CEUR-WS.org/Vol-3497/paper-090.pdf

ZaRa-IU-NLP at EXIST 2023 - Sexism Identification: Specialized or Generalized?

Zackary Leech[1], Ravi Regulagedda[2] and Sandra Kübler[1]

[1]*Department of Linguistics, Indiana University, Bloomington, IN, USA*
[2]*Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Abstract

We present our approach to EXIST (sEXism Identification in Social neTworks) Task 1, at CLEF 2023, comprising automatic sexism detection in both English and Spanish on Twitter data. We compare two methods, the first being a bi-ensemble method that combines two pre-trained BERT architectures, BETO and RoBERTa, for each specific language, Spanish and English, respectively. The second method utilizes the larger multilingual transformer, RoBERTa-XLM-base, and considers the entire dataset despite language differences. We show that the language-specific ensemble performs better than the generalized model and is a better choice when looking at sexism detection in mixed Spanish and English data.

Keywords

sexism detection, RoBERTa, BETO, language-specific classification, language-agnostic classification

1. Introduction

With the rise in global social media interaction, there is also a rise in the shared expression of the darker side of humanity, including sexist sentiment. Santos-Rios et. al. [1] point out that “in 2021, 56% of the global population were social media users". UN Women [2] point to the COVID-19 Pandemic as one contributing factor for the rise, specifically for women, as “... women’s lives shifted online for work, education, access to services and social activities” leading to “rapidly escalated” violence against them. A global study by the Economist Intelligence Unit [3] found that “38% of women have personal experiences of online violence and 85% of women who are online have witnessed digital violence against other women”.

As such, collaborative work to address this growing problem has gained urgency. The sEXism Identification in Social neTworks (EXIST). EXIST 2023 [4, 5] is the third edition of EXIST at CLEF, it builds on EXIST 2021, the first shared task to address sexism in a broad sense, including detection of implicit bias. The shared task defines sexism as: The tweet is sexist itself, describes a sexist situation, or criticizes a sexist behavior, rather than solely explicit bias sexism, typically including direct name calling or harmful, stereotypical generalizations. The current EXIST shared task expands the task to include bilingual data. This is important as the Economist Intelligence Unit reports that “... there are regional differences in the prevalence of online

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece
zleech@iu.edu (Z. Leech); raregul@iu.edu (R. Regulagedda); skuebler@indiana.edu (S. Kübler)
https://cl.indiana.edu/~skuebler/ (S. Kübler)
0000-0003-0885-5436 (S. Kübler)
© 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)

violence against women” with “the overall prevalence rate by region at 76% in North America and 91% in Latin America and Caribbean” [3].

From outwardly explicit misogyny to more subtle implicit misogyny, automatic classification of sexism will create a more efficient process of creating a safer environment online. With rising hate speech towards women online and little research into the detection of sexism, contributions to this issue are urgent.

We present the approach by team ZaRa-IU-NLP. This approach was developed during a course on machine learning in NLP. Our work focuses on comparing a multilingual neural model with an ensemble of language specific models.

The remainder of this report will proceed as follows: Section 2 outlines previous research for prior EXIST shared tasks, Sections 3 and 4 provide details on the task description, the dataset, and the preprocessing methods, respectively. Section 5 describes the two rival model architectures. In section 6 we analyze the results of the experiments before concluding in section 7.

2. Related Work

Team avacaondata [6] provided the winning approach to the 2022 EXIST Task 1; they utilized an ensemble of transformer models using BERTweet-large, RoBERTa, and DeBERTa v3 for English, and BETO, BERTIN, MarIA-base, and Robertuito for Spanish. With this combination, the team achieved an overall F1 of 0.7996. To reach optimal performance given the computational load of the models and to avoid overgeneration, the training was carried out in 2 phases, choosing to optimize parameters with smaller amounts of data before expanding to the entire dataset [4, 5].

Team CIMATCOLMEX ranked first in the evaluation of Spanish tweets, with an accuracy of 0.7801 [4]. This team utilized an ensemble of 10 RoBERTuito and 10 BERT models. They reached the highest scores for Spanish, surpassing avacaondata in F1 score by 2.27% absolute. Using a bi-ensemble method, Villa-Cueva et. al [7] merged two transformers ensembles, one for Spanish and one for English. Though there is a high computational cost, this model scored second in Task 1 with a difference of 0.0038 in F1 to the winning team [4]. In regard to preprocessing, team CIMATCOLMEX used the following normalization steps: lowercasing, removing emojis, replacing usernames with “@user”, replacing any URL with the token “<URL>”, and removing any whitespace at the beginning or end of the tweet [4].

3. Task Description and Dataset

EXIST provides an opportunity to address the issue of online sexism with a wider reach, by providing participants with data containing bilingual, sexist speech, in both Spanish and English. Task 1 focuses on a binary classification of whether the text is sexist or not for both explicit or implicit examples. The data is constructed out of tweets with any form of oppression or prejudice against women because of their sex, explicit or implicit. It is worth noting that while the dataset labels every tweet as either English or Spanish, it includes tweets mixing both of the languages.

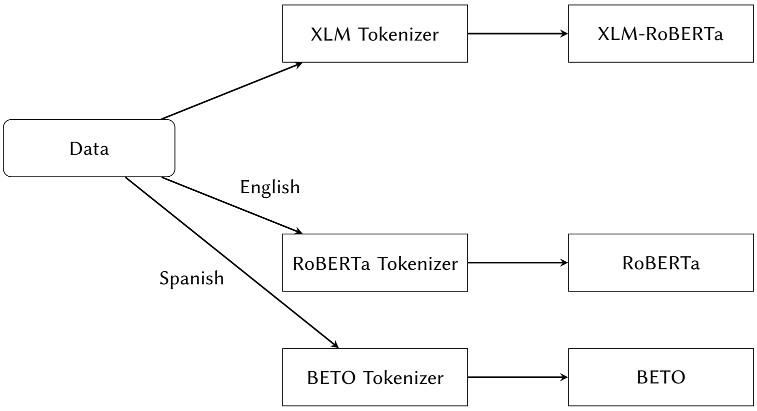

Figure 1: Data Flow during training for both approaches.

The EXIST 2023 dataset consists of 6 920 tweets for training, 1 038 tweets for validation, and 2 076 tweets for testing, where all sets are randomly selected from the 9 000 and 4 000 sets sampled and created by the CLEF 2023 organizers, for balance. The annotation process was carried out by a balanced group of 3 women and 3 men, in order to avoid gender bias. When choosing the label for a given tweet from these 6 labels, we used a simple majority. In the case of ties, we randomly selected either of YES or NO.

4. Preprocessing

Our focus is on a comparison of two model architectures, a generalized model, processing both English and Spanish tweets, and a specialized model using language specific classifiers for each language. As such, preprocessing was kept to a minimum. Tweets were used with URLs, emojis, duplications of characters, etc in place. Further, the dataset included a set of tweets containing varying levels of both languages. No alterations were made to the content of these individual bilingual tweets. Rather, each architecture was given an opportunity to make judgement on tweets based on the language information provided by the shared task. I.e., tweets labeled English were passed through RoBERTa and tweets labeled Spanish were passed through BETO. The fluidity of language, specifically in the informal context of twitter, creates ample opportunity for bilingual users to hide implicit sexism through language switching and as such, should be included in each model's ability to detect implicit bias.

5. Model Architecture

In our work, we compare two different approaches, a multilingual model, and a language specific model. The multilingual model uses XLM-RoBERTa, while the language specific model, RoBERTa for English, and BETO for Spanish. Figure 1 shows an overview of the data flow in the two models.

5.1. XLM-RoBERTa

One of the first approach to sexism detection in a Spanish-English context was using XLM-RoBERTa [8]. This is a multi-lingual version of RoBERTa [9] trained on a database of 100 languages. It was initially pre-trained for masked language modeling (MLM).

We also use this model for our multilingual approach and finetune it on the EXIST dataset on the binary classification task. The data is tokenized using XLM-RoBERTa's tokenizer before passing into the model that was finetuned in 3 epochs. The finetuned model is used for inference.

5.2. RoBERTa + BETO

This model is a pipeline composed of two models pre-trained on English and Spanish texts respectively. For English, we chose RoBERTa, a model trained on English texts using an MLM objective. RoBERTa is an improved model over the core BERT architecture with more parameters and trained on a larger corpus to provide a more robust performance than the base BERT model.

We chose BETO [10], a BERT-based Spanish language model for the Spanish language tweets in the dataset. BETO was trained on a large Spanish corpus [11]. Similar to RoBERTa, it was pre-trained on an MLM objective.

We fine-tuned both these models using the data for their respective languages. Since the data was split fairly evenly between English and Spanish texts, RoBERTa and BETO had about half the training data to finetune in comparison to XLM-RoBERTa.

For this ensemble, we performed tokenization based on the specific language using the pre-trained tokenizer. The data is then passed into the corresponding model.

5.3. Evaluation

We participated in the HARD-HARD evaluation, i.e., we provided a single label per tweet, which was evaluated against the gold label. For the official evaluation on the test set, we report ICM and F1 for the positive class (sexist). For our internal results on the development set, we report the macro-averaged F1 score, the F1 score per class, and ICM. ICM is a score developed to measure the similarity between two datapoints more accurately [12]. It is a generalization of Pointwise Mutual Accuracy (PMA) and compares the closeness between two different outputs by comparing them to a ground truth value.

6. Results

6.1. Official Results

Table 1 shows the official scores on the test set. These results show the ICM metric scores and the F1 scores of the positive class for the two models in the HARD-HARD evaluation ZaRa-IU-NLP_1 is based on the multilingual XLM-RoBERTa, and ZaRa-IU-NLP_2 on a combination of RoBERTa and BETO. Our scores show that the combination of language specific models outperforms the generalized model in each metric, even though the individual language models were trained on only half the data. The models produce consistent results across languages for sexist tweets, with the language specific models producing an F1 score of 0.7332 for Spanish

Table 1
Official results on the test set.

Model	Language	Rank	ICM Hard	ICM Hard Norm	Macro F1
ZaRa-IU-NLP_1	All	47	0.2842	0.5471	0.6955
	Spanish	53	0.1935	0.4661	0.6956
	English	45	0.3692	0.6287	0.6954
ZaRa-IU-NLP_2	All	42	0.3914	0.6154	0.7305
	Spanish	44	0.3056	0.5404	0.7332
	English	38	0.4609	0.6844	0.7263
Baseline	All		-0.4261		0.3272

Table 2
Results comparing the two approaches on the validation set.

Model	Language	F1 sexist	F1 non-sexist	Macro F1	Precision	Recall	ICM
ZaRa-IU-NLP_1	All	0.7248	0.7382	0.7315	0.7155	0.7910	0.3168
	Spanish	0.6823	0.7631	0.7631	0.7207	0.8108	
	English	0.7750	0.7349	0.7349	0.7081	0.7638	
ZaRa-IU-NLP_2	All	0.7610	0.7828	0.7719	0.7597	0.8027	0.4444
	Spanish	0.7061	0.7912	0.7912	0.7341	0.8581	
	English	0.8289	0.7639	0.7639	0.8051	0.7268	

tweets and 0.7263 for English tweets, while the single, large model reached 0.6956 and 0.6954, respectively.

6.2. Results on the Validation Set

Table 2 compares the performance of our two models on the validation set, including language specific results for both architectures. The results are in line with the official results, ZaRa-IU-NLP_2, the combination of language specific models, outperforms the multilingual ZaRa-IU-NLP_1 with an ICM of 0.4444 as compared to 0.3168. This shows that it is possible to obtain solid results given a small set of data and that the quality of the data (wrt. the language) is more important than the size of the training data. A look at the F1-scores per class shows a balanced performance, the score for the sexist class is only slightly lower than the one for the majority class of non-sexist tweets.

We then had a closer look at precision and recall per language and per class. These results are shown in Table 3. These results show that the multilingual model ZaRa-IU-NLP_1 shows the same preference for the non-sexist class given the higher recall for this class. However, this trend is much more pronounced for Spanish where the recall for non-sexist reaches 81.08% while the recall for the sexist class is at 63.24%. In the combined model ZaRa-IU-NLP_2, the Spanish classifier has a preference for the non-sexist class (recall: 85.81) while the English classifier has a preference for the sexist class (recall: 86.08). These results suggest that we might be able to gain better performance for Spanish if we upsample the sexist class. We leave this for future research.

Table 3
Results comparing precision and recall per language.

		Sexist		Non-sexist	
Model	Language	Precision	Recall	Precision	Recall
ZaRa-IU-NLP_1	Spanish	74.07	63.24	72.07	81.08
	English	80.08	75.09	70.82	76.39
ZaRa-IU-NLP_2	Spanish	79.31	63.64	73.41	85.81
	English	79.93	86.08	80.51	72.63

7. Conclusion

Overall, the study addresses the difficulties of online sexism detection in social networks in a bilingual setting. We carried out a comparison between two distinct architectures: the first, a large, multilingual model that processes the entire data set, and the second, a combination of language specific, smaller models that implement a split in language categorization for processing. Given our minimal pre-processing, the higher accuracy of the language specific model can be attributed to the more specialized language models. Our results, when tested on the developmental set, show that the combination of two specialized models outperforms the single, generalized one by a difference of 3 percent points in macro-averaged F1 score, and a difference in ICM of 0.13. Overall, this study contributes to the process of automatic sexism detection in social networks in highlighting the greater efficiency and accuracy of two specialized models, rather than a multi-lingual model.

As described above, this system was developed as a project in a course on machine learning. For this reason, the system was intentionally kept simple, so that it could be carried out in a short amount of time. This is the reason, for example, why we focused on comparing the two systems without preprocessing the data, and without investigating other language models, etc.

References

[1] R. Santos-Rios, J. Vilares, M. A. Alonso, Some experiments on the use of natural language processing for sexism detection and classification in social media, in: A. Leitao, L. Ramos (Eds.), Proceedings of V XoveTIC Conference (XoveTIC), volume 14 of *Kalpa Publications in Computing*, EasyChair, 2023, pp. 24–27. URL: https://easychair.org/publications/paper/rdrm. doi:10.29007/8z61.

[2] U. N. Women, Accelerating efforts to tackle online and technology-facilitated violence against women and girls, https://www.unwomen.org/en/digital-library/publications/2022/10/accelerating-efforts-to-tackle-online-and-technology-facilitated-violence-against-women-and-girls, 2022.

[3] The Economist Intelligence Unit, Measuring the prevalence of online violence against women, The Economist (2021). URL: https://onlineviolencewomen.eiu.com/.

[4] 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, Thessaloniki, Greece, 2023.

[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 (Extended Overview), in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[6] A. Vaca-Serrano, Detecting and classifying sexism by ensembling transformers models, in: Proceedings of IberLEF, A Coruña, Spain, 2022.

[7] E. Villa-Cueva, F. Sánchez-Vega, A. P. López-Monroy, Bi-ensembles of transformer for online bilingual sexism detection, in: Proceedings of IberLEF, A Coruña, Spain, 2022.

[8] 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, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[9] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[10] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: Practical ML for Developing Countries Workshop @ ICLR 2020, Addis Ababa, Ethiopia, 2020.

[11] J. Cañete, Compilation of large Spanish unannotated corpora, 2019. URL: https://doi.org/10.5281/zenodo.3247731. doi:10.5281/zenodo.3247731.

[12] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 5809–5819.