HateU at SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification

Aymé Arango, Jesus Perez-Martin, Arniel Labrada
Millennium Institute for Foundational on Data, Chile
University of Chile, Chile
{aarango,jperez,alabrada}@dcc.uchile.cl

Abstract

Hate speech expressions in social media are not limited to textual messages; they can appear in videos, images, or multimodal formats like memes. Existing work towards detecting such expressions has been conducted almost exclusively over textual content, and the analysis of pictures and videos has been very scarce. This paper describes our team proposal in the Multimedia Automatic Misogyny Identification (MAMI) task at SemEval 2022. The challenge consisted of identifying misogynous memes from a dataset where images and text transcriptions were provided. We reported a 71% of F-score using a multimodal system based on the CLIP model.

1 Introduction

Expressions of hate are common in online environments, and they can appear in different types of multimedia content (Bhattacharya et al., 2020). However, the related work on hate-speech and offensive language detection is primarily focused on textual English content (Agrawal and Awekar, 2018; Hosseinmardi et al., 2015). But, even for the English language, the task is still not solved. Evidence of that is recent reports of the increasing amount of hateful content in social media[1] following the occurrence of social or political events. Recent events like the COVID pandemic have brought a new wave of hate (Vishwamitra et al., 2020), with new targets and expressions including hateful memes (Pramanick et al., 2021). Therefore, the techniques for hate speech detection need to evolve towards new types of hate, representations, and languages.

The lack of generality of existing resources along with the emergence of new nets of hate makes current systems quickly outdated[2].

Most of the available datasets contain tweets (Waseem, 2016; Basile et al., 2019), Facebook and Youtube comments (Bosco et al., 2018) and, in general, textual content. Similar to The Hateful Memes Challenge [3] hosted by Facebook in 2020, The Multimedia Automatic Misogyny Identification (MAMI) challenge (Elisabetta Fersini, 2022) is an excellent opportunity for covering the hate speech detection task beyond written expressions.

In this competition, the organizers provided a training set of 10,000 memes labeled as hate speech in two different forms: binary (misogynous, not misogynous) and multi-class (stereotype, shaming, objectification, and violence). The competition comprises two tasks: Task A, for binary identification of misogyny, and Task B, for fine-grained classification of misogynous memes. For final system evaluation, the organizers published a set of 1000 extra unlabeled memes. Each meme in training and testing sets consists of an image with an overlay text. Each object in the dataset consists of an image and a transcription of the overlay text.

This paper describes our team participation in Task A of the MAMI challenge. We encoded images and texts using a pre-trained multi-modal model based on the CLIP model (Radford et al., 2021). We combined the encoded vectors in different ways to obtain a final classification output. Our best result reported was 71% of *f-score*.

In Section 2 we describe the work that has been done on hate speech detection using multi-modal content. In Section 3 we described the training dataset provided in the competition. Then, in Section 4 we describe our system and experiments. Our conclusions can be read in Section 6.

2 Background

Most of the research in hate speech detection has been conducted over textual datasets (Davidson

[1] https://www.channel4.com/news/george-floyd-death-has-led-to-increasing-online-hate-speech-report-claims

[2] https://whatsnewinpublishing.com/the-rise-of-hate-speech-and-what-the-media-can-do-about-it/

[3] https://www.drivendata.org/competitions/64/hateful-memes/page/205/

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 581 - 584
July 14-15, 2022 ©2022 Association for Computational Linguistics

et al., 2017; Agrawal and Awekar, 2018; Founta et al., 2018). Several strategies based on machine learning models and Natural Language Processing have been used to solve the task, though without success.

On the other hand, the identification of hate other than text formats of multimedia content has been treated only in a few works. Hosseinmardi et al. in 2015 and Singh et al. in 2017 have took advantages of the multi-modal information they could extract from *Instagram*[4] for they work on cyberbullying detection. While Perez-Martin et al. (2020) used the multi-modal representations for retrieving *Twitter* memes from textual queries.

Fortunately, in recent years the multi-modal detection of hateful content has gained popularity due to competitions like "*The Hateful Memes Challenge*" (Velioglu and Rose, 2020) hosted by *Facebook* where different models were proposed to detect hateful content on memes.

The proposed approaches encompass different visual state of the art models like VisualBert (Li et al., 2019), LXMERT (Tan and Bansal, 2019), VilBert (Lu et al., 2019) among others. The winning system combined some of these models with predefined rules (Zhong, 2020) for improving the classification accuracy of difficult samples.

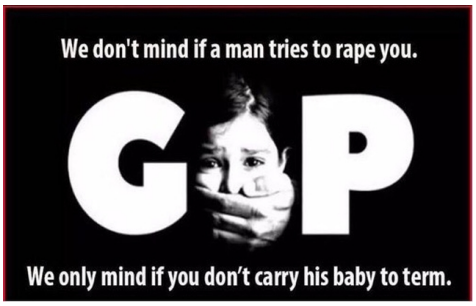
Another recent result on multimodal detection of offensive content has addressed the detection of harmful memes related to the COVID pandemic, also contributing with a new meme dataset (Pramanick et al., 2021).

There is much to do in the multimedia offensive language detection in images and video, considering the popularity of social networks like *Instagram* and *Tik Tok*[5].

3 Dataset Description

The dataset is composed of 10 000 memes, 5000 of which are labeled as *misogynous* and 5000 as *not misogynous*. For each meme is provided the corresponding image in *jpg* format and meme text transcription. All texts are in English; the most extensive text transcription found in the dataset contains 252 words, while the shortest contains one word. A characteristic of this dataset is that in some examples, only the text is enough for determining the nature of the comment (see Figure 1). We do not have evidence of an example where only the image would be necessary for identifying the nature of the meme. This characteristic may be detrimental to the multi-modal intention of the competition.

Figure 1: Meme example 17082. In this example only the texts is necessary for identifying the nature of the meme. The text transcription is: "We don't mind if a man tries to rape you. We only mind you don't carry his baby to term."



System	F-Score
Text_Only	69.23
Image_Only	65.37
CLIP_concat	70.50
CLIP_sum	**71.20**

Table 1: The results obtained in our experimentation. The details of each system is described in Section 4.

4 Experiments and results

Though we experimented with several models for texts and images, our best result was obtained using the CLIP model as the core of our system.

CLIP model: The CLIP model proposed by Radford et al. 2021 exploits the state-of-the-art textual and visual approaches for learning about images from texts. The general idea of the CLIP training strategy is to jointly learn image and text representations and predict the most similar pairs (image, text). According to the authors, the model can competitive transfer to different vision tasks.

CLIP based systems: We use a pre-trained CLIP model[6] for learning text ($text\_clip$) and image ($text\_clip$) representations from the texts transcriptions and images provided for the competition. We combined these outputs in different ways to obtained a vector x used as input for a classification final classification.

$$output = FFN(x)$$

[4]https://www.instagram.com/
[5]https://www.tiktok.com/
[6]https://github.com/OpenAI/CLIP

The different results can be found in Table 1.

Text_Only: Based on the characteristic of the dataset spotted in Section 3, we investigated if only the text transcriptions were enough for successfully detecting misogyny using this dataset.

$$x = text\_clip$$

With this system, we obtain a 69% of f-score after three epochs. This result is very close to our best result using both types of information (71%). One of the reasons could be the percentage of memes that can be classified by only using the texts, but more need to be studied to obtain a conclusive explanation.

Image_Only: Similar to the *Text_Only* system, we investigated if only the images were enough for successfully detecting misogyny using this dataset. The f-score obtained after five epochs is 65%, a lower result than the *Text_Only* system.

$$x = image\_clip$$

CLIP_concat: This system considered both image and text representations by concatenating them into a single vector in one single vector.

$$x = concat(image\_clip, text\_clip)$$

The results improved by using both representations to a 70% of f-score.

CLIP_sum_system: In this variant of the system, we sum both image and text representation in one single vector. This sum was pondered by a trainable parameter of the model a. The idea of this combination is to give the possibility to the model of using the necessary weights for image and text.

$$x = sum(a * image\_clip, (1 - a) * text\_clip)$$

With this combination we obtained our best reported result for the competition.

5 Error Analysis

We observed the memes miss classified by our best model (*CLIP_sum*). The most common type of error was the *false negative* error, examples wrongly classified as *not misogynist*, we noticed that most of them represent male figures or inanimate objects. Only a small number of memes picture a woman as the central figure (see Figure 2).

On the other hand, the female figures in the *false positive* examples is very common (see Figure 3).

Figure 2: Meme example 15115 from the testing set. Our model *CLIP_sum* wrongly classified it as a *not misogynist* meme. The text transcription is: "YOU DON'T WORK,COOK, CLEAN OR GIVE HEAD? LMAOBRUH LMAOBRUH.com LEGALLY, MY CLIENT IS ENTITLED TO A SIDE B*TCH OR TWO"

Figure 3: Meme example 15977 from the testing set. Our model *CLIP_sum* wrongly classified it as a *misogynist* meme. The text transcription is: "2020 BEFORE AND AFTER"

This phenomenon could be caused by a particular bias in the training set that relates misogyny memes with the images of women. But a deeper analysis has to be conducted in this regard.

6 Conclusions

This paper describes our team participation in Task 1 of the Multimedia Automatic Misogyny Identification (MAMI) at SemEval 2022. The purpose of this task was to identify memes as misogynists or not. Images and texts were provided in a training set of 10000 examples. Our team implemented a system based on the pre-trained CLIP approach and reported a 71% of f-score.

The multimodal hate speech detection has been under-addressed through the years and recently is

gaining popularity, though there is still much for research in this regard. Moreover, other types of multimedia, like videos, need to be analyzed since they are popular ways of communicating on social networks.

Acknowledgements

This material was supported by the Millennium Institute for Foundational Research on Data (IMFD).

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 141–153.

Valerio Basile, Cristina Bosco, Viviana Patti, Manuela Sanguinetti, Elisabetta Fersini, Debora Nozza, Francisco Rangel, and Paolo Rosso. 2019. Shared task on multilingual detection of hate. SemEval 2019, Task 5.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Giulia Rizzi Aurora Saibene Berta Chulvi Paolo Rosso Alyssa Lees Jeffrey Sorensen Elisabetta Fersini, Francesca Gasparini. 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.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *CoRR*, abs/1503.03909.

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*.

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.

Jesus Perez-Martin, Benjamin Bustos, and Magdalena Saldana. 2020. Semantic search of memes on twitter. *arXiv preprint arXiv:2002.01462*.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Akhtar, Preslav Nakov, Tanmoy Chakraborty, et al. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413*.

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.

Vivek K Singh, Souvick Ghosh, and Christin Jose. 2017. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2090–2099. ACM.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *CoRR*, abs/2012.12975.

Nishant Vishwamitra, Ruijia Roger Hu, Feng Luo, Long Cheng, Matthew Costello, and Yin Yang. 2020. On analyzing covid-19-related hate speech using bert attention. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 669–676. IEEE.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science, NLP+CSS@EMNLP 2016, Austin, TX, USA, November 5, 2016*, pages 138–142.

Xiayu Zhong. 2020. Classification of multimodal hate speech - the winning solution of hateful memes challenge. *CoRR*, abs/2012.01002.