UMUTeam at SemEval-2023 Task 10: Fine-grained detection of sexism in English

Ronghao Pan[1], **José Antonio García-Díaz**[1], **Salud María Jiménez Zafra**[2], **Rafael Valencia-García**[1]

[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{ronghao.pan@um.es, joseantonio.garcia8, valencia}@um.es
[2] Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain
sjzafra@ujaen.es

Abstract

In this manuscript, we describe the participation of UMUTeam in the Explainable Detection of Online Sexism shared task proposed at SemEval 2023. This task concerns the precise and explainable detection of sexist content on Gab and Reddit, i.e., developing detailed classifiers that not only identify what is sexist, but also explain why it is sexism. Our participation in the three EDOS subtasks is based on extending new unlabeled sexism data in the Masked Language Model task of a pre-trained model, such as RoBERTa-large to improve its generalization capacity and its performance on classification tasks. Once the model has been pre-trained with the new data, fine-tuning of this model is performed for different specific sexism classification tasks. Our system has achieved excellent results in this competitive task, reaching top 24 (84) in Task A, top 23 (69) in Task B, and top 13 (63) in Task C.

1 Introduction

The term sexism refers to any abuse or negative sentiment directed at women on the basis of their sex, or on the basis of their sex combined with one or more identity attributes. It should be noted that sexism is a growing problem on the Internet and can have harmful effects on women and other marginalized groups, making online spaces less accessible and unwelcoming, which can perpetuate asymmetries and social injustices. Automated tools are already widely used today to identify and evaluate sexist content online. It is important that these tools provide specific ratings and explain why content is considered sexist to improve the interpretability, reliability, and comprehensibility of decisions made by automated tools. However, most only provide generic high-level rating without further explanation. Thus, the Explainable Detection of Online Sexism (EDOS) shared task (Kirk et al., 2023), proposed at SemEval 2023, supports the development of English-language models for sexism detection that are more accurate and explainable, with precise classifications of sexist content on Gab and Reddit. To this end, the task is divided into three hierarchical subtasks:

- **Binary sexism detection (Task A)**: a binary classification where systems have to predict whether a post is sexist or not sexist.
- **Category of sexism (Task B)**: for posts which are sexist, a four-class classification where systems have to predict one of four categories (threats, derogation, animosity, and prejudiced discussions).
- **Fine-grained vector of sexism (Task C)**: for post which are sexist, an 11-class classification where systems have to predict one of 11 fine-grained vectors[1] (1.1 threats of harm, 1.2 incitement and encouragement of harm, 2.1 descriptive attacks, 2.2 aggressive and emotive attacks, 2.3 dehumanising attacks and overt sexual objectification, 3.1 casual use of gendered slurs, 3.2 profanities and insults, 3.3 immutable gender differences and gender stereotypes, 3.4 backhanded gendered compliments, condescending explanations or unwelcome advice, 4.1 supporting mistreatment of individual women, 4.2 supporting systemic discrimination against women as a group).

Advances in deep learning have had a major impact on Natural Language Processing (NLP), significantly improving the ability of computation to understand and process human language. Transformers is a neural network architecture that has had a major impact on NLP and has outperformed many tasks, such as automatic classification, sentiment analysis, and automatic translations (Bozinovski, 2020). Fine-tuning of pre-trained Transformer models has proven to be a highly effective

[1] https://codalab.lisn.upsaclay.fr/competitions/7124

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 589–594
July 13-14, 2023 ©2023 Association for Computational Linguistics

technique for improving performance on a wide range of NLP tasks. By taking the advantage of the model's pre-trained knowledge and adapting it to the specific task, the model can significantly improve its ability to perform the specific task, such as classification tasks. However, most Masked Language Models (MLM) are pre-trained on massive datasets from different domains, so according to Arefyev et al. (2021), additional training with the objective MLM on domain or task specific data prior to fine-tuning for the final task is known to improve the final performance.

We participate in all proposed EDOS subtasks using an approach that is based on extending the MLM task of the RoBERTa-large using new unlabeled sexism-related data and then fine-tuning this model for different specific sexism classification tasks. Our system has achieved excellent results in this competitive task, reaching top 24 (84) in Task A, top 23 (69) in Task B, and top 13 (63) in Task C. Besides, it is worth noting that our research group has previously experience dealing with hate-speech and sexism identification in works focused on Spanish (García-Díaz et al., 2022).

The remainder of this paper is organized as follows. Section 2 provides a summary of important details about the task setup. Section 3 offers an overview of our system for the subtasks. Section 4 presents the specific details of our systems. Section 5 discusses the results of the experiments, and finally, the conclusions are shown in Section 6.

2 Background

We only used the dataset given by organizers. The dataset provided consists of 20,000 entries. Of these, 10,000 are sampled from Gab and 10,000 from Reddit. All entries are first labeled by three trained annotators and disagreements are adjudicated by one of two experts. Our first experiments consisted of examining the number of examples of each label to see if the dataset is balanced or not. Table 1 shows the sexist and not sexist posts in the training and development sets for Task A. It can be seen that the dataset is unbalanced, as the number of non-sexist posts is three times higher than the number of sexist posts.

For Task B, we only used the sexist posts from Task A. Sexist post are classified into 4 general categories: (1) *Threats, plans to harm and incitement*, (2) *Derogation*, (3) *Animosity*, and (4) *Prejudiced discussions*. In Table 2, the distribution of each category in the training and dev sets is shown, and we can see that there are few threats-type sexist posts compared to the two other categories.

Table 1: Corpus statics by split for the Task A

Split	Sexist	Not sexist	Total
Train	3,398	10,602	14,000
Dev	486	1,514	2,000
Test	970	3,030	4,000
Total	4,854	15,146	20,000

Table 2: Corpus statics by split for the Task B, including Threats (T), Derogation (D), Animosity (A), and Prejudiced discussions (PD)

Split	T	D	A	PD	Total
Train	310	1,590	1,165	333	3,398
Dev	44	227	167	48	486
Test	89	454	333	94	970
Total	443	2,271	1,665	475	4,854

Each category of a sexist post can be divided into a set of sexist content vector. Thus, Task C consists of classifying and identifying the type of sexist content in the posts. Figure 1 depicts the number of posts of each content type.

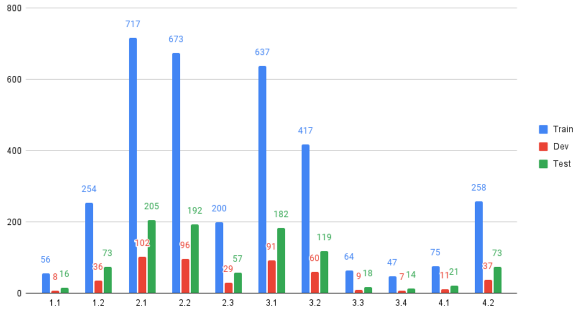

Figure 1: Corpus statics by split for the Task C

Furthermore, the organizers have published 2 unlabeled auxiliary datasets, consisting of 1 million Gab posts and 1 million Reddit posts. The texts of both datasets have been cleaned and prepared following the same procedure as the labeled data.

3 System Overview

For solving EDOS shared task, we built a system, which architecture is depicted in Figure 2. In a nutshell, our system works as follows. First, both labeled and unlabeled datasets were pre-processed.

Second, we trained the RoBERTa-large MLM with the unlabeled datasets provided by the organizers. Finally, once we pre-trained the RoBERTa-large with the unlabeled data, we fine-tuned this model with labeled data for different specific sexism classification task.

3.1 Dataset preprocessing

Our preprocessing stage consists of the following processes to clean the text of the EDOS dataset:

1. Replacement of all hashtags and mentions with *#[HASHTAG]* and *@[USER]*.
2. Social media posts require a lot of cleaning up, but it is inefficient to clean up each post, so a general clean-up approach was applied in this case:
    - All emojis have been removed, as our system does not use emoji features.
    - Some general contractions have been expanded, such as "*lmao*" to "*laughing my ass off*", "*y'all*" to "*you all*", "*i'd*" to "*i would*", etc.

3.2 Classification Model

A Transformer model pre-trained with MLM (Masked Language Modeling) is a type of natural language model that has been pre-trained on large amounts of unlabeled text data, with the goal of learning useful patterns and representations of words and phrases in the text. This allows the model to better perform specific natural language tasks, such as text classification, text generation and language translation (Bozinovski, 2020). RoBERTa-large is one of the pre-trained Transformer models, was developed by Facebook AI Research (FAIR) and has 355 million parameters, making it significantly larger than most of the pre-trained models available at the time (Liu et al., 2019). However, this model has been pre-trained with a corpus that is not closely related to the domain of sexism, so we have trained the RoBERTa-large MLM model on a provided unlabeled dataset. To train the RoBERTa-large with the unlabeled data, the following steps were performed:

- Preprocess the texts and separate them into sentences, i.e., each sentence occupying one line.
- Use the *LineByLineTextDataset* class with a block size of 128 to transform the dataset to the correct format for training.
- Use the data collator object to form batches of input data to train the MLM.
- Finally, the model has been trained with an epoch of 3 and a training batch size of 32.

Once the RoBERTa-large has been pre-trained with the new unlabeled data, fine-tuning of this model is performed for different specific sexism classification tasks. The processes to be performed are as follows.

3.2.1 Tokenization

RoBERTa-large tokenization consists of splitting a text into smaller units called tokens, which are used as input for the model. In this case, RoBERTa-large use the byte-level BPE (Bype-Pair Encoding), which is a hybrid between character-level and word-level representations that allows handling the large vocabularies common in natural language corpora.

3.2.2 Fine-Tuning

Fine-tuning is a deep learning model training process in which a pre-trained model is taken as a basis and additionally trained on a specific task with a smaller dataset. In this way, the prior knowledge of the pre-trained model can be exploited and its behaviour adjusted. In this case, in the absence of sufficient training data, as shown in Table 2, and Figure 1, this process improves the overall performance of the model compared to training from scratch with the corpus provided by the organizers.

4 Experimental Setup

For the three subtasks of EDOS shared task, the corpus we used to train the model is from the organizer, without other external data (see Section 2). System performance on all Tasks (A, B and C) is evaluated using Macro F1 score. Macro F1 is calculated as an equal-weighted average of F1 scores for each class in a classification task. It is widely used for imbalanced classification tasks because it assigns equal importance to minority and majority classes.

To obtain the classification model for each task, we conducted an hyperparameter optimization 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.

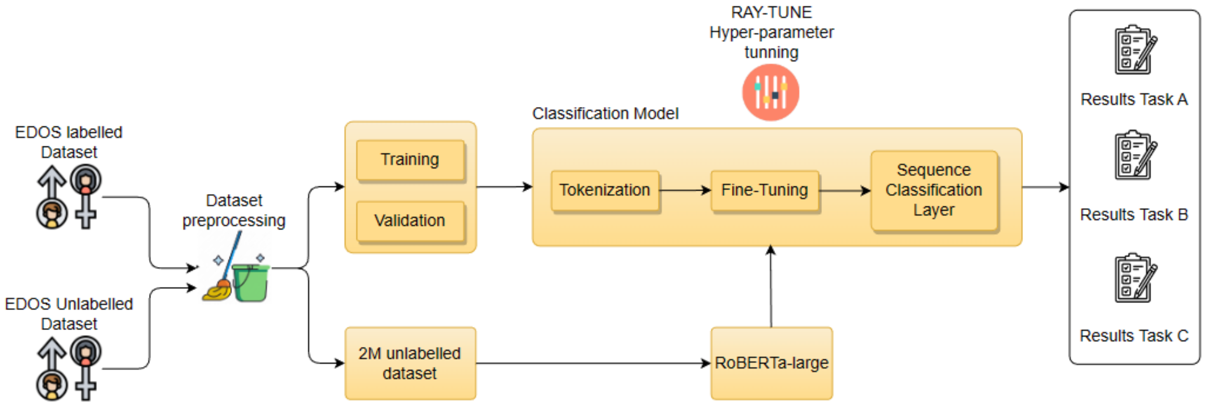

Figure 2: System architecture proposed by the UMUTeam for solving the EDOS shared task

The hyperparameters evaluated, and their interval range are: (1) weight decay (between 0 and 0.3), (2) training batch size ([8,16]), (3) number of training epoch ([1-20]), and (4) learning rate (between 1e-5 and 5e-5). In this case, the best configuration for extended RoBERTa-large with new unlabeled data is 15-20 epoch, a training batch size of 16, a weight decay of 0.183556, and a learning rate of 1.186663e-5 for all Tasks (A, B and C).

5 Results

Each system was evaluated using the validation split and tested using the test split (see Section 2). The official results for the Task A, B and C are depicted in Table 3, 4, and 5. In this shared task, the organizers have published the golden labels of the test split, so all error analysis will be based on the examples in the test split. To analyse the errors and check in which cases the models of different tasks give erroneous predictions, a normalized confusion matrix with truth labels has been used, which consists of a table showing the distribution of the predictions of a model compared to the truth label of the data. The confusion matrix of the model for each task is shown in Figure 3, 4, and 5.

For Task A, we achieved 24/84 position with a macro F1 score of 0.8495, which is only 2.51% worse than the first (see Table 3). Particularly, taking into account the confusion matrix of our model (see Figure 3), it can be observed that the model confuses 11.44% of sexist and not-sexist posts and 3.63% of non-sexist post with sexist post. Looking at the training dataset of the Task A (see Table 1), there is an imbalance of more than 50% between sexist and non-sexist cases. However, by our approach of fine-tuning an extended RoBERTa-large with an unlabeled sexist dataset, we have significantly improved the classification of sexist posts despite not having sufficient training data.

Table 3: Comparison of our results on the test dataset with other participants for Task A

#	Team	Macro F1
1	PingAnLifeInsurance	0.8746
2	stce	0.8740
2	FiRC-NLP	0.8740
4	PALI	0.8717
24	**UMUTeam**	**0.8495**
25	A2Z	0.8479
	...	
84	NLP_CHRISTINE	0.5029

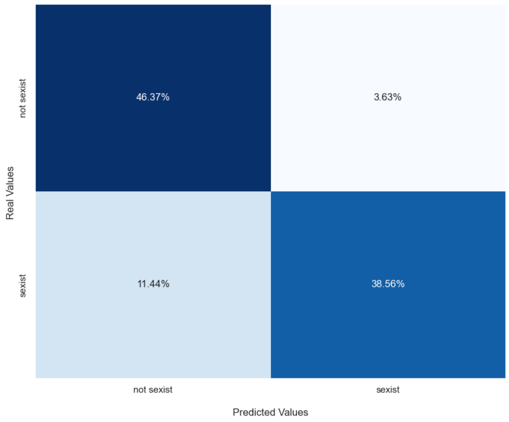

Figure 3: Confusion matrix of our system in the Task A

Regarding Task B, we achieved the position 24/70, which is 9.41% worse than the first one, as shown in Table 4. Using the confusion matrix of the model (see Figure 4), it can be seen that the most common error made by the model was in the classification of prejudiced discussions and animosity. In the classification of prejudiced discussions, it is often confused with other categories, such as derogation, and animosity. This is because the model does not have sufficient training data on the case of prejudiced discussions cases, as shown in Table 2. Regarding the classification of animosity, the model is confused with derogation, despite being one of the categories with the most items in the training set, due to their semantic and syntactic similarity.

Table 4: Comparison of our results with other participants for Task B

#	Team	Macro F1
1	JUAGE	0.7326
2	PASSTeam	0.7212
3	stce	0.7203
4	PALI	0.7194
23	**UMUTeam**	**0.6395**
24	DH-FBK	0.6385
	. . .	
69	NLP_CHRISTINE	0.2293

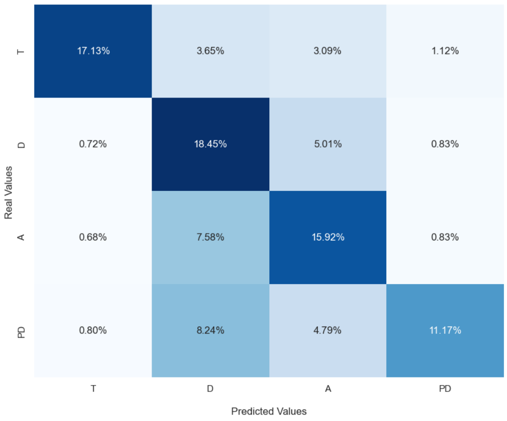

Figure 4: Confusion matrix of our system in the Task B

As for the Task C, we achieved the position 13/63, which is 8.13% worse than the first one, as shown in Table 5. This task consists of classifying the subcategories of each sexist category (Task B), so that thought the confusion matrix (see Figure 5) it can be observed that the most errors made by the model have been in the classification of subcategories of the prejudiced discussions and animosity categories, which are precisely the categories that the models of Task B has committed more errors in the classification.

Table 5: Comparison of our results with other participants for Task A

#	Team	Macro F1
1	PALI	0.5606
2	stce	0.5487
3	PASSTeam	0.5412
4	FiRC-NLP	0.5404
13	**UMUTeam**	**0.4793**
14	JUST_ONE	0.4774
	. . .	
63	shm2023	0.0632

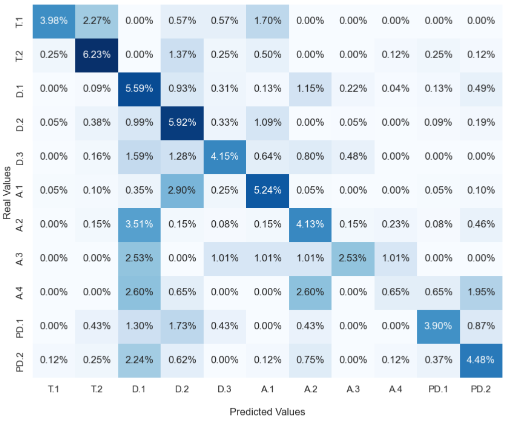

Figure 5: Confusion matrix of our system in the Task C

Finally, we compared the performance of the RoBERTa-large model before and after domain adoption. The macro F1 score results for the different tasks are shown in Table 6. It can be seen that the model with domain adoption has improved the overall performance compared to the generic model. In particular, the RoBERTa-large model with domain adaption has achieved an improvement of 0.691% in Task 1, 0.817% in Task 2, and 4.44% in Task 3.

6 Conclusion

In this working notes we have described the participation of UMUTeam in the EDOS shared task of SemEval 2023. In this shared task, the participants were required to perform precise and explainable detection of sexist content on Gab and Reddit, i.e., developing detailed classifiers that not only identify

Table 6: Comparison of results (macro F1) between RoBERTa-large and RoBERTa-large with domain adaptation for different tasks.

Model	Task A	Task B	Task C
RoBERTA-large	0.84259	0.63133	0.4349
RoBERTA-large (domain adaption)	**0.84950**	**0.63950**	**0.4793**

if a post is sexist, but also explain why it is sexism. Our system has achieved excellent results in this competitive task, reaching top 24 in Task A, top 23 in Task B, and top 13 in Task C.

As commented at subsection 3.1, our system does not use emoji features. However, we are planning to incorporate these features as well as others related to figurative language to our pipeline (García-Díaz and Valencia-García, 2022; del Pilar Salas-Zárate et al., 2020). We consider that these kind of features can improve the overall performance of the system in which the words in a text differs from its literal meaning.

7 Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme, and Salud María Jiménez-Zafra is partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC_01073).

References

Nikolay Arefyev, Dmitrii Kharchev, and Artem Shelmanov. 2021. NB-MLM: Efficient domain adaptation of masked language models for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9114–9124, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA. PMLR.

Stevo Bozinovski. 2020. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica (Slovenia)*, 44(3).

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.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

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*, 8(2):1723–1736.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.