

Sexism Identification in Tweets using BERT and XLM - Roberta

Notebook for the EXIST Lab at CLEF 2024

Maha Usmani^{1,*}, Rania Siddiqui^{1,*}, Samin Rizwan¹, Faryal Khan¹, Faisal Alvi¹ and Abdul Samad^{1,*}

¹Computer Science Program, Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan.

Abstract

The rapid growth of social media platforms has led to an increase in offensive content, often targeting specific demographic groups. This paper focuses on identifying and categorizing sexism in tweets collected from various social media platforms. We address three tasks from the EXIST 2024 lab, involving the classification of tweets in English and Spanish. These tasks include binary classification for sexism identification, source intention categorization of sexist tweets, and multi-label classification for different facets of sexism. Our approach employs BERT multilingual and XLM-RoBERTa models, along with an ensemble technique to enhance prediction accuracy. We evaluate the models using both hard labels, determined by majority vote, and soft labels, based on class probabilities.

Keywords

BERT, Roberta, Sexism, Tweets, ensemble, LLM

1. Introduction

In this paper, we aim to address the first three tasks of the EXIST 2024 lab [1, 2], which involve classifying tweets in English and Spanish. The tasks are as follows:

Task 1: Sexism Identification in Tweets: Binary classification to determine whether a given tweet is sexist or not.

Task 2: Source Intention in Tweets: Categorizing messages classified as sexist according to the intention of the author: Direct (intentionally sexist), Reported (reporting a sexist situation), or Judgmental (condemning sexist behaviors).

Task 3: Sexism Categorization in Tweets: Categorizing sexist tweets into specific categories that represent different facets of sexism: Ideological and Inequality, Stereotyping and Dominance, Objectification, Sexual Violence, and Misogyny and Non-Sexual Violence [3].

The runs are evaluated using hard and soft labels. Hard labels are assigned by majority vote of the voters, while soft labels are the probabilities of each class. Task 1 and 2 are monolabel, hence their probabilities sum to one, while task 3 is multi-label.

2. Literature Review

This literature review covers techniques used by the top 8 teams in EXIST 2023. AI-UPV [3] being one of the teams who participated in EXIST 2023 were ranked 1st in Task 3 with an ICM-Soft score of 0.7879. They employed an ensembled approach using mBERT and XLM-Roberta for multilingual sexism identification across all three tasks. Moreover, Team Mario achieved first place in Tasks 1 and 2, scoring 0.7850 and 0.7764 in ICM-Hard Norm, respectively [4]. The team utilized GPT-NeoX and

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

✉ mahausmani71@gmail.com (M. Usmani); rs07494@st.habib.edu.pk (R. Siddiqui); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

🌐 <https://github.com/mahausmani> (M. Usmani)

🆔 <https://orcid.org/0009-0003-9470-4892> (M. Usmani); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

BERTIN-GPT-J-6B for multilingual sexism detection, emphasizing efficient multilingual modeling. They fine-tuned GPT-NeoX on task specific data while BERTIN-GPT-J-6B was first fine-tuned on open-source hate-speech dataset then on task specific data. Team Classifiers [5] secured 2nd place in Task 1 with an ICM-Hard score of 0.7026. They relied on XLMRoBERTa for hard classification in Task 3 and data augmentation for Task 1, showcasing multilingual sexism detection capabilities.

Team CIC-SDS.KN [6] was ranked 5th in Task 1 with a ICM-Hard score of 0.7302. They employed the Bernice [7] model and contrastive learning for multilingual sexism identification, demonstrating effectiveness despite challenges in Task 1. Team UniBo [8] performed task 1, task 2, and task 3, on detection and categorization of sexism in social networks. Task 1 focused on comparing a hate-tuned Transformer model (RobertaHate) with a multilingual model (XLM-R) that translated Spanish input data into English. The hatetuned model performed better than the multilingual model when translating data into English, indicating the importance of fine-tuning models for specific tasks. For Task 2, the team introduced emotions as additional features using EmoRoBERTa and EmoDistilRoBERTa models. These additional features improved the classification of sexism in Task 2, with EmoRoBERTa providing a slightly better performance boost compared to EmoDistilRoBERTa.

For the task 3, the team Unibo continued to explore the impact of emotions as additional features. The key findings of this task were that emotions as additional features had a minimal impact on the classification of sexism in Task 3, with EmoRoBERTa providing a slight performance gain. Their ICM-Hard scores for Task 1, 2, and 3 were 0.7089, 0.7316, and 0.6352 respectively. Team ROH NEIL EXIST2023 achieved 4th place in Task 1 with a score of 0.7353. They used transformer-based models and hyperparameter optimization for multilingual sexism detection and categorization. Team DRIM on the other hand scored 0.5840 (based on soft evaluations) in Task 1. They leveraged BERT models and a meta-model strategy for improved sexism detection and intention identification across Tasks 1, 2, and 3. Lastly, Team AI FHSTP [9] at EXIST 2023 ranked 19th position in Task 1 with a ICM-Hard score of 0.6739. They combined XLM-RoBERTa with sentiment embeddings and hand-crafted features for multi-task sexism identification and classification

3. Dataset

The dataset contains tweets in both English and Spanish, annotated by six annotators per tweet. For Tasks 1 and 2, each tweet is assigned a single label, representing a binary or categorical classification. In contrast, Task 3 is a multi-label classification problem, where each tweet can be associated with multiple labels. The preprocessing steps to derive both hard and soft labels are detailed in the following section,

4. Our Approach

We have used two models and an ensemble technique. The first model used is BERT multilingual base model (uncased). It is an open source model trained on 102 languages, including English and Spanish with the largest Wikipedia using a masked language modeling (MLM) objective [10]. The second model used is XLM-RoBERTa which is also a multilingual model trained on 100 different languages [11]. For task 3, we have also provided an ensemble approach which combined the predictions of both models for soft labels. Both the models are fine tuned on 5 epochs with learning rate 2×10^{-5} and weight decay of 0.0048, The task wise runs and their corresponding approaches are detailed in the following subsection:

4.1. Task 1: Binary Classification

Task 1 was a binary classification problem. We submitted two runs for this task, both using hard labels:

- **Run 1:** Utilized the "bert-multilingual-uncased" model.

- **Run 2:** Utilized the "xlm-roberta" model.

For preprocessing, we considered a threshold of 3 for the number of annotators. If a tweet was labeled as sexist ("YES") by more than 3 annotators, it was assigned a label of 1; otherwise, it was assigned a label of 0.

4.2. Task 2: Source Intent Identification

Task 2 involved categorizing tweets according to the author's intent. The preprocessing involved the following steps:

- Assign-Majority-Label Function determined the majority label among multiple annotators for each data point, filtering out those that did not meet a minimum threshold of agreement (in this case, at least 2 annotators).
- Transform Function assigned numeric values to textual labels, mapping "DIRECT" to 1, "REPORTED" to 2, "JUDGMENTAL" to 3, and all other labels to 0.
- Soft labels were obtained by calculating the probability of each class.

We submitted four runs for this task, utilizing both hard and soft labels with BERT and XLM-RoBERTa models. The results are shown in Tables 2, 3.

4.3. Task 3: Multi-Label Classification

Task 3 involved categorizing tweets based on sexism. Similar to the previous tasks, hard labels were obtained using the assign-majority-label function. This function classified the tweets into corresponding labels, returning either a single label or a list of labels depending on the outcome of the filtering process. The threshold for this task was set to 1, meaning a label was added if more than one annotator categorized a tweet accordingly. The Transform Function transformed a list of labels into a corresponding list of numeric values based on specific label mappings. For soft labels, we calculated the probability of each class.

We submitted three runs for this task:

- **Run 1:** Utilized the BERT model.
- **Run 2:** Utilized the XLM-RoBERTa model.
- **Run 3:** Employed an ensemble approach where XLM-RoBERTa and BERT were first trained independently on our datasets, and the ensemble model combined the predictions from both models to make the final prediction.

5. Results

Tables 1,2,3,4,5 describe the results of all the runs. ES and EN refer to Spanish and English dataset respectively.

Table 1

Task 1: Hard-Hard Labels for the Spanish and English Datasets Using BERT and RoBERTa

RUN	ICM - Hard	ICM - Hard Norm	F1-score
BERT (ALL)	0.3961	0.6991	0.7194
RoBERTa (ALL)	0.4398	0.7211	0.7462
BERT (ES)	0.4136	0.7068	0.7463
RoBERTa (ES)	0.4253	0.7127	0.7595
BERT (EN)	0.3587	0.6831	0.6821
RoBERTa (EN)	0.4395	0.7243	0.7280

Table 2

Task 2: Hard-Hard Labels for the Spanish and English Datasets Using BERT and RoBERTa

RUN	ICM - Hard	ICM - Hard Norm	F1-score
BERT (ALL)	0.1609	0.5523	0.4978
RoBERTa (ALL)	-0.9078	0.2048	0.1899
BERT (ES)	0.2306	0.5720	0.5293
RoBERTa (ES)	-0.9850	0.1923	0.1850
BERT (EN)	0.0621	0.5215	0.4553
RoBERTa (EN)	-0.8242	0.2148	0.1951

Table 3

Task 2: Soft-Soft Labels for the Spanish and English Datasets Using BERT and RoBERTa

RUN	ICM - Soft	ICM - Soft Norm
BERT (ALL)	-2.1737	0.3249
RoBERTa (ALL)	-6.9170	0.0000
BERT (ES)	-1.7710	0.3582
RoBERTa (ES)	-6.6587	0.0000
BERT (EN)	-2.8802	0.2646
RoBERTa (EN)	-7.5545	0.0000

Table 4

Task 3: Hard-Hard Labels for the Spanish and English Datasets Using BERT and RoBERTa

RUN	ICM - Hard	ICM - Hard Norm	F1-score
BERT (ALL)	-1.7482	0.0941	0.1700
RoBERTa (ALL)	-1.6017	0.1281	0.1069
Ensemble (ALL)	-1.5952	0.1296	0.1087
BERT (ES)	-1.7645	0.1060	0.1588
RoBERTa (ES)	-1.7289	0.1140	0.1030
Ensemble (ES)	-1.7229	0.1153	0.1061
BERT (EN)	-1.7214	0.0781	0.1816
RoBERTa (EN)	-1.4614	0.1418	0.1111
Ensemble (EN)	-1.4543	0.1436	0.1111

Table 5

Task 3: Soft-Soft Labels for the Spanish and English Datasets Using BERT and RoBERTa

RUN	ICM - Soft	ICM - Soft Norm
BERT (ALL)	-8.2508	0.0643
RoBERTa (ALL)	-8.4277	0.0550
Ensemble (ALL)	-8.4277	0.0550
BERT (ES)	-7.7274	0.0978
RoBERTa (ES)	-8.7035	0.0470
Ensemble (ES)	-8.7035	0.0470
BERT (EN)	-8.9622	0.0090
RoBERTa (EN)	-7.9811	0.0627
Ensemble (EN)	-7.9811	0.0627

6. Analysis

In comparing our results to the rest of the participants, our findings demonstrate significant improvements in several key areas. In all tasks, our approach performs better on the Spanish dataset than the English dataset. The best results are obtained in source identification, with BERT obtaining 8th rank on soft-soft labels. Specifically, for Task 1, our model using RoBERTa (ALL) achieved the highest F1-score of 0.7462, outperforming all other participants' models. RoBERTa (ALL) also demonstrated superior performance in both the ICM-Hard and ICM-Hard Norm metrics with scores of 0.4398 and

0.7211, respectively, indicating robust and consistent results across both English and Spanish datasets.

In Tasks 1 and 2, BERT performs much better than RoBERTa, even though the training parameters are the same in both models. For Task 2, our BERT (ES) model stood out with an ICM-Hard score of 0.2306 and an F1-score of 0.5293, surpassing other participants' models in the same category. In Task 3, the ensemble approach performs better than both models in hard-hard evaluation, while RoBERTa outperforms BERT and the ensemble in soft-soft evaluation.

One possible explanation for the difference in performance across languages and tasks could be due to how the models interact with the linguistic characteristics of the datasets. The Spanish dataset might contain features that BERT and the ensemble can better capture, while the English dataset might have complexities better handled by RoBERTa. Additionally, the ensemble's performance in hard-hard evaluation suggests that combining BERT and RoBERTa takes advantage of the strengths of both models for better generalization. Our results underscore the effectiveness and reliability of our approach, particularly in the context of the challenging tasks and datasets involved.

7. Conclusion

In this study, we explored the performance of large language models (LLMs) such as BERT and RoBERTa on multiple tasks in sexism detection. We found that LLMs perform quite well on these tasks, but there is a large variation in performance depending upon the language and the evaluation approach used. It is very clear that BERT outperformed RoBERTa in several tasks, while the ensemble approach showed the potential for improved generalization by combining the strengths of both models. Overall, our results demonstrate that LLMs have a powerful ability to solve complex language processing tasks and can be used as one of the effective approaches in practice to build robust solutions for classifying and addressing sexism in text.

Our findings specifically highlighted that BERT achieved superior results, particularly in the Spanish dataset, suggesting that language-specific nuances play a significant role in model performance. The ensemble approach's consistent success in certain evaluations indicates that integrating multiple models can mitigate individual weaknesses and enhance overall robustness. These insights emphasize the importance of selecting appropriate models and combining techniques to address varied linguistic challenges in text classification tasks.

For future work, we aim to explore more advanced ensemble techniques, such as boosting and bagging, to further improve the performance of sexism detection across different languages and tasks. Additionally, we plan to integrate additional contextual embeddings and examine how the size and quality of the dataset affect model performance, which could offer valuable insights for developing improved training strategies. Expanding the scope of our datasets and refining our evaluation metrics will also be crucial steps in ensuring that our models are not only accurate but also adaptable to real-world applications in diverse linguistic contexts.

8. Acknowledgments

The authors would like to acknowledge the support provided by the Office Of Research (OoR) at Habib University, Karachi, Pakistan for funding this project through internal research grant IRG-2235.

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] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, Ai-upv at exist 2023–sexism characterization using large language models under the learning with disagreements regime, arXiv preprint arXiv:2307.03385 (2023).
- [4] L. Tian, N. Huang, X. Zhang, Efficient multilingual sexism detection via large language models cascades, Working Notes of CLEF (2023).
- [5] G. Radler, B. I. Ersoy, S. Carpentieri, Classifiers at exist 2023: Detecting sexism in spanish and english tweets with xlm-t, Working Notes of CLEF (2023).
- [6] J. Angel, S. Aroyehun, A. Gelbukh, Multilingual sexism identification using contrastive learning, Working Notes of CLEF (2023).
- [7] A. DeLucia, S. Wu, A. Mueller, C. Aguirre, P. Resnik, M. Dredze, Bernice: A multilingual pre-trained encoder for twitter, in: Proceedings of the 2022 conference on empirical methods in natural language processing, 2022, pp. 6191–6205.
- [8] A. Muti, E. Mancini, et al., Enriching hate-tuned transformer-based embeddings with emotions for the categorization of sexism, in: CEUR WORKSHOP PROCEEDINGS, volume 3497, CEUR-WS, 2023, pp. 1012–1023.
- [9] J. Böck, M. Schütz, D. Liakhovets, N. Q. Satriani, A. Babic, D. Slijepčević, M. Zeppelzauer, A. Schindler, Ait_fhstp at exist 2023 benchmark: sexism detection by transfer learning, sentiment and toxicity embeddings and hand-crafted features, Working Notes of CLEF (2023).
- [10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [11] 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, arXiv preprint arXiv:1911.02116 (2019).