Team Aditya at EXIST 2024 – Detecting Sexism in Multilingual Tweets using Contrastive Learning Approach

Notebook for the EXIST Lab at CLEF 2024

Aditya Shah[1,*,†], Aditya Gokhale[2,†]

[1]Pune Institute Of Computer Technology, Pune India
[2]Pune Institute Of Computer Technology, Pune India

Abstract

Due to the growing impact of social media, the necessity for automated mechanisms that can identify sexism and other forms of disrespectful and hateful conduct is rising, aiming to create a more inclusive and respectful digital space. However, it poses significant challenges due to the variety of hate categories and the complexity of interpreting the author's intent, particularly under the multilingual learning framework. This paper describes Team Aditya's participation in the EXIST (sEXism Identification in Social neTworks) Lab at CLEF 2024 . The proposed system makes use of large language models (i.e., Bertweet, mBERT and XLM-RoBERTa) for identifying sexism in English and Spanish language. This work describes our participation in EXIST task 1. Considering a hard evaluation, we obtained F1 score of 0.7691 using best epoch trained with XLM-Roberta. We are ranked 14th in the given task.

Keywords

Sexism, Disrespectful and hateful conduct, Large language models,

1. Introduction

Sexism is prejudice or discrimination based on one's sex or gender, often targeting women due to their gender. This harmful mindset causes inequality, limits opportunities, and reinforces oppressive power dynamics, limiting progress toward a fairer society.

The rise of social media platforms such as Twitter and Facebook has led to a significant change in communication methods. Identifying and reducing hate speech on these platforms can be a daunting task, due to large volumes of data generated. This requires using automated techniques and advanced technologies to efficiently process and classify the content.

The EXIST 2024 [1][2] shared task is focused on detecting sexism, which ranges from blatant misogyny to more subtle, implicit forms of sexist behavior. This task differentiates itself from other related tasks on sexism detection by encompassing not only posts that are explicitly identified as sexist but also posts that document reported incidents of sexism.

2. Dataset Details

The dataset provided by the Exist2024 initiative consists of about 7K tweets, equally split between Spanish and English language. To mitigate label bias, organisers have considered two different social and demographic parameters: gender (MALE/FEMALE) and age (18-22 y.o./23-45 y.o./+46 y.o). The dataset was split into train, dev, and test sets, roughly distributed as 70%, 10%, and 20%, respectively, for both languages.

The labels for the tweets in Subtask 1 were categorized as "YES" or "NO" to indicate whether they conveyed a sexist meaning.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
aditya02shah@gmail.com (A. Shah); adityangokhale@gmail.com (A. Gokhale)

Table 1
Distribution of majority vote per language combining the Train and Dev datasets

Majority Vote	*English*	*Spanish*
Sexist	47.1%	55.7%
Non-Sexist	52.9%	44.3%

3. System Description

To derive the definitive hard label, we utilize the annotations by multiple annotators and apply majority voting. Only if 3 or more of the 6 annotators unanimously agree to a label YES, then the label is set to YES. Prior to training the model, preprocessing steps were carried out to remove any emoji's, URLs and mentions from the samples. This was done to remove any undue bias.

We conducted experiments with various models and found XLM-R to be the most effective, particularly due to its strong performance on multilingual data, as shown in Table 2.

Table 2
Evaluating Performance of Different Models on the Dev Test

Model	*F1*
BerTweet	0.7858
RoBERTa	0.7463
XLM-R-Large	0.7961
UML-T5	0.7683
mBert	0.7432

We finetune XLM-Roberta Large, a multilingual version of Roberta, trained on 2.5TB of filtered CommonCrawl data. This allows us to handle both English and Spanish samples by utilizing a single model.

The model was trained using contrastive learning, enabling it to differentiate between samples effectively. It learns an embedding space where similar pairs are positioned in close proximity, while dissimilar pairs are distinctly separated.

To improve the representation of each example in a batch we created label-aware embeddings by prefixing the text with its corresponding label [3]. A contrastive loss function was then used to align the text features closer to the representations of their correct labels, to improve classification capabilities of the system.

We utilized three datasets during the finetuning process: The training dataset was used to learn an embedding space using contrastive learning. The validation dataset was used to retain the most effective checkpoint and the test dataset was used to evaluate performance on unseen data.

We submitted the following systems for evaluation, with each system being trained with the following hyperparameters:

- Batch size: 16
- Learning rate: 1e-6
- Dropout: 0.35

These hyperparameters were chosen using the Optuna library. We selected the hyperparameters with the best performance on the validation set, after running 50 trials with varying configurations. These optimal hyperparameters help improve performance and reliability.

1. **ADITYA1:** The model was trained for 30 epochs, leveraging the combined train and dev datasets. The optimal epoch was subsequently saved based on the system's performance on the test set. The saved model was used to make predictions on the unseen test set.

2. **ADITYA2:** The model was trained for 30 epochs on the train set. The optimal epoch was subsequently saved based on the system's performance on the test set. The saved model was used to make predictions on the unseen test set.

3. **ADITYA3:** The model was trained for 12 epochs on the combined train and dev datasets. This model was used to make predictions on the unseen test set.

4. Results

For the exist-Task1 we utilized three systems as described in **Section 3.** The evaluation metric used for these systems was the ICM Metric.
The ICM metric [4] is a similarity function that generalizes Pointwise Mutual Information (PMI) to compute the similarity between a model's output and the ground truth categories. To calculate the normalized ICM, the "Minority class" baseline (that classifies all instances as the minority class) is considered the lowest score (i.e., 0) and the "Gold standard" is considered the highest score (i.e., 1). Additionally, the models of sexism identification provided two types of outputs, "Hard" labels that classify samples into sexist or not-sexist and "Soft" labels that specify a value between 0 and 1 in order to measure "the degree of sexism" involved in the sample.
These labels were used to evaluate the models across three schemes, described as follows:

- Hard-hard evaluation: the ICM similarity between the hard system output and hard ground truth
- Soft-soft evaluation: the ICM similarity between the soft system output and the soft ground truth

All three of our submitted systems generated Hard Labels, which were subsequently utilized for the Hard-hard evaluation scheme. A summary of our experiments is presented in Table 3.

Table 3
Results for the hard-hard evaluation for Task 1

Run	*Rank*	*ICM-Hard*	*ICM-Hard Norm*	*F1_YES*
Gold	0	0.9948	1.0000	1.0000
Best Score	1	0.5973	0.8002	0.7944
ADITYA_1	34	0.4680	0.7320	0.7447
ADITYA_2	18	0.5246	0.7636	0.7669
ADITYA_3	14	0.5418	0.7723	0.7691

5. Related Works

The rise of social media has led to an increase in sexist content in society, necessitating the de- velopment of automated systems to detect and counteract sexism. However, the discrepancy in the composition of the tweets and the multilingual nature of the dataset cause problems . To address these problems, we used pre-processing to improve the effectiveness of our system. To solve the sexism identification challenge, contrastive learning with RoBERTa language model [5] has been used. Previous research has shown that deep learning algorithms, such as those used in [6], can outperform machine learning models for sexism detection in Spanish datasets collected from Twitter. Other studies, such as [7] and [8], have applied multilingual transformer models, including multilingual BERT and XLM-R, to detect sexism in multiple languages. Meanwhile, [9] has used pre-trained transformers for sexism detection in low-resource languages such as Romanian, and [10] has employed ensemble models for multilingual classification. Despite these efforts, there is still limited exploration of modeling and analyzing sexism in Spanish and English datasets, highlighting the need for further research in this area. There has been

relatively limited research on modelling and analyzing sexism in datasets that contain both Spanish and English language content.

6. Conclusion and Future Scope

This paper presents the participation of team Aditya in the Task1 of the EXIST2024 [2] lab at CLEF, which focuses on sexism identification. We investigated a contrastive learning based approach for fine-grained analysis. The use of contrastive learning improved the classification capabilities of our models.
Throughout our experimentation, we evaluated various models, including BerTweet, mBERT and T5. However, XLM-Roberta Large consistently demonstrated superior results compared to its counterparts and the best performance was obtained with its help.

There is tremendous scope for advancements and progress in this field. In future works, we would like to refine and enhance our approach with different contrastive learning strategies, to improve the model's ability to distinguish between different classes. This approach can be also used in multiclass classification and multilabel classification problems. Large Language Models trained specifically on Spanish text could also be utilized to further improve performance.

References

[1] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[2] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[3] Q. Chen, R. Zhang, Y. Zheng, Y. Mao, Dual contrastive learning: Text classification via label-aware data augmentation, 2022. URL: https://arxiv.org/abs/2201.08702. arXiv:2201.08702.

[4] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), 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.

[5] J. Angel, S. Aroyehun, A. Gelbukh, Multilingual sexism identification using contrastive learning, Working Notes of CLEF (2023).

[6] 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.

[7] M. Schütz, J. Boeck, D. Liakhovets, D. Slijepčević, A. Kirchknopf, M. Hecht, J. Bogensperger, S. Schlarb, A. Schindler, M. Zeppelzauer, Automatic sexism detection with multilingual transformer models, arXiv preprint arXiv:2106.04908 (2021).

[8] H. H. Hemati, S. H. Alavian, H. Beigy, H. Sameti, Sutnlp at semeval-2023 task 10: Rlat-transformer for explainable online sexism detection, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 347–356.

[9] A. Moldovan, K. Csürös, A.-m. Bucur, L. Bercuci, Users hate blondes: Detecting sexism in user comments on online Romanian news, in: K. Narang, A. Mostafazadeh Davani, L. Mathias,

B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 230–230. URL: https://aclanthology.org/2022.woah-1.21. doi:10.18653/v1/2022.woah-1.21.

[10] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, 2021. arXiv:2111.04551.