did for sub-taskA.

For our textual models, we trained BERT-based multilabel classification models. We use cross-entropy loss to train our models. Since there is a significant class imbalance, we add weights to our positive data samples while calculating the loss function as done by researchers at (Gupta et al., 2021).The formula is given below:

$$
\begin{aligned}
\ell(\mathbf{x}, \mathbf{y}) &= -\frac{1}{Nd} \sum_{n=1}^{N} \sum_{k=1}^{d} \left[p^k y_n^k \log x_n^k + (1 - y_n^k) \log(1 - x_n^k) \right] \\
p^k &= \frac{1}{f^k} (|K| - f^k)
\end{aligned}
\tag{1}
$$

Where N is the batch size, n index denotes n^{th} batch element, d is the number of classes, f stands for a vector of class absolute frequencies calculated on the train set, $\mathbf{x}$ is the output vector from the last sigmoid layer, $\mathbf{y}$ is a vector of multi-hot encoded ground truth labels and $|K|$ is the size of the train

[1] vilio-repository,https://github.com/Muennighoff/vilio/blob/master/SCORE_REPRO.md

set.

4 Experimental setup

Parameter	Text Only	VL
Dropout	0.3	-
BatchSize	8	4
Epochs	5	3
Learning Rate	1e-05	1e-05
Warmup	-	0.1
Optimizer	Adam	AdamW

Table 1: Hyperparameters

The dataset contained 10,000 images along with the corresponding texts. Half of the data is marked positive for being misogynous. 85% of the dataset was used to train the model, and the rest was used to validate the model for both subtasks.

We use the VL model implementations of OSCAR and UNITER from the library vilio and for image feature extraction. [2]. We use huggingface [3] library for our transformers trained on just text.

The information about the hyperparameters can be found in Table 1. All models were trained on a GeForce RTX 2080 Ti GPU.

4.1 Evaluation Metrics

We use f1-macro scores as our primary evaluation metrics for both the tasks. We also calculate the accuracy scores for both tasks.

Model	Accuracy	F1-Macro
RoBERTa$_{\text{large}}$	68.4	68.3
BERT$_{\text{large}}$	64.7	63.7
OSCAR$_{\text{ens}}$	68.7	67.2
OSCAR $_{\text{pretrained\_ens}}$	69.5	67.8
OSCAR $_{\text{hm\_pretrained\_ens}}$	**70**	**68.5**
OSCAR $_{\text{hm\_finteuned\_ens}}$	59.9	59.3
UNITER $_{\text{ens}}$	65.8	63.3
OSCAR + UNITER $_{\text{ens}}$	68.1	66.5
OCSAR36	69.5	67.9
OSCAR50	68.2	66.3
OSCARV50	67.1	64.7

Table 2: Results: Sub-TaskA

5 Results And Discussion

The detailed results from all our experiments conducted can be seen in Table 2 and 3.

We here use the F1 macro scores to judge our models. For subtaskA, We see that OSCAR ensemble models, task-specifically pre-trained on hateful memes dataset perform the best. Another interesting thing to notice is the textual only RoBERTa large model performs almost as good as our best performing VL model and better than all other VL models and is significantly better than BERT large.

We also see that simple average ensemble models for OSCAR perform better than each of its constituent models, and using transfer learning methods with model fine-tuned on hateful memes dataset performed unexpectedly worse. It means that even though hateful memes detection and detecting misogyny in memes are closely related in their idea, they are still not necessarily similar to predict.

In sub-taskB, we see that the ensemble of models with task-specific pretraining on our dataset worked the best and slightly better than the ensemble with task-specific pretraining on the hateful memes dataset. We also see that our OSCAR VL models worked significantly better here than text-only models like BERT and RoBERTa, which is unexpected since the text-only models worked very well compared to VL models in sub-taskA.

Model	Accuracy	F1-Macro
BERT $_{\text{large}}$	31.9	45.8
RoBERTa $_{\text{large}}$	35.8	45.7
OSCAR $_{\text{ens}}$	41.4	**52.6**
OSCAR $_{\text{hm\_pretrained\_ens}}$	42.3	52
OSCAR $_{\text{pretrained\_ens}}$	**45.5**	31.3
RoBERTa $_{\text{large\_misogynous\_labels}}$	12.5	41

Table 3: Results: Sub-TaskB

As we observe that BERT Based models give comparable, and in the case of RoBERTa, better performance than almost all the VL models, it indicates that detecting misogyny might not be an utterly multimodal problem, and just the textual cues are enough in identifying the misogyny.

We also observe that even though text-only models performed very well on misogyny detection, they performed poorly on more fine grained classification tasks, showcasing that the visual cues mattered as well to figure out the subtleties in the classification of the type of misogyny.

[2] vilio,https://github.com/Muennighoff/vilio

[3] Transformers,v4.16.2,https://huggingface.co/docs/transformers/index

We also trained both textual, and VL models on just the data points marked for misogyny as those are the only ones where at least one of the subcategories of misogyny will be marked positively. However, in this case, the models performed much more poorly. It is because they are not trained on examples that are not misogynous in nature and thus perform poorly on them in the test dataset.

The scores according to the official metrics for our best performing unimodal and multimodal models were as follows: Sub-taskA: RoBERTa $_{\text{large}}$: 68.3; OSCAR $_{\text{hm\_pretrained\_ens}}$: 68.6; Sub-taskB: RoBERTa $_{\text{large}}$: 63.6; OSCAR $_{\text{hm\_pretrained\_ens}}$: 69.1

6 Conclusion

In this paper, our experiments indicate that although misogyny detection in memes is designed as a multi-modal setting, the textual cues also perform very well and, in some instances, better than Visual Linguistic models. We also found out that when it comes to detecting more subtle forms of misogyny, visual cues seem to help in the classification task and perform better than transformer models with just textual cues. More work can be done to improve the results. Future work like experimenting with more upcoming VL models, employing better techniques to address the class imbalance, and using more advanced ensembling techniques like Rank Averaging, Power Averaging & Simplex Optimization can improve results.

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2020. A multimodal memes classification: A survey and open research issues.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jessica Drakett, Bridgette Rickett, Katy Day, and Kate Milnes. 2018. Old jokes, new media – online sexism and constructions of gender in internet memes. *Feminism & Psychology*, 28(1):109–127.

Fabio Fasoli, Andrea Carnaghi, and Maria Paola Paladino. 2015. Social acceptability of sexist derogatory and sexist objectifying slurs across contexts. *Language Sciences*, 52:98–107. Slurs.

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.

Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. 2018. Multimodal classification of sexist advertisements.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1075–1081, Online. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

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.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

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.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Nasrina Siddiqi, Anjuman Bains, Arbaaz Mushtaq, and Sheema Aleem. 2018. Analyzing threads of sexism in new age humour: A content analysis of internet memes. *Indian Journal of Social Research*, 59(3):355–367.