TIB-VA at SemEval-2022 Task 5: A Multimodal Architecture for the Detection and Classification of Misogynous Memes

Sherzod Hakimov[1,2], **Gullal S. Cheema**[1], and **Ralph Ewerth**[1,2]
[1]TIB – Leibniz Information Centre for Science and Technology
[2]Leibniz University Hannover, L3S Research Center
Hannover, Germany
{sherzod.hakimov, gullal.cheema, ralph.ewerth}@tib.eu

Abstract

The detection of offensive, hateful content on social media is a challenging problem that affects many online users on a daily basis. Hateful content is often used to target a group of people based on ethnicity, gender, religion and other factors. The hate or contempt toward women has been increasing on social platforms. Misogynous content detection is especially challenging when textual and visual modalities are combined to form a single context, e.g., an overlay text embedded on top of an image, also known as *meme*. In this paper, we present a multimodal architecture that combines textual and visual features to detect misogynous memes. The proposed architecture is evaluated in the *SemEval-2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification* challenge under the team name *TIB-VA*. We obtained the best result in the *Task-B* where the challenge is to classify whether a given document is misogynous and further identify the following sub-classes: *shaming*, *stereotype*, *objectification*, and *violence*.

1 Introduction

Detection of hate speech has become a fundamental problem for many social media platforms such as Twitter, Facebook, and Instagram. There have been many efforts by the research community and companies to identify the applicability of advanced solutions. In general, hate speech is defined as *a hateful language targeted at a group or individuals based on specific characteristics such as religion, ethnicity, origin, sexual orientation, gender, physical appearance, disability or disease*. The hatred or contempt expressed towards women has been drastically increasing, as reported by Plan International (2020) and Vogels (2021). Detection of such misogynous content requires large-scale automatic solutions (Gasparini et al., 2018; Suryawanshi et al., 2020; Menini et al., 2020) and comprehensive annotation processes (Zeinert et al., 2021).

The detection of hateful content has been mainly studied from the textual perspective based on the Computational Linguistics and Natural Language Processing (NLP) fields. However, hateful content on social media can be found in other forms, such as videos, a combination of text and images, or emoticons. Misogynous content detection is especially challenging when textual and visual modalities are combined in a single context, e.g., an overlay text embedded on top of an image, also known as *meme*. Recent efforts in multimodal representation learning (Lu et al., 2019; Radford et al., 2021) pushed the boundaries of solving such problems by combining visual and textual representations of the given content. Several datasets have been proposed using multimodal data (Gomez et al., 2020; Kiela et al., 2020b; Sharma et al., 2020; Pramanick et al., 2021; Suryawanshi et al., 2020; Menini et al., 2020) for various tasks related to hate speech. Each dataset includes an image and corresponding text, which is either an overlay text embedded on an image or a separate accompanying text such as tweet text. In contrast to existing datasets based on memes (Kiela et al., 2020b; Sharma et al., 2020; Pramanick et al., 2021; Suryawanshi et al., 2020), the addressed task in this paper aims to identify misogyny in memes specifically. Among the previously mentioned work, only the dataset from Menini et al. (2020) is intended for misogyny detection, in which the text is in the Italian language. In terms of dataset size, the dataset by Gomez et al. (2020) contains approximately 150,000 image-text pairs, while other datasets have moderate sizes that range between two and ten thousand image-text pairs. Moreover, existing model architectures that are evaluated on such benchmark datasets use a combination of various textual and visual features extracted from pre-trained visual and textual models (Kiela et al., 2020a).

The *SemEval-2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification* (Fersini et al.,

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 756 - 760

Figure 1: Four data samples from *MAMI - Multimedia Automatic Misogyny Identification* with their corresponding class labels. Misogynous samples have additional sub-classes from *stereotype*, *shaming*, *objectification*, and *violence*.

2022)[1] is a new challenge dataset that focuses on identifying misogynous memes. The memes in this dataset are composed of an image with an overlay text. Some samples with their corresponding class labels are shown in Figure 1. The dataset includes two sub-tasks as described below.

- Task-A: a basic task about misogynous meme identification, where a meme should be categorized either as misogynous or not misogynous
- Task-B: an advanced task, where the type of misogyny should be recognized among potential overlapping categories such as stereotype, shaming, objectification and violence.

[1] https://competitions.codalab.org/competitions/34175

In this paper, we present our model architecture for which we submitted results under the team name *TIB-VA*. The model architecture is based on a neural model that uses pre-trained multimodal features to encode visual and textual content and combines them with an LSTM (Long-short Term Memory) layer. Our proposed solution obtained the best result (together with two other teams) on the *Task-B*.

The remainder of the paper is structured as follows. In Section 2, we describe the proposed model architecture. In Section 3, the experimental setup, dataset details, as well as evaluations of the model architecture are described in detail. Finally, Section 4 concludes the paper and outlines areas for future work.

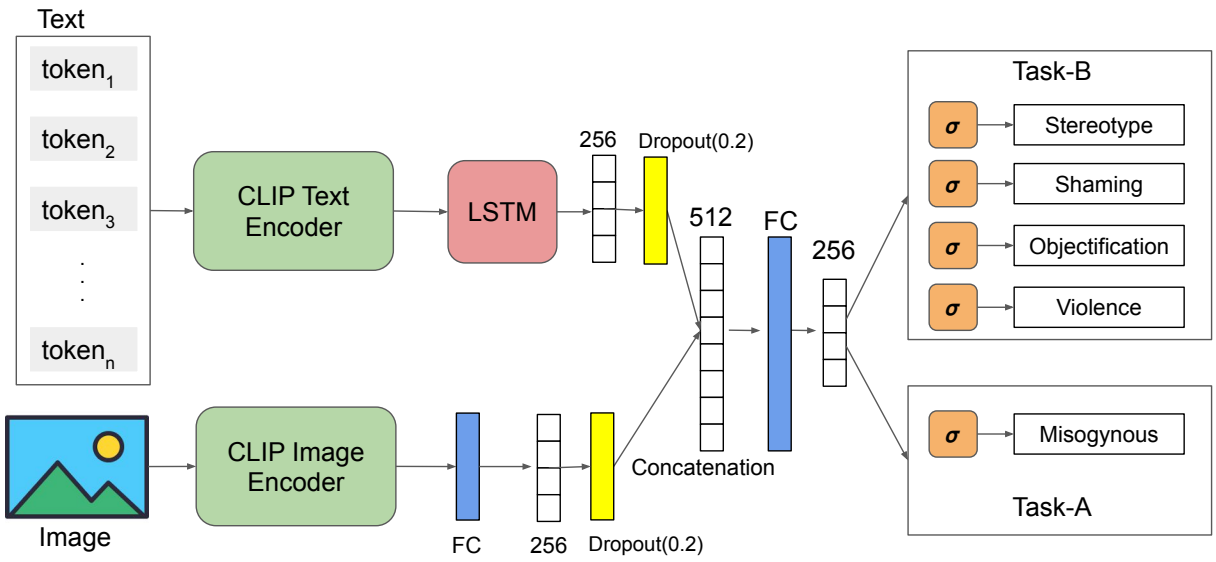

Figure 2: The model architecture that combines textual and visual features to output probabilities for Task-A (misogynous) and Task-B (stereotype, shaming, objectification, violence). FC: Fully connected layer, σ: sigmoid function.

2 Multimodal Architecture

Our model architecture is a neural model that uses a CLIP (Radford et al., 2021) model to extract textual and visual feature representations, which is pre-trained on over 400 million image-text pairs. The goal is to investigate whether this model is applicable to identify misogynous content in memes where both visual and textual content are considered. We used recently available *ViT-L/14* variant of CLIP. The tokens in the overlay text and the image are fed into *CLIP Text Encoder* and *CLIP Image Encoder* respectively. The text encoder outputs a sequence of 768-dimensional vectors for each input token. These token vectors are then fed into an LSTM layer with a size of 256. This layer is another essential part of the proposed architecture. It learns the contextual relatedness among tokens in the text by combining all token representations extracted from the CLIP text encoder branch. The output from the image encoder is fed into a fully-connected layer with a size of 256. The output from an LSTM layer for text and output from the fully-connected layer for the image are fed into separate dropout layers (dropout rate of 0.2), the outputs are concatenated, and then fed into another fully connected layer with a size of 256. The final vector representation is then fed into separate sigmoid functions for each task. For Task-A, the sigmoid outputs a single value that indicates the probability of misogyny. For Task-B, each sub-class of misogyny (stereotype, shaming, violence, objectification) has a separate sigmoid function that outputs a probability value for the corresponding class. The model architecture is shown in Figure 2. The source code of the described model is shared publicly with the community[2].

3 Experimental Setup and Results

3.1 Dataset

The *SemEval-2022 Task 5: MAMI - Multimedia Automatic Misogyny Identification* (Fersini et al., 2022) aims at identifying misogynous memes by taking into account both textual and visual content. Samples from the dataset are given in Figure 1. The dataset includes the overlay text extracted from an image. The challenge is composed of two sub-tasks. Task-A is about predicting whether a given meme is misogynous or not. Task-B requires models to identify sub-classes of misogyny (stereotype, shaming, violence, objectification) in cases where a given meme is misogynous. The samples in Task-B can have multiple labels where a meme can have a single or all of the above sub-classes of misogyny. The train and test splits have 10 000 and 1000 samples, respectively. The distribution of samples for the corresponding two sub-tasks is given in Table 1.

[2] https://github.com/TIBHannover/multimodal-misogyny-detection-mami-2022

Splits	Task-A		Task-B				Total
	Misogynous	NOT	Shaming	Objectification	Violence	Stereotype	
Train	5000	5000	1274	2202	953	2810	10 000
Test	500	500	146	348	153	350	1000

Table 1: Distribution of samples in Task-A and Task-B for train and test splits in the *MAMI - Multimedia Automatic Misogyny Identification* dataset.

3.2 Experimental Setup

Training Process: The model architecture is trained using Adam optimizer (Kingma and Ba, 2015) with a learning rate of *1e-4*, a batch size of *64* for maximum of *20* epochs. We decrease the learning by half after every five epochs. We use 10% of the training split for validation to find the optimal hyper-parameters.

Implementation: The model architecture is implemented in Python using the PyTorch library.

Team	Task-A	Task-B
Ours (TIB-VA)	0.734	**0.731**
SRC-B	**0.834**	**0.731**
PAFC	0.755	**0.731**
DD-TIG	0.794	0.728
NLPros	0.771	0.720
R2D2	0.757	0.690

Table 2: Experimental results for the selected top-performing teams on the *MAMI* dataset. The results on Task-A and Task-B are Macro-F1 and Weighted F1 measures, respectively.

3.3 Results

The official evaluation results[3] for the top-performing teams are presented in Table 2. The results on Task-A and Task-B are macro-averaged F1 and weighted-average F1 measures, respectively. Our model architecture (team *TIB-VA*) achieves the best result (0.731) on Task-B along with other two teams: *SRC-B* and *PAFC*. The *SRC-B* team has the highest performance on Task-A. Our results on Task-A are ten points below the best result from the team *SRC-B*. Despite this gap in Task-A, our result is still among the top 20 percentile of all submitted results.

4 Conclusion

In this paper, we have presented a multimodal model architecture that uses image and text features to detect misogynous memes. The proposed solution is built on the pre-trained CLIP model to extract features for encoding textual and visual content. While the presented solution does not yield top results on Task-A, it achieves the best performance in Task-B for identifying sub-classes of misogyny such as stereotype, shaming, objectification, and violence. In future work, we will explore the combination of multiple multimodal features that measure different aspects of visual content such as violence, nudity or specific objects and scene-specific content.

[3] https://competitions.codalab.org/competitions/34175#results

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 812997 (CLEOPATRA ITN).

References

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. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018 - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS, Porto, Portugal, July 26-28, 2018*, pages 565–572. SciTePress.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1459–1467. IEEE.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Casey A. Fitzpatrick, Peter Bull, Greg Lipstein, Tony Nelli, Ron Zhu, Niklas Muennighoff, Riza Velioglu, Jewgeni Rose,

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, Helen Yannakoudakis, Vlad Sandulescu, Umut Ozertem, Patrick Pantel, Lucia Specia, and Devi Parikh. 2020a. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track, 6-12 December 2020, Virtual Event / Vancouver, BC, Canada*, volume 133 of *Proceedings of Machine Learning Research*, pages 344–360. PMLR.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2020. A multimodal dataset of images and text to study abusive language. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Plan International. 2020. Free to be online? Online; accessed February 2022.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2783–2796. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 759–773. International Committee for Computational Linguistics.

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, TRAC@LREC 2020, Marseille, France, May 2020*, pages 32–41. European Language Resources Association (ELRA).

Emily A. Vogels. 2021. The State of Online Harassment. Pew Research Center. Online; accessed February 2022.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3181–3197. Association for Computational Linguistics.