Mitra Behzadi at SemEval-2022 Task 5 : Multimedia Automatic Misogyny Identification method based on CLIP

Mitra Behzadi and **Ali Derakhshan** and **Ian G. Harris**
Department of Computer Science
University of California, Irvine
Irvine, California, USA
mbehzadi@uci.edu , aderakh1@uci.edu , harris@ics.uci.edu

Abstract

Everyday more users are using memes on social media platforms to convey a message with text and image combined. Although there are many fun and harmless memes being created and posted, there are also ones that are hateful and offensive to particular groups of people. In this article present a novel approach based on the CLIP (Radford et al., 2021) network to detect misogynous memes and find out the types of misogyny in that meme. We participated in Task A and Task B of the Multimedia Automatic Misogyny Identification (MaMi) challenge (Fersini et al., 2022) and our best scores are 0.694 and 0.681 respectively.

1 Introduction

In the past few years more and more people have been using memes on social media platforms to express their thoughts and sometimes their beliefs. Although there are countless memes that are humorous and fun without expressing hate towards any certain group of people, there are also memes that aim to attack people.

Misogynistic memes use a combination visual and textual content to put down women and some of them are so violent that can be triggering to previous sexual abuse victims. That is why it would be so useful if these instances could be identified automatically.

One of the main challenges of this problem is that the images of these memes come in various forms. Also, the text on the memes adds occlusion to the objects in the images which makes the image understanding part of the problem even more challenging.

To address these challenges, we present a novel multi-modal classification approach based on CLIP (Radford et al., 2021) to identify misogynistic memes and also determine which 4 subcategories of shaming, stereotype, objectification or violence they belong to. CLIP is a multi-modal network trained for object detection. We use a multi-label classification method and detect misogynistic memes and all the subcategories using one single pipeline.

2 Related Work

There have been numerous researches conducted to solve hateful speech detection online in the past decade. Many of these approaches such as (Samghabadi et al., 2020) and (Cao et al., 2020) focus on only one modality which is text. More recently, researchers such as (Gomez et al., 2020) gathered multi-modal data-sets to be able to detect those instances of hate that go undetected by using only textual context.

Memes are also a form of multi-modal data that are widely used to indirectly communicate some meaning online. For the past few years, researchers have tried to solve the problem of hateful meme detection in various ways. In (Suryawanshi et al., 2020) a small data-set was gathered and an approach based on VGG16 (Simonyan and Zisserman, 2014) neural network was proposed. In 2020 Facebook gathered a the Hateful Meme data-set with 10,000 instances which was part of a Hateful Meme Detection challenge (Kiela et al., 2020). The winner of that challenge (Zhu, 2020) used an ensemble of visual-linguistic models such as VisualBert (Li et al., 2019) and Ernie-Vil (Yu et al., 2020) and fine-tuned them to solve the problem.

3 Proposed Method

3.1 Data Description

The Multimedia Automatic Misogyny Identification (MaMi) data-set which was gathered for SemEval Task5 challenge (Fersini et al., 2022), consists of 10,000 instances for training phase and 1000 instances for test. An example meme from this dataset is shown in Figure 1. Each instance has 5 binary labels depicting if it is misogyny, shaming,

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 724 - 727

stereotype, objectification and violence.

It is important to note that these subcategories are not mutually exclusive, that is more than one label can be 1 for each instance. So we are dealing with a multi-label classification problem.

Label	Training Data	Test Data
Non-Misogynous	5000	500
Misogynous	5000	500
Shaming	1274	146
Stereotype	2810	350
Objectification	2202	348
Violence	958	153

Table 1: Training and Test Data-set Distributions

The challenge (Fersini et al., 2022) has two parts, in Task A the goal in to detect the misogynous memes. In the second part, Task B, the goal is to determine the type of misogyny that was present. We participated in both parts of the challenge using a multi-label classification scheme.

3.2 Classification Pipeline

In our proposed model, the pre-trained multi-modal object detection network CLIP has an important role. CLIP was initially introduced as multi-modal way of object detection. It was trained on 400,000 million pairs of images and text. It has proven to be much more efficient than many other state-of-the-art object detection techniques.(Radford et al., 2021)

We use CLIP to encode the image and text separately and concatenate the features as can be see in Figure 2 .The main idea is that there is a non-linear function $f(X_I, X_T)$ between the image feature space X_I and text feature space X_T that will help determine if an instance belongs to each one of the 5 categories or not. To find out the parameters of this non-linear function we create a feed-forward neural network and feed the concatenated features to that. The network has 5 output nodes, each for one of the labels.

Before feeding the image input to the CLIP image encoder we had to resize the image to 224x244 with 3 channels to match the input shape requirements. The text was also truncated to a sequence of length 77 to be seamlessly used with the CLIP tokenizer.

Additionally, we use Sigmoid layer as the activation function because in contrast to softmax, the probabilities of each instance belonging to a class do not have to sum up to 1 and so they can be independent of each other. Therefore, it is more suitable for multi-label classification. After getting the output of the pipeline, we determine the binary value for the predictions based on a threshold of 0.5.

3.3 Training Process

As no validation data-set was provided, we randomly split the 10,000 instances into 9000 for training and 1000 for validation purposes. We used binary cross-entropy loss function and used Adam optimizer for the process. The optimal hyper-parameters were found empirically with learning rate of 0.001, batch size of 128 and the training was done for 10 epochs.

At the end of each training epoch, the evaluation metrics including precision, recall and F1-score with macro averaging was calculated on the validation data-set. If a higher F1-score was found then the state of the model was saved as the best state.

As CLIP comes with different options for image feature extractions, we made sure to try two different ones, Visual Transformer (ViT) (Dosovitskiy et al., 2020) with 32x32 patches and Residual Network(He et al., 2016) with 101 layers to see if they have an impact on the classification results.

4 Evaluation and Results

4.1 Evaluation Metrics

All the submissions to the challenge were automatically evaluated by a script that was programmed by the organizers. After the challenge was over, the script was released and we investigated how the F1-scores for each task was calculated.

First the confusion matrix was calculated resulting in $M = \begin{pmatrix} tp & fp \\ fn & tn \end{pmatrix}$. Then, as shown in Equations 1 - 7, positive precision P^+, positive recall R^+, negative precision P^- and negative recall R^- was calculated separately and the final F1-score is the average of positive F1-score and negative F1-score.

$$P^+ = \frac{tp}{tp + fp} \tag{1}$$

$$R^+ = \frac{tp}{tp + fn} \tag{2}$$

$$F1Score^+ = \frac{2 \times (P^+ \times R^+)}{P^+ + R^+} \tag{3}$$

$$P^- = \frac{tn}{tn + fn} \tag{4}$$

Figure 1: Sample number 152 of Training data which is Misogynous and Stereotype

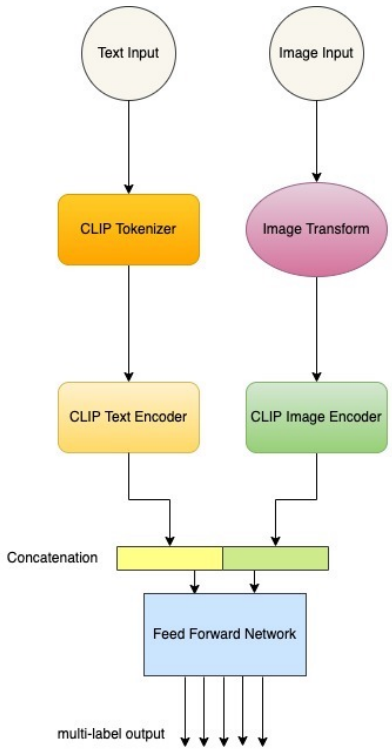

Figure 2: Proposed Classification Pipeline

$$R^{-} = \frac{tn}{tn + fp} \tag{5}$$

$$F1Score^{-} = \frac{2 \times (P^{-} \times R^{-})}{P^{-} + R^{-}} \tag{6}$$

$$F1Score = \frac{F1Score^{+} + F1Score^{-}}{2} \tag{7}$$

4.2 Experiments and Results

We conducted multiple experiments in the evaluation phase of the challenge. We tried different architectures for the feed forward network of our pipeline to get the best results.

In our initial experiments we had more layers in the feed forward network and did not include dropout layer and batch normalization. As the challenge progressed we realized that those models lacked in generalization, so we started using a simpler architecture and added dropout and batch normalization to get better results. Additionally, we tried switching the image encoder and discovered that using Resnet-101 based encoder results in much better scores, especially in Task B.

4.3 Error Analysis

As can be seen in the Table 2 our best result significantly outperforms all the baseline provided by the organizers. As informed by the organizers, the baseline models are grounded upon VGG-16 model for image feature extraction and USE model for textual feature extraction. Our best model achieves 0.694 score in Task A and 0.681 in Task B. It uses the pre-trained Resnet101 as the image encoder and also has 200 nodes in the hidden layer of our feed forward network.

5 Conclusion

In this paper we demonstrated how a successful model such as CLIP can be used to detect misogynous memes. We presented a novel architecture based on that multi-modal model and used multi-label training. We were able to achieve good results using this approach.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1813858. This research was also supported by a generous gift from the Herman P. & Sophia Taubman Foundation.

Our Models: Image Encoder - Feed Forward Network	**Task A F1-score**	**Task B F1-score**
ViT/32 - (1024,200,10,5)	0.678	0.497
ViT/32 - (1024,200,10,5) Dropout = 0.2	0.682	0.513
ViT/32 - (1024,200,5) Dropout = 0.2	0.681	0.629
ViT/32 - (1024,200,5) Dropout = 0.2 + Batch Norm	0.687	0.633
ViT/32 - (1024,100,5) Dropout = 0.2 + Batch Norm	0.674	0.639
ViT/32 - (1024,400,5) Dropout = 0.2 + Batch Norm	0.687	0.617
Resnet101 - (1024,200,5) Dropout = 0.2 + Batch Norm	**0.694**	**0.681**
Resnet101 - (1024,400,5) Dropout = 0.2 + Batch Norm	0.659	0.694
Resnet101 - (1024,100,5) Dropout = 0.2 + Batch Norm	0.693	0.673
Baseline Models		
Baseline Image	0.639	0.0
Baseline Text	0.640	0.0
Baseline Image-Text	0.543	0.0
Baseline Flat Multi-label	0.437	0.421
Baseline Hierarchical Multi-label	0.650	0.621

Table 2: Evaluation Results

References

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *12th ACM conference on web science*, pages 11–20.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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 preprint arXiv:2005.04790*.

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*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas Pykl, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using bert: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 1:12.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.