UniBO at SemEval-2022 Task 5: A Multimodal bi-Transformer Approach to the Binary and Fine-grained Identification of Misogyny in Memes

Arianna Muti and **Katerina Korre** and **Alberto Barrón-Cedeño**
Department of Interpreting and Translation
Alma Mater Studiorum–Università di Bologna
Forlì, Italy
{arianna.muti2,aikaterini.korre2,a.barron}@unibo.it

Abstract

We present our submission to SemEval 2022 Task 5 on Multimedia Automatic Misogyny Identification. We address the two tasks: Task A consists of identifying whether a meme is misogynous. If so, Task B attempts to identify its kind among shaming, stereotyping, objectification, and violence. Our approach combines a BERT Transformer with CLIP for the textual and visual representations. Both textual and visual encoders are fused in an early-fusion fashion through a Multimodal Bidirectional Transformer with unimodally pretrained components. Our official submissions obtain macro-averaged F_1=0.727 in Task A (4th position out of 69 participants) and weighted F_1=0.710 in Task B (4th position out of 42 participants).

1 Introduction

Evolving from the two previous editions of the Automatic Misogyny Identification initiatives (Fersini et al., 2018, 2020), the Multimedia Automatic Misogyny Identification shared task at SemEval 2022 (MAMI) targets multimodal data for the first time. Within MAMI, Fersini et al. (2022) propose two classification tasks:

Task A A basic task about misogynous meme identification, where a meme should be categorized either as misogynous or not.

Task B An advanced task, where the type of misogyny should be recognized among potentially overlapping categories: stereotyping, shaming, objectification and violence.

The increasing volume of meme posts on social media renders the development of models able to identify malicious instances timely. The task is more challenging than when dealing with text alone because, in general, both the textual and the visual channels play an indivisible role in conveying the desired message.[1]

We build upon our previous experience in identifying misogyny and aggressiveness in text (Muti and Barrón-Cedeño, 2020) and approach both multimodal tasks with a supervised multi-modal bi-transformer model (MMBT) (Kiela et al., 2020a). We use *bert-base-uncased-hatexplain* (Mathew et al., 2020) and *bert-base-uncased* (Devlin et al., 2019) for the textual embeddings, and CLIP (Radford et al., 2021) for the visual ones. We also build two unimodal baselines to compare against.[2]

Our experiments aim at understanding if and to what extent our multimodal model outperforms the two unimodal ones that address the problem separately. Since meme classification is a challenging task due to its multimodal nature, we shed some light on which component should weigh more in the decision process —text or image— by observing the impact of both modalities in the predictions.

Our official submission for Task A achieved a macro-averaged F_1 score of 0.727, whereas that for Task B obtained a weighted F_1 score of 0.710, positioning our team in the 4th position in the task, for both tasks.

In addition, we identify the linguistic and visual elements which are potentially responsible for the misclassification. We perform an error analysis to check whether misclassified memes rely heavily on external world knowledge and/or are subtle and implicit, as we believe that those two aspects cause struggle to the models.

The rest of the paper is structured as follows. Section 2 provides essential definitions for this task, such as misogyny and memes. We then move on to dataset description as well as a summary of related work. Section 3 describes our models for both tasks.

[1] This is different from other multimodal scenarios, such as visual question answering or image captioning, where one of the two modalities tends to be the dominant one (Zhu, 2020).

[2] Our model is publicly available at https://github.com/TinfFoil/Unibo-at-SemEval-2022-MAMI.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 663 - 672

	train	test
Not Misogynous	5,000	500
Misogynous	5,000	500

Table 1: Class distribution for the binary Task A misogynous or not.

Section 4 describes the experimental setup. Our results are presented and discussed in Section 5. Section 6 presents our error analysis. Section 7 concludes with a summary of our findings.

2 Background

2.1 Definitions

Memes are relatable acts of communication made of visual and textual artifacts, where often an image is superimposed with text with a humorous purpose (MacDonald and Wiens, 2022). To be fully understood, memes require context and real-world knowledge. They are often satirical, implying humour and sarcasm in a subtle way (Sharma and Pulabaigari, 2020). These factors cause the identification of phenomena in them —such as expressions of misogyny— difficult.

Humour does not always come as harmless fun and that is the case with misogynous memes. Such memes contribute to the establishment of a rape culture (Ridgeway, 2014), where violence and sexual harassment are tolerated, belittled, normalized, excused and transformed into jokes. Therefore, developing automatic approaches to tackle misogyny has both technological and social value.

According to the MAMI guidelines (Fersini et al., 2022), a meme is misogynous when it conveys an offensive, sexist or hateful message (be it weak or strong, implicitly or explicitly) targeting a woman or a group of women. Four kinds of misogyny are considered for this task:

Shaming occurs when memes insult or offend women because of their physical aspect.

Stereotyping represents a fixed idea or set of characteristics; physically or ideological.

Objectification represents a woman like an object through the over-appreciation of her physical appeal (sexual objectification) or by depicting her as an object (a human being without any value as a person).

Violence shows physical or verbal violence toward women.

	train	test	preds
Shaming	1,274	146	130
Stereotype	2,810	350	379
Objectification	2,202	348	334
Violence	953	153	102

Shaming	Stereotyping	Objectification	Violence	train	test	preds
☑				400	24	32
	☑			1,247	32	152
		☑		992	37	96
			☑	250	19	21
☑	☑			286	32	20
☑		☑		161	25	40
☑			☑	11	2	0
	☑	☑		412	152	118
	☑		☑	302	40	31
		☑	☑	116	38	23
☑	☑	☑		301	45	32
☑	☑		☑	55	3	1
☑		☑	☑	12	5	0
	☑	☑	☑	162	36	20
☑	☑	☑	☑	45	10	5
			Total	4,752*	500	591

* 248 of the misogynous memes lack type annotation.

Table 2: Number of instances per class for the multi-label Task B (top). Class distribution (bottom). Column **preds** shows the predictions of our best submitted model (Multi; cf. Section 5).

2.2 Datasets

The datasets are balanced with respect to Task A (see Table 1). The same instances include the multi-label annotation for Task B. Table 2 shows the number of instances for each label combination.

Stereotyping is the most represented class, with $3.2k$ instances overall, followed by objectification ($2.2k$) and shaming ($1.2k$); violence is the least represented, with less than $1k$. The label overlapping plays an important role in our results analysis (c.f., Section 6). As Table 2 shows, stereotyping and objectification tend to come together, whereas shaming and violence do not. We will explore whether our models capture this intersection in Section 6.

2.3 Related Work

The identification of misogyny in textual form was explored in the series of shared tasks on Automatic Misogyny Identification (AMI), held at IberEval (Anzovino et al., 2018) and EVALITA (Fersini et al., 2018, 2020). AMI at IberEval 2018 focused on identifying misogyny on English and Spanish tweets and on classifying the misogynous tweets into seven categories: discredit, stereotype, objectification, sexual harassment, threats of violence, dominance, and derailing. AMI at EVALITA 2018 targeted the analysis of Italian and English tweets. Task A addressed misogyny identification as well. Task B aimed at recognizing whether the target of a misogynous post was a specific person or women in general, and at classifying the positive instances in the aforementioned categories. AMI at EVALITA 2020 targeted the detection of misogyny and aggressiveness in Italian tweets (Task A) and the identification of unintended bias (Task B).

Multimodality has been explored for the automatic analysis of memes. Sharma and Pulabaigari (2020) worked in the task of identifying whether an image is a meme or not. Two recent SemEval tasks have targeted memes as well. Sharma et al. (2020) proposed an emotion identification task. The best performing system consisted of a text-only approach, a feed-forward neural network (FFNN) wth word2vec embeddings (Keswani et al., 2020). Dimitrov et al. (2021) proposed a shared task on the identification of propaganda techniques. Feng et al. (2021) approached it with a pre-trained transformer using text with visual features. They extracted grid features using ResNet50 (He et al., 2016) and salient region features using BUTD (Anderson et al., 2018). They further used these grid features to capture the high-level semantic information in the images. Moreover, they used salient region features to describe objects and to caption the event present in a meme. They built an ensemble of fine-tuned DeBERTA+ResNet, DeBERTA+BUTD, and ERNIEVIL (Yu et al., 2021) models.

Multimodal hate speech has attracted the interest of the research community only recently. In 2019, Facebook AI launched the Hateful Memes Challenge (Kiela et al., 2021b), which consisted in identifying hate speech in memes: hateful vs not. It is constructed such that unimodal models struggle and only multimodal models can succeed: difficult examples ("benign confounders") are added to the dataset to make it hard to rely on unimodal signals. The most successful approaches used both *early fusion* and *late fusion* (Kiela et al., 2021a), with the former achieving the best results. Those include VilBERT (Lu et al., 2019), VisualBERT (Li et al., 2019), MMF (Singh et al., 2020), MMBT (Kiela et al., 2020a) and CLIP (Radford et al., 2021). In late fusion approaches, systems for each modality are trained independently. The scores produced by each model are joined during inference to produce the final prediction (Kiela et al., 2020b).

In early fusion the different modalities are combined at an early stage to learn one single classification model (Kiela et al., 2020b). The top-performing model applied optical character recognition to find and remove the text from the input images in order to improve the quality of object detection, named entity identification and human race detection, using all these tags as input for different transformer models (Zhu, 2020).

Although MAMI is the first shared task on misogyny detection on memes, there has been preliminary work on automatic detection of sexist memes. Fersini et al. (2019) explored unimodal and multimodal approaches both with late and early fusion to understand the contribution of textual and visual cues on the MEME dataset, a dataset containing 800 sexist and not sexist memes. The sexist memes were also annotated according to aggressiveness and irony. From their work emerged that a unimodal textual model performs better than image-based ones. Concerning multimodality, late-fusion worked better.

3 System Overview

Our approach is based on the multimodal bi-transformer model (MMBT) (Kiela et al., 2020a). MMBT fuses image and text embeddings in an early fashion. MMBT jointly finetunes unimodally pretrained text and image encoders by projecting image embeddings to the text token space. Figure 1 represents the model architecture. MMBT combines two segments: segment 0 corresponds to the text, whereas segment 1 corresponds to the picture. They are fed together to use attention over both modalities at the same time. Each token is indexed according to its position from 0 to the maximum text length, which we set to 80. Each image representation is indexed from 0 to 640.

The original MMBT combines BERT (Devlin et al., 2019) and ResNet (He et al., 2016). We con-

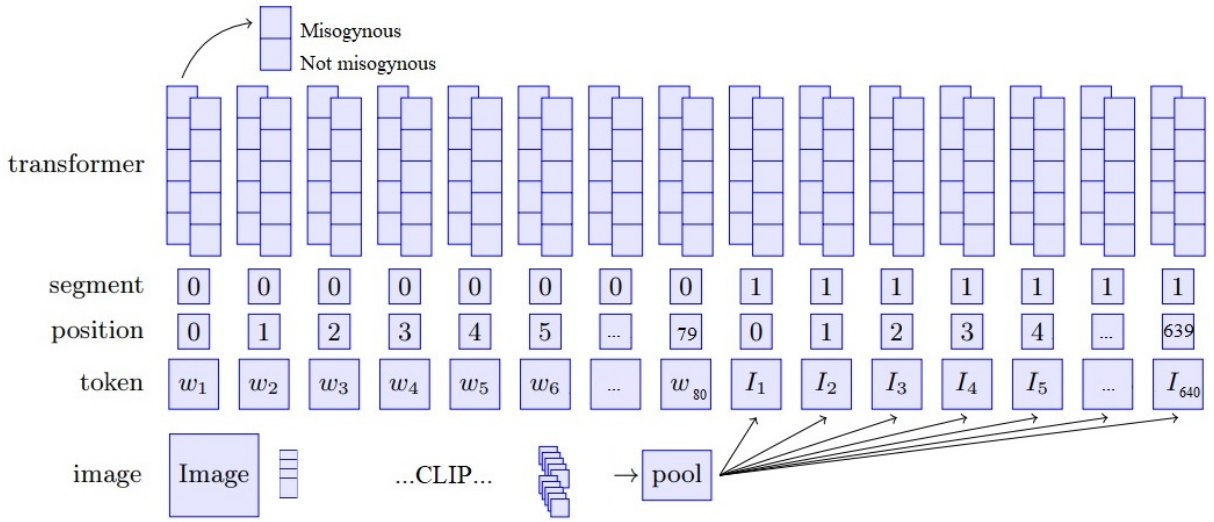

Figure 1: Representation of the MMBT model architecture combining CLIP and BERT; adapted from (Kiela et al., 2020a).

sider other models. For the textual embedding we tried with *bert-base-uncased-hatexplain* (Mathew et al., 2020), a version of BERT trained on identifying hate speech. For the image embedding we used CLIP, since it has outperformed all alternative models in a large variety of multimodal tasks, including OCR, action recognition in videos, geolocalization, and various types of fine-grained object classification (Radford et al., 2021). CLIP is pre-trained on the task of predicting which caption should be tied together with a given image. In this way it learns state-of-the-art image representations from scratch, enabling zero-shot transfer of the model to downstream tasks.

The two embeddings are fused through MMBT. For Task A we use the sigmoid activation function for the output layer and threshold at 0.5 to discriminate between misogynous or not. For Task B we adopt a binary relevance approach (Zhang et al., 2017), combining four binary classification models. The output for each classifier is a sigmoid function too. We opt for this approach after observing that treating the classes separately increased the performance in a multi-class model predicting misogynous, misogynous-aggressive or none (Muti and Barrón-Cedeño, 2020). This approach allows us to predict multiple mutually non-exclusive classes.

We apply a heuristic to refine the multi-label decisions in Task B. All four decisions are turned off if an instance had not been predicted as misogynous by our Task-A model.

Pre-processing Since CLIP requires square images, following Neskorozhenyi (2021) we produce 288×288 pixel versions of all memes. The memes come in different sizes and orientations, hence we rescale them until the largest side reaches 288 pixels respecting the aspect ratio and pad to fill the empty pixels in the square. Then, we slice the resized images into three equal parts horizontally if the image orientation is landscape and vertically if it is portrait to obtain both global and local image features. We extracted four vectors for each image: a vector for each part encoding spatial information and one for the whole image. We used the Pillow library (Clark, 2015) to perform these operations.

No preprocessing is applied to the text, other than applying the BertTokenizer (Devlin et al., 2019).

4 Experimental Setup

We shuffled the training set and take 10% of the data for development preserving the class distribution through stratified random sampling (Pedregosa et al., 2011).

We trained three alternative models to identify the best possible configuration. **Uni$_{\text{txt}}$** is a BERT-based unimodal system that considers the text alone. **Uni$_{\text{img}}$** is a CLIP-based unimodal system that considers the image alone. **Multi** is a multimodal system, fusing BERT and CLIP embeddings through MMBT.[3]

[3]We tried a variation of **Uni$_{\text{txt}}$** for Task A. We augmented the training data with the tweets corpus from AMI at Evalita

model	variation	macro F_1
$Multi_5$	after 5 epochs	0.703
$Multi_6$	after 6 epochs	**0.727**
Uni_{txt}	bert-base-uncased	0.656
Uni_{txt}	bert-Hatexplain	0.569
Uni_{img}	with fine tuning	0.703
Uni_{img}	zero-shot	0.417

Table 3: Official macro-averaged F_1-measures for our submissions to Task A.

model	masked with	weighted F_1
Multi	$Multi_5$	**0.710**
Multi	$Multi_6$	0.588
Uni_{txt}	Uni_{txt} bert-base-uncased	0.660

Table 4: Official weighted F_1-measures for our three submissions to Task B. Column *masked with* specifies the model from Task A used to mask the output labels.

Hyperparameters For Multi, we tried learning with different numbers of learning epochs, in range [3, 6]. The best validation performance was obtained after 5 epochs in Task A in the development set. We trained over 5 epochs in Task B. For both Uni_{txt}we train over 4 epochs. For the Uni_{img}we train over 5 epochs. In all cases we saved the model only when an increase in the performance was obtained. Since we also aimed at assessing how CLIP performs in making zero-shot predictions in this task, we used CLIP without fine-tuning on the training set, to check whether it could be effectively used to detect misogyny without prior annotation. We considered batch sizes of 16 and 32, with the former consistently performing better. We used a learning rate of 2e-4, the MADGRAD optimizer (Defazio and Jelassi, 2021), and a binary cross-entropy loss function.

The results reported in Section 5 are obtained with a model trained during 5 epochs for Task B and 6 epochs for Task A with 16 as the batch size.

Evaluation metrics We stick to the official MAMI evaluation metrics: macro-averaged F_1-measure for the binary Task A and weighted-averaged F_1-measure for the multi-label Task B.

5 Results

In this section we present the results obtained by our submissions to both Task A and Task B.

5.1 Task A

Table 3 shows the results of our submitted runs for Task A. The highest score is obtained after training the multimodal model Multi during 6 epochs: F_1=0.727. Considering the textual information alone runs short; the highest performance being obtained when the Uni_{txt} model is trained upon a *generic* BERT: 0.656. As expected, the zero-shot Uni_{img}model performs the worst, with a performance lower than that of a random model. A proper fine-tuning of the Uni_{img} model turns into the runner-up performance with F_1=0.703. The improvement of the Uni_{img}over the Uni_{txt}model by five points suggests that the visual information is captured better than the textual one. The reason might be that the text is too short and out of context to be captured effectively by BERT.

5.2 Task B

Table 4 shows the results of our submitted runs for Task B. In this case, we trained one single Multi model during 5 epochs. The difference between the two configurations is in the masking of the multi-label classification. The most successful multimodal model gets F_1=0.710, after masking with respect to Task A's $Multi_5$ model. Masking on the basis of Task A's $Multi_6$ model causes a performance drop of twelve points. $Multi_5$ predicts more misogynous instances than $Multi_6$ (678 vs 653). $Multi_6$ blacks out more predictions which are false positives and hence a potentially correct decision by the multi-label model gets ignored. The text-alone approach, masked by the corresponding Task A model, runs short by five points.

In Table 5 we zoom into the performance of our Task B Multi model for each of the four classes. The model struggles the most when trying to spot stereotyping and shaming. This reflects the nature of misogyny. Stereotyping and shaming tend to be less explicit, and hence harder to spot —even for human beings. On the contrary, violence, which is the most explicit, is more likely to be identified. Stereotyping is the class that has been over-predicted the most (cf. Table 2).

6 Qualitative Analysis

In this section we present a qualitative analysis of the results to further examine the strengths and weaknesses of our approach.

2018 (Fersini et al., 2018). Since no improvement was observed in the model, the results are neglected.

	prec	recall	F_1
Shaming	0.52	0.46	0.49
Stereotype	0.54	0.58	0.56
Objectification	0.69	0.66	0.67
Violence	0.73	0.48	0.58

Table 5: Per-class performance on the positive class for model Multi; our best submission to Task B.

	Uni_{txt}	Uni_{img}	Multi
false positives	0.20	0.23	0.21
false negatives	0.14	0.06	0.06
true positives	0.36	0.44	0.44
true negatives	0.30	0.27	0.29

Table 6: Error analysis across all models for Task A showing relative frequencies.

6.1 Analysis on Task A

To address the question of which component for detecting misogyny in multimodal settings is more important, we looked at the distribution of the kind of errors made by the different models, as well as the overlapping instances among the four categories. Table 6 shows the relative frequencies. The amount of false negatives is much lower than that of false positives across all models. Considering a practical application, false negatives could have a greater impact as they are the misogynous instances that could not be detected, and could therefore lead to harm. On the other hand, blocking instances that were not misogynous but were classified as such could be considered censorship.

Table 6 shows the prediction analysis of the best runs for each modality. Looking at each model individually, Uni_{txt} has less false positives but more false negatives than the other two models. Uni_{img} has the highest number of false positives, and the same number of false negatives as the Multi model. This means that the textual model performs worse than the others in capturing misogyny, while the visual one tends to overpredict misogyny more than the other two models.

Figure 2 shows the intersections and differences in both false positives and false negatives by the three models. There are more false positive than false negative instances across all models, as observed in Table 6. Indeed, the number of common false positives by all models is almost 4 times as high as the number of common false negative values. This indicates that the models tend towards

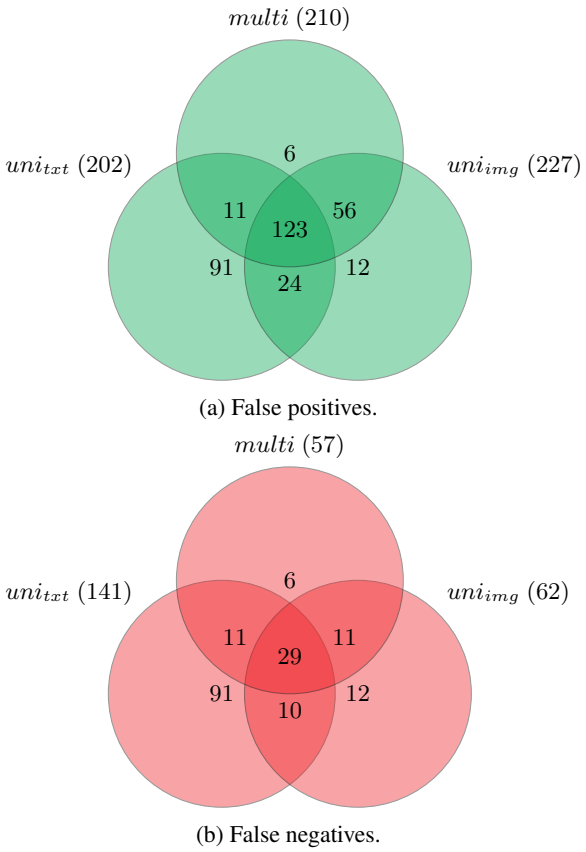

Figure 2: Venn diagrams representing the false positive and false negative errors by the three top Multi, Uni_{txt} and Uni_{img} models during the testing stage.

over-predicting misogyny. Taking into account the differences among the sets, Uni_{txt} accounts for the fewest false positive instances (Figure 2a), while it accounts for the most false negative instances (Figure 2b). Therefore, in this specific multimodal task, where we can be more lenient with false positives than false negatives, a textual model does not seem to be an optimal alternative.

Since the model does not allow for a great interpretability of the results, we performed a manual inspection of some interesting instances and the potential reasons behind the errors when classifying them. As Figure 2 shows, 132 instances are misclassified by all three models: 123 are false positives and 29 are false negatives. We observe the following trends after looking at the false negatives:

1. The level of misogyny is low or subjective, as the meme is not directly referred to women (e.g., Figure 3a) or misogyny is expressed in a subtle way (e.g., Figure 3b implies the stereotype that women are complicated);

2. Real-world knowledge is required to under-

(a) Instance 15846.

(b) Instance 16132.

(c) Instance 17028.

(d) Instance 16232.

Figure 3: Instances of Task A false negatives by all three models

stand the meme (Figure 3c can be better understood if we know Sarah Jessica Parker and the Twisted Sister band).

Figure 4: An example of false positive (instance 15094).

Figure 5: An example of meme properly labeled by text models only (instance 15802).

3. The stance of the text with respect to the image is relevant in order to convey the general meaning (see Figure 3d);

Among the false positives, memes mostly contain:

1. Compliments, which are often associated to objectification (e.g. Fig. 4).

2. Images or phrases that often occur in misogynous contents (e.g., women in underwear, kitchen-related terms).

3. Identity terms (e.g., *wife, women, girls*), that tend to co-occur with misogynous contents in the training set.

We also performed an analysis on memes that have been correctly classified by only one model. Among the instances that only the textual model got right, 11 were true positives and 56 true negatives. True positive cases mostly share a strong textual component in conveying misogyny, while the image is either irrelevant, or it is used only to make the sentence ironic (Fig. 5).

Among instances that only the visual model got right, both true positives and true negatives are 11.

Figure 6: An example of benevolent sexism that tends to confuse the classifier.

Most true positives have a explicit visual component. For instance, beaten women and texts justifying an aggression or glorifying violence. Among instances that only the multimodal model got right, 10 are true positive and 24 true negative. By observing the true positive instances, contrarily to what is expected, misogyny is not always conveyed by the integration of text and image, as in most of the cases the text is actually dominant.

6.2 Analysis on Task B

We performed a manual inspection focusing on the errors in predicting stereotyping and observed a relatively large amount of compliments toward women, which tend to confuse the classifier. In particular, false negatives are often caused by the presence of *benevolent sexism* (Glick and Fiske, 1997), which shows a subjectively positive attitude towards women that conceals inferiority compared to men, and it is often disguised as a compliment. Figure 6 shows an example.

Now we analyse the label overlaps to determine if our model captured the intersection of the classes. We compare our predictions to the gold labels in Table 2. The size of the intersection between stereotype and objectification is in the same order for gold and predictions: 152 vs 118. The intersection between cases of shaming and violence is practically null, which is well reflected in the model (2 vs 0). Less cases of both shaming and stereotyping than expected are identified (32 vs 20). The same applies to the combinations stereotyping–violence (40 vs 31) and objectification–violence (38 vs 23). The pair shaming–objectification tends to be overpredicted (25 vs 40).

7 Conclusions

We presented our participation to the Multimedia Automatic Misogyny Identification shared task. We addressed two problems: spotting whether a meme is misogynous and, if it is, what kind of misogyny it expresses. We compared unimodal models (text only and image only) with a multimodal model based on Multimodal bi-Transformers. Our image-only model performs better than the text-only one, suggesting that the visual information might be easier to capture than the textual one. Our multimodal approach performs the best in both tasks. The errors come from more false positives than false negatives.

From our error analysis we observed that stereotyping and shaming are the most misclassified categories. This proves that more focus on subtle and implicit forms of misogyny and sexism is needed.

Acknowledgements

A. Muti's research is carried out under project "DL4AMI–Deep Learning models for Automatic Misogyny Identification", in the framework of *Progetti di formazione per la ricerca: Big Data per una regione europea più ecologica, digitale e resiliente*—Alma Mater Studiorum–Università di Bologna, Ref. 2021-15854.

K. Korre's research is carried out under the project "RACHS: Rilevazione e Analisi Computazionale dell'Hate Speech in rete", in the framework of the PON programme FSE REACT-EU, Ref. DOT1303118.

References

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. *arXiv*.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Alex Clark. 2015. Pillow (pil fork) documentation.

Aaron Defazio and Samy Jelassi. 2021. Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization. *arXiv*.

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, MN. ACL.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Zhida Feng, Jiji Tang, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, and Li Chen. 2021. Alpha at SemEval-2021 task 6: Transformer based propaganda classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 99–104, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, and Silvia Corchs. 2019. Detecting sexist meme on the web: A study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on automatic misogyny identification (AMI). In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*, pages 59–66. Torino: Accademia University Press.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Peter Glick and Susan T. Fiske. 1997. Hostile and benevolent sexism: Measuring ambivalent sexist attitudes toward women. *Psychology of Women Quarterly*, 21(1):119–135.

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.

Vishal Keswani, Sakshi Singh, Suryansh Agarwal, and Ashutosh Modi. 2020. IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of Internet memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1135–1140, Barcelona (online). International Committee for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020a. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv preprint arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, and Aravind Mohan. 2020b. Hateful Memes Challenge and dataset for research on harmful multimodal content. *MetaAI*.

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. 2021a. The hateful memes challenge: Competition report. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, 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. 2021b. 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*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv preprint arXiv:1908.02265*.

Shana MacDonald and Brianna I. Wiens. 2022. Feminist memes: Digital communities, identity performance and resistance from the shadows. *Materializing Digital Futures: Touch, Movement, Sound and Vision*, page 123. Publisher: Bloomsbury Publishing USA.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289*.

Arianna Muti and Alberto Barrón-Cedeño. 2020. UniBO @ AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop 12-13 December 2018, Naples*.

Rostyslav Neskorozhenyi. 2021. How to get high score using MMBT and CLIP in Hateful Memes Competition. *Towards Data Science*.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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. *arXiv 2103.00020*.

Shannon Ridgeway. 2014. 25 everyday examples of rape culture. *Everyday Feminism*.

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 visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Chhavi Sharma and Viswanath Pulabaigari. 2020. A Curious Case of Meme Detection: An Investigative Study. In *International Conference on Web Information Systems and Technologies (WEBIST)*, pages 327–338.

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`.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph. *arXiv preprint arXiv:2006.16934*.

Min-Ling Zhang, Yukun Li, Xu-Ying Liu, and Xin Geng. 2017. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12:191–202.

Ron Zhu. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *CoRR*, abs/2012.08290.