UMUTeam at SemEval-2022 Task 5: Combining image and textual embeddings for multi-modal automatic misogyny identification

José Antonio García-Díaz
Facultad de Informática,
Universidad de Murcia,
Campus de Espinardo,
30100, Spain
joseantonio.garcia8@um.es

Camilo Caparros-Laiz
Facultad de Informática,
Universidad de Murcia,
Campus de Espinardo,
30100, Spain
camilo.caparrosl@um.es

Rafael Valencia-García
Facultad de Informática,
Universidad de Murcia,
Campus de Espinardo,
30100, Spain
valencia@um.es

Abstract

In this manuscript we describe the participation of the UMUTeam on the MAMI shared task proposed at SemEval 2022. This task is concerning the identification of misogynous content from a multi-modal perspective. Our participation is grounded on the combination of different feature sets within the same neural network. Specifically, we combine linguistic features with contextual transformers based on text (BERT) and images (BEiT). Besides, we also evaluate other ensemble learning strategies and the usage of non-contextual pretrained embeddings. Although our results are limited, we outperform all the baselines proposed, achieving position 36 in the binary classification task with a macro F1-score of 0.687, and position 28 in the multi-label task of misogynous categorisation, with an macro F1-score of 0.663.

1 Introduction

This manuscript describes the participation of the UMUTeam in the Multimedia Automatic Misogyny Identification (MAMI) shared task (Fersini et al., 2022), proposed at SemEval 2022. This shared-task consists in the identification and categorisation of misogynous content from a dataset composed of memes (Dawkins and Davis, 2017). A meme is essentially a pictorial content with an overlaying text that pretends to be funny. However, some of these memes are being used as a form of hate against women, with sexist messages in online social networks that amplify misogynous traits such as sexual stereotyping or gender inequality.

MAMI shared task proposes two challenges. A binary classification task, in which each meme should be labelled as misogynous or not misogynous, and a multi-label classification task, to categorise different misogynous traits, namely shaming, stereotype, objectification, and violence.

2 Background information

From the last years, the number of shared tasks in workshops regarding hate-speech and misogyny detection are increasing. To name just but a few, in Italian there is the EVALITA 2018 dataset (Bosco et al., 2018); in Spanish, the AMI 2018 dataset (Fersini et al., 2018) and the EXIST dataset (Rodríguez-Sánchez et al., 2021). In German, the GermEval 2021 (Risch et al., 2021) dataset.

The common approaches for misogyny detection and categorisation consists in the training of an automatic machine learning classifier. For example, the authors of (Anzovino et al., 2018) compiled and labelled a corpus from Twitter focused on misogynous content, and evaluate several feature sets and machine-learning models. In Spanish, a similar approach was conducted in (García-Díaz et al., 2021), in which the authors released the Spanish MisoCorpus 2020. This dataset is organised into three splits: (1) VARW (Violence Against Relevant Women), focused on aggressive messages on Twitter to women who have gained social relevance; (2) SELA (European Spanish vs that of Latin America), focused on distinguish between misogynistic messages from Spain and Latin America; and (3) DDSS (Discredit, Dominance, Sexual harassment and Stereotype), focused on general traits related to misogyny. The Spanish MisoCorpus 2020 is balanced and contains 3841 misogynous documents, annotated by three human annotators.

It is worth noting that our research group evaluated of a set of hand-crafted linguistic and negation features along with Spanish pre-trained contextual and non-contextual embeddings for detecting hate-speech (García-Díaz et al., 2022b). These work included two datasets concerning misogyny and sexist behaviour.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 742 - 747

3 Dataset

Table 1 depicts the dataset proposed by MAMI. For the training and validation split, the labels were balanced, with 5000 misogynist memes and 5000 safe memes that were manually annotated using crowd sourcing platforms. Our first experiments consider a balance between both labels. However, during the final evaluation phase we suspect that there was a strong imbalance among the labels. As our first results were limited, we sub sampled the dataset removing some misogynous documents with less than 8 words. We are aware that there are better techniques for handling class imbalance but, due to time constraints, we could not evaluate them.

Split	Original dataset	Subsampled
training	8000	6709
val	2000	1677
test	1000	1000
total	11000	9386

Table 1: Dataset statistics of the MAMI dataset. We show the original distribution (left) and our sub sampled distribution (right)

Table 2 depicts the label distribution per misogynous trait for the second challenge. It should be noticed that shaming and violence traits are the traits with less instances, hinder their detection.

Split	(OBJ)	(SHA)	(STE)	(VIO)
training	1762	1020	2248	763
val	440	254	562	190
total	2202	1274	2810	953

Table 2: Misogynous trait distribution: Objectification (OBJ), Shaming (SHA), Stereotype (STE), and Violence (VIO)

4 Methodology

For solving the challenges proposed in MAMI, we build a system which architecture is depicted in Figure 1. It a nutshell, our system works as follows. First, we select some of the documents of the training MAMI dataset to create a custom validation dataset. Next, we extract a subset of language-independent linguistic features (LF), non-contextual sentence (SE) and word embeddings (WE) from fastText, the contextual word embeddings from BERT (BF), and the image embeddings from BEiT (BI). Second, we train several neural network models by performing an hyperparameter tuning process. The models evaluated included one model per feature set, and models based on several feature sets together. Besides, we evaluate two ensembles based on soft voting (mode) and averaging all the probabilities (mean) of the neural networks trained with each feature set. To handle the multi-label challenge, we repeat this process per trait. That is, we evaluate the problem as a binary classification problem per trait.

Next, some insights of the feature sets involved are given. As the memes are images with overlaying text, this shared-task has a multi-modal perspective. Our proposal uses several feature sets based on texts, and one for the images. First, we use the UMUTextStats tool to obtain a set of relevant psycho-linguistic features (LF). This tool has already been used in studies related to misogyny, such as (García-Díaz et al., 2021, 2022a). The LF included low-level linguistic categories concerning phonetics and syntax's, and high-level features related to semantics and pragmatics, including features proper from figurative language (del Pilar Salas-Zárate et al., 2020). Moreover, these kinds of features have proven to be effective for performing other automatic classification tasks such as irony and satire identification (García-Díaz and Valencia-García, 2022). As some of the dictionaries of UMUTextStats are not translated to English, we select a subset of language-independent linguistic features, based on linguistic metrics, Part-of-Speech features and the usage of social media jargon. Second, we extract sentence and word embeddings using the pre-trained fastText model (Joulin et al., 2016). Third, we use contextual sentence embeddings from BERT (Devlin et al., 2018), which sentence embeddings are obtained in a similar manner as described at S-BERT (Reimers and Gurevych, 2019). Forth, we use visual embeddings from BEiT (Bao et al., 2021), which is a self-supervised model trained with ImageNet-21k with more than 21000 labels. BEiT learns the embeddings images as a sequence of fixed-size elements using relative position. This allows us to perform the classification using a mean-pooling strategy from the final hidden states of the patches instead of placing a linear layer on top of the final classification token. However, the suggested way to fine-tune the model for performing downstream tasks is to attach a new linear layer that uses the last hidden state of the

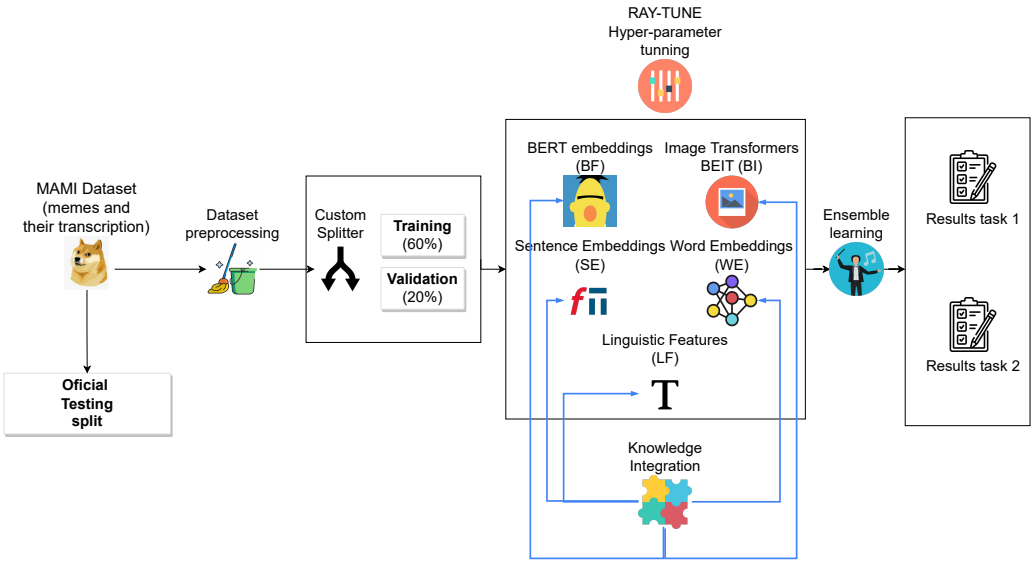

Figure 1: System architecture proposed by the UMUTeam for solving the MAMI shared task

classification token as a representation of the whole image.

To obtain the embeddings from BERT (for text) and BeIT (for images) we conduct an hyperparameter optimisation stage using RayTune (Bergstra et al., 2013) with a Tree of Parzen Estimators (TPE) to select the best combination of the hyperparameters over 10 trials. The hyperparameters evaluated and their interval range are: (1) weight decay (between 0 and .3), (2) training batch size ([8, 16]), (3) warm-up steps ([0, 250, 500, 1000]), (4) number of training epochs ([1-5]), and (5) learning rate (between 1e–5 and 5e–5).

Once all features are obtained, we train a neural network per feature set. The training of each neural network is performed with hyperparameter optimisation. Each training involved: (1) 20 shallow neural networks, that are multi-layer perceptrons (MLP) composed by one or two hidden layers with the same number of neurons per layer connected with one activation function (`linear`, `ReLU`, `sigmoid`, and `tanh`); (2) 5 deep-learning networks, that are MLP between 3 and 8 hidden layers, in which the neurons per layer are disposed for each layer in different shapes, namely brick, triangle, diamond, rhombus, and funnel, and connected with an activation function (`sigmoid`, `tanh`, `SELU` and `ELU`). The learning rate of the deep-learning models is `10e-03` or `10e-04`. Besides, on the neural networks with the pre-trained word embeddings from fastText (WE) we also evaluate 10 convolutional neural networks (CNN) and 10 bidirectional recurrent neural network layers (BiLSTM). In all experiments, we evaluate two batch sizes: `16` and `32`. These small values were selected because the training split was balanced, and a dropout mechanism (`[False, .1, .2, .3]`) for regularisation.

Apart of the neural networks trained with the feature sets separately, we evaluate different forms for combining the strengths of each feature set in the same system. The combination of the feature sets is performed using two strategies: (1) knowledge integration, in which each feature set is used as input of the same neural network. For this, we train another neural network repeating the hyperparameter optimisation stage; and (2) ensemble learning, in which the output of each neural network model trained with a feature set is combined by averaging the predictions or calculating the mode of the predictions.

5 Results and discussion

Each system was evaluated using the custom validation split (see Section 3).

The results for the first challenge are depicted in Table 3. Note that we remove some of the documents with fewer words to sub sample the dataset. For that reason, our validation split contains 677 misogynous and 1000 non-misogynous documents.

As it can be observed in Table 3, the performance of LF in isolation is limited. This fact is not surprising because not all of the features from UMUTextStats are available in English. For the rest of the textual embeddings (SE, WE, and BF), BF obtains the best results, with a macro F1-score of

Feature set	MIS	N-MIS	F1
LF	58.956	75.771	67.363
SE	73.286	80.860	77.073
WE	76.614	83.367	79.991
BF	79.690	87.306	83.498
BI	65.744	81.984	73.864
LF,BF,BI	**80.277**	**87.549**	**83.913**
LF,BF,BI (mode)	75.874	86.524	81.199
LF,BF,BI (mean)	76.853	85.183	81.018

Table 3: Results for the Task A with the custom validation split, reporting the F1-score of the misogynous (MIS) and non-documents (N-MIS) and the macro F1-score (F1).

83.498, outperforming the non-contextual sentence and word embeddings from fastText (SE and WE). Due of this, we decided to discard non-contextual embeddings and use BERT for the knowledge integration strategy. The combination of LF, BF, and BI in the same neural network outperformed the results achieved by BF, increasing slightly the F1-score of both labels and the macro f1-score. However, the ensemble learning strategy achieved lower results for the F1-score of the misogynous label, regardless of the strategy employed for combining the predictions.

For the second challenge, the results with the custom validation split are depicted in Table 4. We report the F1-score of each binary model that we train for each misogynous trait. Similar to the first challenge, the results achieved by LF are limited, achieving results below 60% in all the traits. Concerning non-contextual embeddings, the results with WE are superior to SE for all traits, but inferior compared to BF. In case of BI, it draws our attention the high macro F1-score achieved in objectification and shaming, outperforming BF. However, their results are inferior to BF for stereotype and violence. When LF is combined with BF and BI embeddings within the same neural network, the results are superior to the ones achieved separately, except in shaming. However, the results achieved with the ensemble learning strategy are quite limited, specially with the mean strategy. The result obtained with the violence trait is especially striking. We observe that the resulting model achieved a perfect recall over the violence class, which suggests that this model is always predicting all tweets as violence.

Table 5 depicts the official results for the first challenge. We achieve position 36/83 with a macro F1-score of 68.7. As it can be observed, our submission outperforms all the proposed baselines that consisted in: (1) sentence embeddings from the USE pre-trained model; (2) image features from VGG-16; (3) a combination of deep image and text representations based a shallow neural network with a single layer. In addition, two baselines focused on the second challenge were also evaluated: (4) a multi-label model, based on the concatenation of deep image and text representations, for predicting simultaneously if a meme is misogynous and the corresponding type; and (5) a hierarchical multi-label model, based on text representations, for predicting if a meme is misogynous or not and, if misogynous, the corresponding type.

The best result for the first challenge is achieved by the *SRCB_roc* team, with an F1-score of 83.4. It is worth mentioning that, although there was a restriction of the number of accounts available in Codalab per team and user, the organisers of the task are not able to control it. Nevertheless, in the official leader board the second best result was also achieved by the team *SRCB_roc*, with an F1-score of 81.1. However, as the team name is the same, we have removed this result from the table. Therefore, for the official results we ask to the reader to check the official results published in the overview of the task.

For this first challenge we send different runs and modify our strategy according to the results. We also send some basic results to obtain some baselines. For example, we achieved an macro F1-score of 52.85 with LF. This result is more limited than the ones achieved in the validation split. With non contextual embeddings, SE and WE, the results are, respectively, 61.30 and 61.96 (vs 77.073 and 79.99 with the validation split), and 64.75 for BF. Because of these results, we suspect that there are relevant different between the training and testing splits. Then, we examined carefully the testing split but we could not find relevant differences so we suspect to imbalanced as an possible explanation of this problem. To confirm this, we send a toy submission with a baseline consisting in all the predictions as non-misogyny and we observed a ratio similar to 1:3 between misogynous and non-misogynous instances. Then, we reduced the training dataset and retrained all models in order to make them stronger against class imbalance (see Section 3). It is worth noting that our methods

Feature set	Objectification	Shaming	Stereotype	Violence
LF	59.449	57.858	57.733	59.052
SE	66.983	66.479	70.923	66.730
WE	69.727	68.418	72.291	69.390
BF	72.849	68.452	73.785	67.815
BI	75.112	69.834	59.529	63.551
LF-BF-BI	**76.911**	68.957	**73.855**	**69.457**
LF,BF,BI (mode)	65.997	**70.046**	64.471	59.740
LF,BF,BI (mean)	45.465	66.330	45.577	8.680

Table 4: Results for the Task B with our custom validation split including the Macro F1-score for each misogynous trait

Rank	Team	F1-score
1	SRCB_roc	83.4
2	DD-TIG	79.4
3	Beantown	77.8
36	UMUTeam	68.7
58	Baseline 1	65.0
61	Baseline 2	64.0
61	Baseline 3	63.9
79	Baseline 4	54.3
83	Baseline 5	43.7

Table 5: Official results for the Task A

Rank	Team	F1-score
1	SRCB_roc	73.1
2	TIBVA	73.1
3	PAFC	73.1
28	UMUTeam	66.3
41	Baseline 1	62.1
48	Baseline 2	42.1

Table 6: Official results for the Task B

already consider some good practises concerning class imbalance, such as setting the initial bias, adding class weights to the model, heavily weight the few examples that are available

Finally, the official results for the second challange are depicted in Table 6. It can be observed that our best result achieved a 66.3 of F1-score, outperforming the two baselines proposed: (1) a hierarchical multi-label model (baseline 1), based on text representations, and (2) a multi-label model (baseline 2), based on the concatenation of deep image and text representations, for predicting the corresponding misogynous type.

The best result was a F1-score of 73.1, with a triple tier between teams *SRCB_roc*, *TIBVA*, and *PAFC*. Our best proposal, however, achieved position 28 in the official leader board, outperforming all baselines.

6 Conclusions

In this paper the participation of the UMUTeam in the MAMI shared task, concerning the identification and categorisation of misogynous content in memes, is described. Our approach for solving the binary and multi-label classification tasks consisted in the combination of a set of language-independent linguistic features with contextual images and textual features obtained from the documents. Our best result was achieved in the misogynous categorisation task, with an macro F1-score of 66.3, reaching position 28 in the ranking.

We consider that the weakest point of our proposal is that we have not handle class imbalance in the testing dataset. However, we have evaluated some strategies that have improve our results, as reducing the number of instances and the application of class weight. As further work, we will evaluate data augmentation techniques, both for images and for text in order to deal class imbalance.

Acknowledgements

This work is part of the research project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/ AEI/10.13039/501100011033. This work is also part of the research project PDC2021-121112-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

References

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.

Hangbo Bao, Li Dong, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR.

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.

Richard Dawkins and Nicola Davis. 2017. *The selfish gene*. Macat Library.

María del Pilar Salas-Zárate, Giner Alor-Hernández, José Luis Sánchez-Cervantes, Mario Andrés Paredes-Valverde, Jorge Luis García-Alcaraz, and Rafael Valencia-García. 2020. Review of english literature on figurative language applied to social networks. *Knowledge and Information Systems*, 62(6):2105–2137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2022a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Rafael Valencia-García. 2022b. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–14.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.