Multilingual Sexism Identification Using Contrastive Learning

Notebook for the CIC-IPN Lab at CLEF 2023

Jason Angel[1,*], Segun Taofeek Aroyehun[2] and Alexander Gelbukh[1,*]

[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico
[2]University of Konstanz, Konstanz, Germany

Abstract

We present our systems and findings for the Exist2023 (subtask 1), a shared task for multilingual sexism identification at CLEF 2023 [1]. Our system aims to accurately identify and evaluate the degree of sexism in social media content in a multilingual setting considering its subjective nature. We successfully integrated two variations of contrastive learning as an intermediate step in a conventional fine-tuning language model pipeline. Our approach not only outperformed the sole fine-tuned method but also achieved competitive results compared to the top scores in the competition. This substantiates the simplicity and benefits of our approach to the task of sexism identification.

Keywords

Sexism identification, contrastive learning, learning with disagreement, multilingual natural language processing

1. Introduction

Sexism is a form of discrimination rooted in biased beliefs, stereotypes, and the oppression of individuals, often targeting women due to their sex/gender. In today's era, where social networks wield significant influence, it is vital to acknowledge and combat sexism. This harmful mindset perpetuates inequality, limits opportunities, and reinforces oppressive power dynamics, hindering progress toward a fairer society.

Nevertheless, the automatic and reliable identification of sexist statements poses significant challenges due to their subjective nature. This research aims to propose an approach to identifying sexism taking into account varying opinions on whether a message can be considered sexist or not. We conducted experiments using a multilingual language model on Spanish and English messages and explored two variations of incorporating contrastive learning in a typical NLP pipeline in order to cluster the "degree of sexism" present in a message.

The document continues as follows: Section 2 outlines the distinctive characteristics of the

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
ajason08@gmail.com (J. Angel); aroyehun.segun@gmail.com (S. T. Aroyehun); gelbukh@cic.ipn.mx (A. Gelbukh)
https://ast123.github.io/ (S. T. Aroyehun); www.gelbukh.com (A. Gelbukh)
0000-0002-7991-1979 (J. Angel); 0000-0002-2571-3731 (S. T. Aroyehun); 0000-0001-7845-9039 (A. Gelbukh)

Exist2023 dataset, enabling the analysis of how perceptions of sexism can be influenced by gender and age groups across two distinct languages. In Section 3, we offer a detailed description of our experimental approach. Section 4 summarizes our results and offers an interpretation of our findings. Section 5 provides a concise overview of related research on sexism identification. Lastly, in Section 6 we conclude by highlighting our contributions and outlining avenues for future research.

2. Sexism Identification Dataset

The dataset provided by the Exist2023 initiative consists of nearly 10K tweets (around 5.3K for Spanish and 4.7K for English) carefully selected (to mitigate terminology, temporal and author biases) from more than 8M tweets from 1st September 2021 till the 30th September 2022. The dataset was roughly split into train, dev, and test 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. What sets this dataset apart is its thoughtful consideration of the subjectivity inherent in identifying sexism. To accommodate this, the dataset follows the learning with disagreements paradigm, where multiple annotators (six in this case) offer diverse perspectives. Furthermore, to address potential "label bias" resulting from socio-demographic differences among annotators, each annotator represents a unique socio-demographic profile, including gender (MALE, FEMALE) and age group (18-22, 23-45, and 46+).

Although there are no gold annotations, the majority vote from the label annotations suggests the proportion of sexist content existing in the dataset, which is further used for evaluation purposes as a "hard label". Table 1 combines samples from train and dev split to showcase the distribution of labels per language in terms of majority votes as Sexist, not-sexist, and undetermined i.e, when three annotators consider the tweet sexist and the other three as non-sexist

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

Majority vote	Spanish	English
Sexist	43.3%	35.5%
Not-sexist	44.3%	52.9%
Tie	12.5%	11.6%

3. System description

We fine-tune Bernice [2] a multilingual RoBERTa language model that specializes in processing language from the Twitter domain, which allowed us to handle two important aspects of the Exist2023 dataset: the presence of English and Spanish samples, and the particularities of the informal language used in social networks such as Twitter, including the processing of emojis and hashtags.

Our experiments differentiate because of the addition of the contrastive learning technique to the typical fine-tuning language model pipeline which learns an embedding space in which similar sample pairs stay close to each other while dissimilar ones are far apart, in our contrastive learning approach we used a regression setting where the labels we used were calculated as the fraction of annotators that answer "Yes" divided by six (the total number of annotators). In this manner, our method takes into consideration the subjectivity and the diversity of views of the task in the paradigm of learning from data with disagreements.

We also leverage the different annotators' labeling for making the final predictions, hence we cast the Sexism identification task as a regression task that predicts the fraction of annotators that indicated "YES". To derive the hard label, we use a rule where the model prediction has to be greater than 0.5 to predict "YES" and "NO" otherwise.

The following summarizes our submitted systems:

1. **FT:** fine-tuning of the language model for a maximum of 30 epochs with early stopping. Listed on the official leaderboard as "CIC-SDS.KN_1"
2. **Freeze_CL:** we added contrastive learning before fine-tuning the model and freezing the model to only train the classifier head. Listed on the official leaderboard as "CIC-SDS.KN_2"
3. **Unfreeze_CL:** the same as the second run except that the fine-tuning step updates all of the model parameters. Listed on the official leaderboard as "CIC-SDS.KN_3"

We train with the contrastive learning objective for 10 epochs with a learning rate of $5e-5$ and a batch size of 32, we follow mainly the settings reported in [3]. In the subsequent fine-tuning step we train for a maximum of 20 epochs (in order to have a comparable setting with the FT setting with 30 epochs) with early stopping, learning rate of $1e-5$, and batch size of 128. We use the AdamW optimizer [4]. We use the transformers library [5] to train our models on an NVIDIA V100 GPU with 32GB memory. We save the model with the lowest root mean square error (rmse) score on the validation set during training. We then use the saved model to make predictions on the unseen test set.

4. Results

Our systems ranked among the top-3 teams according to the normalized ICM metric (Information Contrast Measure). The ICM metric [6] 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 the hard ground truth

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

A summary of our experiments is presented in Table 2 where we use the Hard-Hard, Soft-Soft, and Hard-Soft evaluation schemes to compare our model results with those obtained by the baseline "majority_class" that classifies all instances as the majority class, and the best models submitted to Exist2023-task-1 (which are publicly available in the original leaderboard) in the competition, which we refer to as the "best score" and correspond to the score obtained by the model with the highest performance in that specific evaluation. We also provide results for Spanish only, English only, and both Spanish and English.

Table 2
Performance scores on the test set using the ICM metric (Information Contrast Measure). The columns H/H, S/S, and H/S refer to the hard-hard, soft-soft, and hard-soft types of evaluation respectively. The best scores obtained by our submissions are highlighted.

	ALL samples			Spanish samples			English samples		
Model	H/H	S/S	H/S	H/H	S/S	H/S	H/H	S/S	H/S
Baseline	0.085	0.115	0.115	0.014	0.006	0.006	0.163	0.233	0.233
Best score	0.785	0.642	0.573	0.801	0.620	0.575	0.769	0.668	0.571
FT	0.704	0.613	0.528	0.688	0.569	0.510	0.721	**0.660**	0.547
Freeze_CL	**0.730**	**0.625**	**0.541**	**0.709**	**0.595**	**0.524**	**0.752**	**0.657**	**0.559**
Unfreeze_CL	0.726	0.618	0.538	**0.709**	0.584	**0.524**	0.742	0.655	0.553

4.1. Analysis of results

Our results show clearly that our systems with contrastive learning (Freeze_CL and Unfreeze_CL) perform better than the just fine-tuning model across all evaluation schemes and language slices for Spanish and English, hence demonstrating that the addition of contrastive learning as an intermediate step benefits the model's ability to correctly identify sexist content. Specifically, between our two approaches for contrastive learning, the "Freeze" is slightly better than the "Unfreeze" model, showing that the knowledge gained by the contrastive learning step is not forgotten to a big extent by updating the previously learned parameters. With respect to the baseline "majority class", its weakness is quite evident and also non-informative, but it suggests how complex the task is without proper modeling of the phenomenon. We also remark on the outstanding performance obtained by the best models in the competition, which we refer to in Table 2 as the "best score" for each evaluation individually, and obtained far superior scores compared with our proposed models in some evaluation scenarios such as the Spanish hard-hard evaluation. We hypothesize that this effect may be attributed to the multilingual nature of our model, which offers the advantage of utilizing a single model for multiple languages. However, as we observed, a multilingual model may also show performance variation across languages. We leave as future work the investigation of factors that are likely to explain the observed variation.

5. Related works

In recent years, NLP tasks promoting tolerance and respect, including Hate Speech detection [7], Stereotype identification [8], and gender bias mitigation [9], have gained significant popularity and strong support within the NLP community. Among these tasks, the identification of Sexism has emerged as a distinct field of investigation, evolving from being a subsection of hate speech detection [10] primarily conducted in English, to a standalone task studied in multiple languages such as French [11], Chinese [12], and even lesser-resource languages like Romanian [13]. However, apart from the previous Exist initiatives [14, 15], which primarily concentrated on English and Spanish datasets, there has been limited exploration of modeling and analyzing sexism phenomena from a multilingual perspective.

6. Conclusion

Sexism continues to be a significant societal concern, gaining increased attention as social media platforms play an ever-growing role in our lives. The need to address and mitigate sexism on these platforms has become paramount. In light of this, our study focused on developing effective multilingual sexism identification systems using contrastive learning. Our findings demonstrate the superiority of our proposed systems, which incorporated contrastive learning with and without updating learned parameters, over the traditional fine-tuning approach.

The results obtained from our experiments exceeded the performance of solely fine-tuned models and proved to be highly competitive compared to the best scores achieved in the competition. This outcome underscores the value of exploring the integration of contrastive learning techniques into the traditional pipelines to further advance the field of content moderation. Moving forward, further exploration and refinement of contrastive learning approaches hold the potential to enhance the accuracy and efficiency of sexism detection systems, leading to more inclusive and equitable online spaces.

6.1. Future work

Further Research in this field holds exciting prospects. Firstly, we intend to extend the evaluation of our contrastive learning approach to additional sexism datasets and explore its applicability in related tasks such as hate speech detection. Secondly, a more comprehensive analysis is needed to understand how language models handle the inherent subjectivity of the task, considering varying perspectives from annotators with diverse socio-demographic profiles. Lastly, while our participation in the binary classification task of Exist2023 was fruitful, we are eager to investigate the potential application of our approach in a multiclass setting. These avenues of exploration promise to deepen our understanding and improve the effectiveness of multilingual sexism identification systems.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20232138, 20232080, 20231567 of the Secretaría de

Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: European Conference on Information Retrieval, Springer, 2023, pp. 593–599.

[2] 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, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6191–6205. URL: https://aclanthology.org/2022.emnlp-main.415.

[3] H. Sedghamiz, S. Raval, E. Santus, T. Alhanai, M. Ghassemi, SupCL-Seq: Supervised Contrastive Learning for downstream optimized sequence representations, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3398–3403. URL: https://aclanthology.org/2021.findings-emnlp.289. doi:10.18653/v1/2021.findings-emnlp.289.

[4] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[6] E. Amigó, 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), 2022, pp. 5809–5819.

[7] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.

[8] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, F. D'Errico, Detecting racial stereotypes: An italian social media corpus where psychology meets nlp, Information Processing & Management 60 (2023) 103118.

[9] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in:

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1630–1640.

[10] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th international workshop on semantic evaluation, 2019, pp. 54–63.

[11] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4055–4066.

[12] A. Jiang, X. Yang, Y. Liu, A. Zubiaga, Swsr: A chinese dataset and lexicon for online sexism detection, Online Social Networks and Media 27 (2022) 100182.

[13] A. Moldovan, K. Csürös, A.-M. Bucur, L. Bercuci, Users hate blondes: Detecting sexism in user comments on online romanian news, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), 2022, pp. 230–230.

[14] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[15] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.