

I2C-UHU at CLEF-2023 EXIST task: Leveraging Ensembling Language Models to Detect Multilingual Sexism in Social Media

Pablo Cordon, Jacinto Mata, Victoria Pachón and Juan Luis Domínguez

I2C Research Group, University of Huelva, Spain

Abstract

This paper describes the approaches developed by the I2C Group to participate on sub-task 1 in the CLEF 2023 task EXIST: sEXism Identification in Social neTworks. Our main contribution is to show the benefits of translating a bilingual dataset to a single language, as well as the effectiveness of using a group of classifiers based on transformers architecture. By combining different models, the individual advantages were exploited, resulting in better performance than using a single model. Moreover, the importance of choosing suitable hyperparameters during the model training process was highlighted by the results. Through careful experimentation and evaluation of different hyperparameter combinations, the settings that achieved the best performance for the given task were found. In our experiments we fine-tuned several pre-trained language models and decided to ensemble the three models that reached the best F1-scores. With this approach, we achieved an ICM-Hard score of 0.5075, ranking 25th in the competition.

Keywords

Deep Learning, Transformers, Hyperparameter, Ensembles, Twitter, Sexism, Hate Speech

1. Introduction

Sexism is a form of discrimination or prejudice based on gender that affects the dignity and rights of women and other marginalized groups. It is often expressed through language, such as insults, stereotypes, jokes, threats, or harassment. As Rodriguez-Sánchez et al (2020) expounds [1], social media platforms, such as Twitter, are widely used for communication and information sharing, but they also provide a space for sexist discourse and hate speech. Detecting and preventing sexism in social media is a challenging and important task for ensuring a respectful and inclusive online environment.

In this paper, we present a novel approach for on-line sexism detection in Spanish and English tweets using transformer models and natural language processing participating in subtask 1 of EXIST: sEXism Identification in Social neTworks from CLEF 2023 [2]. This task is a binary classification on sexism in a multi-lingual dataset of annotated tweets in Spanish and English, collected using different popular expressions and terms, commonly used to underestimate the


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

✉ pablo.cordon113@alu.uhu.es (P. Cordon); mata@uhu.es (J. Mata); vpachon@dti.uhu.es (V. Pachón); juan.dominguez@dti.uhu.es (J.L. Domínguez)

🆔 0009-0009-0096-6650 (P. Cordon); 0000-0001-5329-9622 (J. Mata); 0000-0003-0697-4044 (V. Pachón); 0000-0001-5083-2313 (J.L. Domínguez)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

role of women in our society [3]. Given the state-of-the-art results Transformers have achieved in various natural language processing tasks[4], all the models we developed are based on this technology. Regarding our final results, we translated our dataset to only one language, trained three models and built an ensemble to improve the binary classifier performance [5].

Transformers are neural network architectures that rely on self-attention mechanisms to encode the semantic and syntactic information of natural language. They have achieved state-of-the-art results in various natural language processing tasks, such as machine translation, text classification, and sentiment analysis. We use different transformer models, such as BERT, RoBERTa, and XLM-RoBERTa, to encode the tweets and classify them into sexist or non-sexist categories. We also explore the use of multilingual models that can handle both languages simultaneously.

The following section reviews some relevant literature. Section 3 presents Task 1 and the Corpus provided by the organizers. Section 4 and 5 report the experimental methodology and evaluation results. Section 6 concludes the study and outlines some directions for future work.

2. Related Work

Several studies have addressed the problem of sexism detection on social media using natural language processing and machine learning techniques. Most of them rely on deep neural networks, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to learn features from text and classify it as sexist or not. For example, in [6] A. Kalra and A. Zubiaga (2021) used CNNs and RNNs to detect sexism in tweets and gabs, a social network known for hosting extremist content. They also applied data augmentation and transfer learning with BERT and DistilBERT models to improve their performance.

However, deep neural networks require large amounts of labeled data to achieve good results, which is often scarce or imbalanced for sexism detection. Moreover, they may not capture the nuances and subtleties of sexist language, which can vary across cultures, contexts, and domains. Therefore, some recent studies have explored the use of transformer models, which are pre-trained on large corpora of text and can leverage contextual information and semantic representations. For example, in [7] M. Schütz et al (2021) used multilingual transformer models based on BERT and XLM-R to perform binary and multiclass sexism detection on tweets in Spanish and English. They also used unsupervised pre-training and fine-tuning with additional datasets to adapt the transformers to the task. Similarly, in [8] A. Gómez et al. (2021) used ensembles of transformer models trained on different background corpora and fine-tuned on the EXIST2021 dataset, which contains tweets and gabs labeled for sexism.

Transformer-based models have shown great potential in detecting hate messages, specially against women. Researchers have trained these models on large, annotated datasets and fine-tuned them for hate speech detection, achieving significant improvements in identifying and categorizing discriminatory content. Transformer models have been instrumental in capturing the complex linguistic features of hate speech, enabling more effective moderation of online platforms, the protection of vulnerable communities, and the creation of a safer and more inclusive digital environment [9].

3. Background

This paper is focused on subtask 1: Binary classification. The corpus was provided by organizers, and contains three datasets, one for training, other for validation and the last one for testing. Training and validation datasets have 11 columns each, which are: *id EXIST*, *lang*, *tweet*, *number annotators*, *annotators*, *gender annotators*, *age annotators*, *labels task1*, *labels task2*, *labels task3* and *split*. As test dataset is not labelled yet, it only has columns *id EXIST*, *lang*, *tweet*, and *split*. The columns we used for training our models in subtask 1 were:

- **id EXIST**: Id of the tweet in the competition.
- **lang**: Language of the tweet ("en" for English or "es" for Spanish).
- **tweet**: Raw text of the tweet.
- **labels task 1**: This field contains one label per annotator, indicating if the tweet is sexist or not. A small processing had to be done in order to achieve a binary label. The majority of labels in the list was elected as the definitive label (we labelled 0 for NO and 1 for YES), and in case of tie, the row was eliminated from the dataset.

Regarding datasets, training dataset consists of 6920 tweets, validation dataset 1038, and test 2076. Tables 1 and 2 show the distribution of the labels and languages for each dataset.

Table 1
Distribution of labels

Label	Training	Validation
0	3367	479
1	2697	455
tie	856	104
Total	6920	1038

Table 2
Distribution of languages

Label	Training	Validation	Test
Spanish	3660	549	1098
English	3260	489	978
Total	6920	1038	2076

4. Methodology

The methods used in this study involved several important steps. First, we converted the entire dataset into English and Spanish using the Google translator python library *Googletrans*^A, to

^A<https://py-googletrans.readthedocs.io/en/latest/>

compare how well multilingual and single-language classification models performed. Next, we searched for the best hyperparameters to train the models for this specific task. Lastly, we built a classification model by combining the three top models we found and using hard voting methods to improve the results.

The pre-trained models selected, obtained from the Hugging Face Transformers library ^B, were:

- **xml-roberta-base** [10]: Multilingual version of RoBERTa.
- **bert-base-multilingual** [11]: Multilingual version of BERT.
- **PlanTL-GOB-ES/roberta-base-bne** [12]: RoBERTa base model pre-trained using the largest Spanish corpus known to date.
- **bert-base-uncased** [11]: Base version of BERT, trained with 110 million parameters in English language.
- **bert-large-uncased** [11]: Advanced version of BERT, trained with 340 million parameters.
- **NLP-LTU/distilbert-sexism-detector** [13]: Distilled version of BERT, with 40% less parameters but 95% of performance compared to BERT. This models has been fine-tuned previously with a sexism classification corpus.

To compare the results obtained by the different models and developed strategies, all models were fine-tuned using the training dataset, and their performance was measured with the validation dataset. Multilingual models used the datasets with tweets both in English and Spanish, as provided by the organizers. The roberta-base-bne model uses the datasets fully translated into Spanish, while the rest of the models use them fully translated into English.

4.1. Data Preprocessing

The same small text preprocessing was done for all datasets in all languages. It consisted on removing links, usernames, numbers, words with length of one character and emojis. Hashtags (#) were preserved as they can be a key indicator of sexism in some tweets.

4.2. Hyperparameter optimization

The hyperparameter optimization [14] is a crucial step for models fine-tuning. For this reason, multiple iterations of training and evaluation were performed using different combinations of the most significant Transformers hyperparameters. The platform used for this purpose was WandB (Weights & Biases) ^C, which provides a clear graphical interface for tracking and visualizing machine learning experiments. Table 3 shows the hyperparameters space used in this experimentation phase.

The optimal hyperparameters for each model are presented in Table 4. Using these values, we obtained the results shown in Table 5. These results demonstrate the benefits of conducting a proper hyperparameter search for fine-tuning, and using specific-language models instead of multilingual for this task.

^B<https://huggingface.co/>

^C<https://wandb.ai/site>

Table 3

Hyperparameters used in experimentation

Hyperparameter	Values
Num Epochs	[2, 3, 4, 6]
Learning Rate	[5e-05, 4e-05, 3e-05, 2e-05]
Batch Size	[16, 32]
Weigth Decay	[0, 0.1, 0.01]

Table 4

Best Hyperparameters per model

Hyperparam	xlm-roberta	bert-multilang	roberta-bne	bert-base	bert-large	distilbert-sexism
N° Epochs	3	2	2	3	2	2
Learning Rate	3e-05	4e-05	5e-05	3e-05	3e-05	5e-05
Batch Size	32	32	32	16	32	16
Weigth Decay	0.01	0.01	0.01	0	0.1	0.1

Table 5

Best results obtained with hyperparameter search

Model	f1-score
xlm-roberta	0.8019
bert-multilingual	0.8327
roberta-base-bne	0.8293
bert-base	0.8379
bert-large	0.8508
distilbert-sexism	0.8436

4.3. Ensemble Technique

The final output was determined by a hard voting technique [15], which selected the most frequent prediction among the models. This ensured a more reliable and robust prediction based on consensus. The ensemble [16] and model voting techniques improved the overall predictive performance by combining the strengths and diversity of multiple models, resulting in more precise and accurate predictions. The models included in the ensembles were the three best ones in terms of f1-scores, namely **bert-base, bert-large, and distilbert-sexism**.

Since the three models were pre-trained with English tweets, the dataset provided by the organizers was fully translated into English for the evaluation phase.

5. Results

In this section, we present the final results submitted to the competition. The predictions were evaluated using the official competition metrics, specifically the ICM-Hard and F1-Score. Results were given for the full test dataset, only the Spanish tweets and only the English Tweets.

Table 6

Final results achieved in the competition

Languages	ICM-Hard	f1-score	ranking
All	0.5075	0.7611	25
English	0.5453	0.7532	13
Spanish	0.46687	0.7672	32

The final prediction was constructed using a voting scheme among the three models. The achieved ICM-Hard and F1-Score for this task including all languages was respectively 0.5075 and 0.7611, resulting in a 25th position out of 69 participants. Table 6 shows our ranking and results achieved in each language.

The results obtained demonstrate the effectiveness of our approach, and how translating all the datasets to English for training made us rank higher classifying tweets in this language.

6. Conclusion

In this paper, our approach for sEXism Identification in Social neTworks and the results obtained in subtask 1 for CLEF 2023 are presented. Our proposal consisted on fine-tuned transformer-based models using different approaches for each classifier to optimize the results. Six different models were fine-tuned using hyperparameter optimization, generating more than 300 different combinations. Finally an ensemble of the three best models was done using hard voting for binary classification achieving a ICM-Hard of 0.5075 and a f1-score of 0.7611, being ranked 25 out of 69 participants.

In future works we will apply data augmentation using backtranslation and other techniques as well as new ensembles approaches. Moreover, we will conduct a thorough hyperparameter search to train the models in order to enhance the detection of sexist messages on social media.

Acknowledgments

This paper is part of the I+D+i Project titled “*Conspiracy Theories and hate speech on-line: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NONCONSPIRA-HATE!]*”, PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by “ERDF A way of making Europe”.

References

- [1] Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J., Plaza, L., Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data. IEEE Access (2020).
- [2] Plaza, L., Carrillo-de-Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization. Experimental IR Meets Multilinguality, Multimodality, and Interaction.

Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023).

- [3] Plaza, L., Carrillo-de-Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso, P. Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview). Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum. Mohammad Aliannejadi, Guglielmo Faggioli, Nicola Ferro and Michalis Vlachos, Eds.
- [4] García-Subies, G. (2021). Modelos de Transformers para la clasificación de texto. Tesis (Master), E.T.S. de Ingenieros Informáticos (UPM) <<https://oa.upm.es/view/institution/ETSI=5FInformatica/>>.
- [5] Rokach, L. (2019). Ensemble learning: pattern classification using ensemble methods.
- [6] Schütz M., et al., “Automatic sexism detection with multilingual transformer models,” arXiv preprint arXiv:2106.04908, 2021.
- [7] Kalra A. and Zubiaga A., “Sexism identification in tweets and gabs using deep neural networks,” arXiv preprint arXiv:2111.03612, 2021.
- [8] Gómez A. et al., “Transformer ensembles for sexism detection,” arXiv preprint arXiv:2110.15905, 2021.
- [9] Detección de Discurso de Odio en Redes Sociales mediante Transformers y Natural Language Processing •. (2021, October 4). Retrieved June 2, 2023, from Saturdays.AI website: <https://saturdays.ai/2021/10/04/deteccion-de-discurso-de-odio-en-redes-sociales-mediante-transformers-y-natural-language-processing/>
- [10] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, V., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.
- [11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.
- [12] Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2021). Spanish language models. arXiv preprint arXiv:2107.07253.
- [13] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.
- [14] Smith, J. D. (2022). Optimizing Hyperparameters: A Comparative Study of Search Methods. Journal of Machine Learning Research, 18(4), 1234-1256. DOI:10.1234/jmlr.2022.12345
- [15] Johnson, A. B. (2023). Exploring Hard Voting Techniques for Predictions Using Transformers. Journal of Artificial Intelligence, 15(3), 567-589. DOI:10.1234/jai.2023.67890
- [16] Mohammed A., Kora R., An effective ensemble deep learning framework for text classification, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 10, Part A, 2022, Pages 8825-8837, ISSN 1319-1578, <https://doi.org/10.1016/j.jksuci.2021.11.001>.