CLassifiers at EXIST 2023: Detecting Sexism in Spanish and English Tweets With XLM-T

Notebook for the Exist Lab at CLEF 2023

Berna Ilke Ersoy[1,*,†], Gian Radler[1,*,†] and Sofia Carpentieri[1,*,†]

[1]University of Zurich, Rämistrasse 71, 8006 Zurich, Switzerland

Abstract

In this paper, we present the submission of our team *CLassifiers* to the EXIST 2023 shared task competition on sexism detection. Our approach involves utilizing multiple techniques based on a pre-trained RoBERTa model. We developed two distinct models: a binary classifier for Task 1 and a multi-label classifier for Task 3. Leveraging the multilingual XLM-T model, we tailored our models to each task and achieved favorable results in our experiments.

Keywords

hate speech detection, sexism detection, sexism categorization, social media

1. Introduction

With increasing numbers of online users and larger digital space, more content is being produced everyday. Along with this surge in content, there has been a parallel rise in the dissemination of hateful content. It appears in various forms, one of which is sexism — a form of prejudice or discrimination based on gender, usually targeted towards women or feminine-presenting people [1]. As it implies harmful ideals not only about femininity, but also about masculinity, it hurts everyone in societies all over the world, unrelated to culture or religion [2]. It is an on-going struggle to combat and can be especially hard to detect for humans socialized within patriarchal structures. For online spaces there is an increased need to create automatic sexism detection systems which allow to filter or censor to avoid the spreading of harmful content. The difficulty and need to develop reliable systems is evident in the amount of published research on sexism or hate speech detection [3].

For our participation in this third edition of the shared task on sexism detection, we used multiple approaches based on a pre-trained XLM-RoBERTa model [4]. This paper describes the submission of our team *CLassifiers* to the EXIST 2023 shared task competition [5, 6]. We created two different approaches: one for Task 1 of the shared task that acts as a binary classifier and a second model for Task 3 that is a multi-label classifier. Both of our models are based on the multilingual XLM-T [7], which we adapted for both tasks.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

†These authors contributed equally.

bernailke.ersoy@uzh.ch (B. I. Ersoy); gian.radler@uzh.ch (G. Radler); sofia.carpentieri@uzh.ch (S. Carpentieri)

2. Related Work

The field of sexism detection has seen a variety of approaches, most of them focused on social media posts. Anonymity, invisibility and accessibility make hateful posts more common amongst social media users, as they often go unnoticed and are not followed up by consequences [8]. Researchers have been attempting to automate the process of identifying sexist content to help social platforms and communities establish safer environments. However, sexism detection is challenging due to its subtle nature, the diversity of its expressions, and the complexity of the language used. Different methods have been used to tackle the sexism detection task. Traditional approaches involve rule-based systems and machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, and Random Forests. These methods usually rely on manually selected features like Bag of Words (BoW), TF-IDF, and sentiment scores [9].

Recently, the focus has shifted towards Transformer-based models like BERT [10], GPT [11], and RoBERTa [4] due to their superior performance on a wide range of NLP tasks. These models leverage the Transformer architecture's strengths to capture complex language semantics and context effectively. XLM-RoBERTa [12] is an extension of RoBERTa that has been trained multilingualy in XLM style [13]. It combines the advantages of RoBERTa, which is a variant of BERT optimized for more robust performance, with the benefits of cross-lingual training, thus enabling the model to understand text in various languages. XLM-RoBERTa uses a masked language modeling objective, similar to BERT, but with modifications in the training process. It employs dynamic masking instead of static masking and removes the next sentence prediction task, which results in a more robust and versatile model. Its multilingual capabilities make it especially suitable for tasks like sexism detection, where the content may be in different languages.

Tweets pose unique challenges to sexism detection due to their casual language use, and the frequent use of slang and abbreviations [14]. The character limit of tweets often leads to condensed expressions of sexism that are difficult to identify using traditional detection methods. The XLM-T model [7], an XLM-RoBERTa model pre-trained on multilingual tweet data, emerged as a promising tool for tackling the challenges of tweet-based content classification tasks. In the context of sexism detection, the use of XLM-T offers several advantages. Firstly, its multilingual pre-training allows it to detect sexist content across different languages, which is critical given the global nature of Twitter. The model's grounding in tweet-based language, coupled with its ability to understand the nuances and idiosyncrasies of this language, lends it robust capabilities in identifying subtle expressions of sexism. Secondly, the model can be fine-tuned on labeled datasets of sexist and non-sexist tweets. Through this fine-tuning, XLM-T can learn to distinguish the markers of sexism embedded within the tweets.

3. Material and Methods

3.1. Dataset

The EXIST 2023 dataset includes postings from Twitter as well as annotations for different categories of sexism. It contains 10,034 labeled tweets, both in English and in Spanish, which are split into training, development and test sets. The two languages are balanced in their

distributions. Each tweet in the dataset is annotated by six annotators and no hard unique labels are given. For an overview, see Table 1.

Table 1
Counts and distrubution of Spanish and English tweets in dataset according to splits

Dataset Split	Spanish Tweets	English Tweets	Total Tweets
Train Set	3660	3260	6920
Evaluation Set	549	489	1038
Test Set	1098	978	2076

Labels for Task 1, a binary classification task, contain the values sexist or non-sexist, indicating whether the tweet is perceived to be sexist or not. Labels for Task 3, a multi-label classification task, are provided as a set of arrays (one array per annotator) indicating the type or types of sexism found in a tweet. The labels for Task 3 are: ideological-inequality, stereotyping-dominance, objectification, sexual-violence, misogyny-non-sexual-violence, - (assigned to non-sexist tweets), and unknown. Additional information on each tweet and its annotation include the tweet ID, language, number of annotators, gender of the annotators, and age group of the annotators.

3.2. External Data

As a data augmentation strategy, in addition to the EXIST 2023 dataset, we used the EXIST 2021 dataset [15]. It consists of Spanish and English tweets, namely 6,977 for training and 3,386 for testing. The data was annotated by five crowd-sourcing annotators each. Final labels were selected by majority decision. The data was used for training on Task 1, since the labels were incompatible with Task 3.

3.3. Evaluation Metrics

The official evaluation metrics used for the results are ICM and ICM-soft. Proposed in 2022 by Amigó and Delgado [16], the ICM metric is based on information theory and draws inspiration from the Information Contrast Model (ICM) [17]. For the shared task, the organizers have extended the original ICM metric and created ICM-soft. This extension enables the evaluation of soft labels and is suited for learning with disagreement scenarios, as is the case in EXIST 2023. Higher values of the ICM and ICM-soft metrics indicate a stronger similarity between system outputs and ground truth; thus, higher values are considered better.

3.4. Preprocessing and Data Preparation

Considering that our dataset for training our models is bilingual, we decided to do minimal preprocessing. All tweets were lower-cased, we removed HTML tags, URLs, and user mentions, and converted emojis into their CLDR (Common Locale Data Repository) short name. For instance, the emoji "💜" is replaced by ":purple_heart:" and the emoji "💫" is replaced by ":dizzy:".

Table 2
Overview of the models utilized.

Task	Labels	Model
1	hard	cardiffnlp/twitter-xlm-roberta-base
1	soft	cardiffnlp/twitter-xlm-roberta-base
3	hard	xlm-roberta-base
3	soft	cardiffnlp/twitter-xlm-roberta-base

3.5. Problem Modeling

In both tasks, the classification process involves hard or soft labels. Soft labels refer to predicting probabilities for each label category, while hard labels only involve a label prediction.

3.5.1. Task 1: Binary Classification

Task 1 required binary classification to determine whether a tweet should be considered sexist and to label them as either "YES", meaning that the tweet is sexist, or "NO". indicating that a tweet is not. A tweet was labeled as sexist if it received agreement from more than three annotators, disregarding evenly split cases. The soft labels provided a representation of the uncertainty by assigning probabilities that summed up to one for each tweet. We chose XLM-T [7] for both soft and hard settings. For an overview, please refer to Table 2. For our Task 1 submissions, we conducted three runs with variations in the number of training epochs. Run 1 involved training for 3 epochs, Run 2 for 4 epochs, and Run 3 for 6 epochs.

3.5.2. Task 3: Multi-Label Classification

Task 3 involved multi-label classification, where a tweet was categorized in one or more label categories. To assign a hard label to a tweet, at least two or more annotators had to agree on the presence of a particular label. Soft labels were provided for each category, indicating the probabilities associated with the presence of that label. For soft label classification, we decided to train one model for each of the five labels to predict the probabilities of said label. For Task 3, we also used the binary classification model developed in Task 1 to enhance the labeling process. If the Task 1 model predicted that a tweet was not sexist ("NOT SEXIST"), any corresponding hard labels predicted by the Task 3 model were replaced with "NO" to align with the model's prediction, improving the overall performance of our multi-label classification model. For our Task 3 submissions, we conducted two runs with differences in the number of training epochs. In Run 1, we trained the model for 5 epochs, while in Run 2, we trained it for 4 epochs.

4. Results

Since our team only participated in Tasks 1 and 3, we will only discuss the results of these two tasks in the following sections. For Task 1 we submitted 3 runs and for Task 3 we handed in 2 runs. All results in Table 3, Table 4, Table 5 and Table 6 are taken from the official leaderboard of the corresponding task from the soft-soft and hard-hard evaluation sheets and ranked according

Table 3
Results for Task 1 soft-soft evaluation

Run	Rank	Team Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
test_gold_soft	0	-	3.1182	1	0.5472
SINAI_3	1	1	0.9030	0.6421	0.7960
CLassifiers_3	2	2	0.9027	0.6421	0.9754
CLassifiers_2	3	2	0.8698	0.6368	0.9823
CLassifiers_1	4	2	0.8172	0.6283	0.9672
CIC-SDS.KN_2	5	3	0.7960	0.6248	0.7770
test_majority_class	47	-	-2.3585	0.1152	4.6115

Table 4
Results for Task 1 hard-hard evaluation

Run	Rank	Team Rank	ICM-Hard	ICM-Hard Norm	F1
test_gold_hard	0	-	2.1533	1	1
Mario_3	1	1	0.6575	0.785	0.8109
CLassifiers_2	12	5	0.539	0.7095	0.7702
CLassifiers_3	18	5	0.5282	0.7026	0.7642
CLassifiers_1	21	5	0.5113	0.6918	0.7615
test_majority_class	66	-	-0.4413	0.0847	0

to their ICM-Soft score along the normal ICM score. We have added the ”Team Rank” column, which takes into account only the best models of a team.

4.1. Task 1

The three submitted runs for this task differed in the number of epochs during training, while using the same base model. The results in Table 3 show that any increase in training epochs correlates directly with an increase in the ICM-Soft score.

In the first run, we achieved an ICM-Soft score of 0.8172 and an ICM-Soft normalized score of 0.6248 whereas the highest possible score is a 3.1183 in ICM-Soft and 1.0 in the ICM-Soft normalized score. In run 2, we managed to achieve an ICM-Soft score of 0.8698 and a normalized score of 0.6368. Lastly, in run 3 we achieved our overall highest ICM-Soft score of 0.9027 and a normalized score of 0.6421. Lagging behind the first rank by only 0.0003 in the ICM-Soft score, we consider it a state-of-the-art result. Training for one more epoch might have improved results even further.

The two lowermost runs in the table are the non-informative baselines provided by the EXIST 2023 committee and ranked 47th and 52nd respectively, with an ICM-soft score of -2.3585 and -3.0717. The EXIST 2023_test_majority_class run set the probability of the class to 1 and classified all instances as the majority class. The EXIST 2023_test_minority_class classified all instances as the minority class and set the probability of the class to 1.

The evaluation results in Table 4 show the ICM-Hard scores for our three runs. As in the soft-soft scenario, the model in run 1 was ranked lowest among the three models. However, in this case, the second run achieved a higher ranking compared to the third run. Run 2 obtained an ICM-Hard score of 0.539 and an ICM-Hard normalized score of 0.7095 and was ranked 12th,

Table 5
Results for Task 3 soft-soft evaluation

Run	Rank	Team Rank	ICM-Soft	ICM-Soft Norm
test_gold_soft	0	-	9.4686	1
AI-UPV_3	1	1	-2.3183	0.7879
DRIM_1	4	2	-3.6842	0.7633
CLassifiers_1	6	4	-6.4072	0.7143
test_majority_class	15	-	-8.7089	0.6729
CLassifiers_2	20	4	-14.7828	0.5636

Table 6
Results for Task 3 hard-hard evaluation

Run	Rank	Team Rank	ICM-Hard	ICM-Hard Norm
test_gold_hard	0	-	2.1533	1
roh-neil_1	1	1	0.4433	0.6763
test_majority_class	27	-	-1.5984	0.2898
CLassifiers_2	29	13	-1.8664	0.2391
CLassifiers_1	30	13	-1.8852	0.2355
M&S_NLP_1	31	14	-2.1587	0.1838

while run 3 had an ICM-Hard score of 0.5282 and an ICM-Soft normalized score of 0.7026 and was ranked 18th.

4.2. Task 3

The two runs for Task 3 vary according to the number of training epochs used for building the model. The difference of one epoch during training resulted in large discrepancies between the ICM-Soft scores of the two runs. Run number 2 (4 epochs) attained an ICM-Soft score of -14.7828 and a normalized score of 0.5636. The first run (5 epochs), achieved an ICM-Soft score of -6.4072 and a normalized score of 0.7143. Our two runs ranked (in order mentioned) 20th and 6th place. The total difference of the ICM-Soft score between our better run and the first ranking place is 4.0889. The difference between the normalized scores equals to 0.0736. Again, we conclude that training for more epochs may have further improved our results.

We also report the results for the hard-hard evaluation scenario in Table 4. Both our models have encountered difficulties in accurately predicting the hard labels. Our first run ranked 30th with an ICM-Hard score of -1.8664, while our second run was ranked 29th with a score of -1.8852. The models in the hard label scenario perform worse than the majority baseline, suggesting that more optimization is needed for both systems that go beyond changing the number of epochs trained.

5. Conclusion

In conclusion, the XLM-T model, pre-trained on a large corpus of multilingual tweet data, has demonstrated substantial promise for the task of sexism detection in the unique linguistic

context of Twitter. Even with minimal configuration and experimentation, it has achieved second place in the binary sexism detection task and sixth place in the more complex multi-label classification task. This showcases the power of the domain-adapted XLM-T foundation model and highlights how little effort is needed to yield promising results when fine-tuning it for specific tasks.

References

[1] J. K. Swim, L. L. Cohen, Overt, covert, and subtle sexism: A comparison between the attitudes toward women and modern sexism scales, Psychology of women quarterly 21 (1997) 103–118.

[2] I. Horowitz, Sexism hurts us all, Agenda 13 (1997) 75–80. doi:10.1080/10130950.1997.9675610.

[3] M. S. Jahan, M. Oussalah, A systematic review of hate speech automatic detection using natural language processing, Neurocomputing 546 (2023) 126232. URL: https://www.sciencedirect.com/science/article/pii/S0925231223003557. doi:https://doi.org/10.1016/j.neucom.2023.126232.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[5] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[6] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – learning with disagreement for sexism identification and characterization (extended overview), in: M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[7] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[8] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[9] O. Istaiteh, R. Al-Omoush, S. Tedmori, Racist and sexist hate speech detection: Literature review, in: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), 2020, pp. 95–99. doi:10.1109/IDSTA50958.2020.9264052.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[11] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.

[13] A. Conneau, G. Lample, Cross-lingual language model pretraining, Advances in neural information processing systems 32 (2019).

[14] S. Sharifirad, S. Matwin, When a tweet is actually sexist. a more comprehensive classification of different online harassment categories and the challenges in NLP (2019). arXiv:1902.10584.

[15] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sEXism Identification in Social neTworks, in: Proces. del Leng. Natural, 2021.

[16] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[17] E. Amigó, F. Giner, J. Gonzalo, M. Verdejo, On the foundations of similarity in information access, Information Retrieval Journal 23 (2020). doi:10.1007/s10791-020-09375-z.