INF-UFRGS at SemEval-2022 Task 5: analyzing the performance of multimodal models

Gustavo A. Lorentz
Inst. of Informatics / UFRGS – Brazil
galorentz@inf.ufrgs.br

Viviane P. Moreira
Inst. of Informatics / UFRGS – Brazil
viviane@inf.ufrgs.br

Abstract

This paper describes INF-UFRGS submission for SemEval-2022 Task 5 Multimodal Automatic Misogyny Identification (MAMI). Unprecedented levels of harassment came with the ever-growing internet usage as a means of worldwide communication. The goal of MAMI is to improve the quality of existing methods for misogyny identification, many of which require dedicated personnel, hence the need for automation. We experimented with five existing models, including ViLBERT and VisualBERT - both uni and multimodally pretrained - and MMBT. The datasets consist of memes with captions in English. The results show that all models achieved Macro-F1 scores above 0.64. ViLBERT was the best performer with a score of 0.698.

1 Introduction

Social media and anonymity enable the spread of hateful speech, which explains why misogyny is prevalent and abundant on the internet. Not only is it present, but also increasingly so, as confirmed by Farrell et al. (2019). The platforms that contribute to the sharing of hateful content dedicate a considerable amount of human effort in detecting, analyzing, and eventually removing these contents. The task is demanding due to the nature of the posts, which are frequently not straightforward – they often contain irony and slang. Additionally, the textual information needed to automatically classify a post as misogynistic might be part of an image, in the form of a meme. That prevents sexist posts from being immediately detected by algorithms that rely solely on textual input.

In this paper, we describe the training and usage of five different multimodal models applied to detecting misogynistic memes in the scope of SemEval-2022 Task 5 Multimodal Automatic Misogyny Identification (MAMI) (Fersini et al., 2022). We explain the distinction between the models and compare their performances in light of differences in pretraining (unimodal or multimodal).

Among the five models, the one which achieved the highest score was ViLBERT, reaching the 32^{nd} position on the leaderboard (out of 83 participants), with a score of 0.698. The one which performed the worst was MMBT-Grid, with a score of 0.649.

The remainder of this paper is organized as follows: Section 2 covers background and related work. Section 3 presents an overview of our system. The experimental setup is described in Section 4. Section 5 presents our results. Then, Section 6 concludes the paper.

2 Background and Related Work

One crucial aspect of this task is the multimodality of inputs. Most of the time, a meme requires both textual and visual information to be correctly understood. Not only because the punch line usually comes in written form, but also because texts and images often contradict each other for humouristic purposes. Take for example Figure 1. The text alone indicates a positive feeling towards an object that makes sandwiches. The image, if one would remove the caption, would show a woman. But when taken into consideration simultaneously, it is a sexist meme implying that women exist to make men sandwiches.

We took part only in Subtask A, in which the goal of the model is to take a meme such as the one in Figure 1 as input and indicate whether it is misogynistic or not.

The Hateful Memes Challenge (Kiela et al., 2021) is similar to MAMI since both address hateful multimodal contents. Participants in the Hateful Memes Challenge received a dataset of memes with visual as well as textual inputs and had to predict whether the memes were hateful.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 695 - 699

Figure 1: Example of a meme from the MAMI dataset

MMF (Singh et al., 2020) is a multimodal framework from Facebook AI Research and it implements state-of-the-art visual and language models, such as VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019), M4C (Hu et al., 2020), and Pythia (Jiang et al., 2018), among others. MMF provides code and model implementations for The Hateful Memes Challenge. Their work served as the primary inspiration for our experiments, in which we apply many of the same models to the Multimedia Automatic Misogyny Identification (MAMI) dataset.

3 System Overview

We used MMF (Singh et al., 2020) to train five models on the MAMI dataset. These models can be briefly described as follows.

1. **MMBT-Grid** is a supervised multimodal bitransformer that jointly finetunes unimodally pretrained text and image encoders by projecting image embeddings to text token space. Its inputs are the concatenation of textual embeddings and the final activations of a ResNet after pooling – the downsampling of dimensions – and positional and segment encodings. The final activations are transformed so that they fit the dimensions of the transformers' hidden layers.

2. **ViLBERT and ViLBERT CC** ViLBERT (Lu et al., 2019) consist of two parallel models, one that operates over visual inputs, and another that operates over textual inputs. Both models operate similarly to BERT, *i.e.*, they are a series of transformer blocks. The difference lies in the *Co-attentional Transformer Layers* introduced by the researchers. During the attention calculation, they compute the usual Q, K, and V matrices. However, the textual K and V are passed to the visual multi-headed attention block, and the visual K and V are passed to the textual multi-headed attention block. The rest of the transformer operations proceed normally, causing multi-modal features since each modality pays attention to the other.

3. **VisualBERT and VisualBERT COCO** - VisualBERT extends BERT by modifying the input it processes. Making use of features extracted from Object Proposals – a set of image regions likely to contain objects – the model can capture the interaction between text and image. The model does that by treating these features as usual BERT input tokens, appending them to the textual tokens. That is, VisualBERT uses the self-attention mechanism to align textual and visual elements implicitly.

Two versions of ViLBERT and VisualBERT were used. The distinction between these two versions lies not in the architecture, but rather in how they were pretrained. The multimodally pretrained versions, ViLBERT CC and VisualBERT COCO, are the official ones published by Lu et al. (2019) and Li et al. (2019), respectively. The unimodally pretrained versions are, as explained by Kiela et al. (2021), *multimodal models that were unimodally pretrained (where, for example, a pretrained BERT model and a pretrained ResNet model are combined in some way).*

4 Experimental Setup

The dataset used to train all models was the one provided by the organization team. The data has not been augmented or modified in any way. Training data consists of 10,000 memes, trial data has 100 memes, and validation data has 1,000 memes.

The MMF framework comes with implementations of state-of-the-art models, preconfigured

with hyperparameters. In our experiments, the default configurations of the models were used. All models share the same values for the main configurations, such as 1e-5 for learning rate, 22,000 for maximum number of steps, 128 tokens at most for text processing and, due to memory limitations, one hyperparameter was set to a fixed value, that is, models used batch sizes of 16 for training. Hyperparameter optimization could, therefore, be applied to the models to obtain better results. The configuration files are open to inspection and change, given that they are publicly available at Facebook Research's Github repository[1]. The specific commit that was used for this work is available here[2]. The training process of all models, excluding MMBT-Grid, uses image features, which were not included with the dataset supplied by the organization team. We relied on a script included in MMF to extract these features, using ResNet-152.

The main evaluation metric used in the task and during training is macro-F1. Here we also report True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) Rates.

5 Results

In this section, we report on our experimental results organized around four questions.

5.1 What are the best and worst models?

The results obtained by each model can be seen in Table 1. The best and worst-performing models were, respectively, ViLBERT and MMBT-Grid, with macro-F1 scores of 0.698 and 0.649. With this score, ViLBERT ranked 32^{nd} on Subtask A. It is worth pointing out that they had very similar values for TP-rate. MMBT-Grid achieved a value of 0.866, despite being the worst-ranked among all five models. That means it had a good performance in identifying misogynistic memes. The problem is evidenced by the TN and FP rates. MMBT-Grid was the worst at classifying memes that are not misogynistic, with a TN-rate of 0.463, the lowest of all. It also has the highest FP rate of 0.537. Analyzing ViLBERT's metrics, we can see that what guaranteed it the first place among the five models was the TP and FN rate, which were, respectively, the highest and the lowest. VisualBERT COCO was the best at correctly classifying the negative class (TN rate = 0.581), but it also had, by far, the highest FN rate (0.21).

The differences in performance can not be explained by the usage of uni or multimodal pretraining. This is evidenced by the similarity between scores obtained by unimodally pretrained models (VisualBERT and ViLBERT) and that by multimodally pretrained models (VisualBERT COCO and ViLBERT CC). Additionally, the mentioned models share the same architecture (ViLBERT with ViLBERT CC and VisualBERT with VisualBERT COCO), and so it can not be the explanation for the differences in performance. However, what seems to have impacted scores the most is the use of image features during training, since MMBT-Grid performs the worst.

5.2 Do multimodally pretrained models perform better?

It is interesting to notice that there was no great difference in performance between unimodally and multimodally pretrained models, such as VisualBERT vs. VisualBERT COCO and ViLBERT vs. ViLBERT CC. This finding is in line with Kiela et al. (2021), who worked on the Hateful Memes dataset. Nevertheless, while multimodally pretrained models were slightly better on Hateful Memes, here the unimodally pretrained version of ViLBERT yielded slightly better results, but the difference was not statistically significant (according to a Wilcoxon signed-rank test).

5.3 Can combining classifiers improve classification performance?

To answer this question we analyzed the predictions of the five models for each instance on the evaluation dataset. Our results have shown that:

- 86.89% of the instances were correctly predicted by at least one model;
- 77.58% of the instances were correctly predicted by at least two models;
- 69.67% of the instances were correctly predicted by at least three models;
- 61.76% of the instances were correctly predicted by at least four models;
- 47.95% of the instances were correctly predicted by all models.

[1] https://github.com/facebookresearch/mmf

[2] https://github.com/facebookresearch/mmf/tree/d31f8776f3bee53e7be722cb6d6c7ecf0827cc30/mmf/configs

Model	Macro-F1	TP	TN	FP	FN
MMBT-Grid	0.649	0.866	0.463	0.537	0.134
VisualBERT	0.666	0.874	0.483	0.517	0.126
VisualBERT COCO	0.679	0.786	0.581	0.419	0.214
ViLBERT CC	0.697	0.836	0.571	0.429	0.164
ViLBERT	0.698	0.874	0.541	0.459	0.126

Table 1: Macro-F1 scores, true positive and negative rates, and false positives and negative rates for our models

This analysis suggests that, if we were to use a simple majority voting system to determine the predicted label for images, the obtained accuracy score would be 69.67%, which does not surpass the score achieved by ViLBERT alone. Additionally, we tried combining the predictions of the classifiers by averaging their output probabilities. Similar to what we found with majority voting, there were no performance improvements in relation to ViLBERT on its own.

5.4 How correlated are the models?

Table 2 shows the Pearson correlation coefficient calculated for all pairs of models to measure their level of agreement, *i.e.,* how many images they classified with the same label. We can see that ViLBERT and ViLBERT CC have the highest correlation coefficient, 0.78. We initially supposed that the reason for their high similarity was that they share the same architecture, but further analysis showed that VisualBERT and VisualBERT COCO, the other models that also share architectures, have low similarities. Therefore, the initial hypothesis was wrong and we can assert that the reason for the difference in similarity resides in the pretraining modality, since that is the only distinction between the models. We see that MMBT-Grid and ViLBERT have a correlation score of 0.61, while the lowest score is between MMBT-Grid and VisualBERT, 0.56. The fact that all correlation scores can be classified between strong and moderate explains why there were no gains in combining the models in an ensemble.

5.5 Is there any pattern in memes that were erroneously classified?

We analyzed images that were wrongly classified by all five models. They were, in total, 131 images. Through visual inspection, we were able to identify a pattern in the captions. We noticed that most false positives contained words like "girl", "girls", "woman", and "women", while false negatives did not present these words. To confirm this, we examined the frequency of these words in training and test datasets. The term "girl" appeared in approximately 4.57% of not misogynous memes in the training dataset, and in 6.37% of misogynous memes, that is, 1.39 times more often. This proportion, however, is almost reversed in the test dataset, in which the term appears in 11.1% of *not* misogynous memes, and only in 7.1% of misogynous memes, that is, 1.56 times *less* frequently. This might explain the high number of wrong classifications for memes that contain this word. For the term "women", training dataset analysis shows that 8.27% of misogynous memes had this word, while appearing in only 2% of not misogynous memes, about 4.13 times less often, while in the test dataset, 5.8% of misogynous memes had it, and 2.4% of not misogynous memes, that is, 2 times less. The change in word frequency for this term might also have contributed to misclassification.

6 Conclusion

In this paper, we described our submission to SemEval-2022 Task 5. Using Hateful Memes and MMF as inspiration, we wanted to replicate their methods in a similar context. Although hateful and misogynistic memes share some overlap, there are important distinctions between them, regarding different vocabulary, context, and targets (*i.e.,* hate can be directed towards anyone, while misogyny cannot).

In our experiments, we trained five models and confirmed that they reach similar performances in this dataset as they do in Hateful Memes. Our best model, ViLBERT, reached a F1 score of 0.698 and ranked 32^{nd} out of 83 on the leaderboard. We showed that using a majority voting system with all models would not be beneficial. The models could be further improved by hyper-parameter

	ViLBERT CC	VisualBERT	VisualBERT COCO	MMBT-Grid	ViLBERT
ViLBERT CC	1.00	0.66	0.58	0.61	0.78
VisualBERT		1.00	0.57	0.56	0.69
VisualBERT COCO			1.00	0.58	0.59
MMBT-Grid				1.00	0.61
ViLBERT					1.00

Table 2: Pearson correlation for each pair of models

tuning. We could also have experimented with late/early fusion, which, as suggested by Hateful Memes (Kiela et al., 2021), has an impact on performance, and we leave this as future work.

Acknowledgments. This work was partially supported by CAPES Finance Code 001.

References

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 87–96, New York, NY, USA. Association for Computing Machinery.

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.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0.1: the winning entry to the vqa challenge 2018.

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.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

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.