IUEXIST: Multilingual Pre-trained Language Models for Sexism Detection on Twitter in EXIST2023

Notebook for the EXIST Lab at CLEF 2023

Yash A. Hatekar[1], Muhammad S. Abdo[1], Snigdha Khanna[1] and Sandra Kübler[1]

[1]*Indiana University, Bloomingtonm, IN, USA*

Abstract

We describe an approach towards sexism detection in tweets, for the EXIST 2023-Task 1, a shared task on sexism identification. The dataset for this task consists of English and Spanish tweets. Task 1 is a binary classification task, where our system needs to decide whether a given tweet contains sexist expressions or behaviors. We describe our experiments with different machine learning algorithms and vector lengths, algorithms including Multinomial Naive Bayes, SVM, XGBoost, transformers, and Distilbert. The best model performance was achieved by an ensemble of transformers including XLM-Roberta small and large and TwHIN-BERT base and large, combined using XGBoost. The ensemble was trained on the original tweets dataset plus additional training data from the 2021 shared task.

Keywords

SEXISM DETECTION, TRANSFORMERS, DEEP LEARNING, PRE-PROCESSING

1. Introduction

The past two decades witnessed an unprecedented surge in the amount of online content produced by social network users. Unfortunately, the rapid growth and ubiquity of this content made them a fertile ground for darker human emotions, including sexism. The definition of sexism often varies, but it generally refers to discriminatory practices or beliefs on the basis of sex or gender. It can take on various forms, which may range from subtle and indirect to overt and hidden expressions. Most often, these forms of discrimination are expressed against women with the aim of humiliating or objectifying them, destroying their reputation, undervaluing their skills and opinions, or making them feel fearful and vulnerable [1, 2, 3, 4]. Hence, hatred, threats, harassment, intimidation, and disparagement may all be the results of such sexist content. Fox et al. [5] argue that sexist behavior is promoted due to the 'online dis-inhibition effect', i.e., online users who remain anonymous may exhibit behaviors that they would not typically display in face-to-face situations or when their identity is known. They also argue that engaging with sexist content online may lead to sexist attitudes offline. As a result, the

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

yhatekar@iu.edu (Y. A. Hatekar); mabdo@iu.edu (M. S. Abdo); snkhanna@iu.edu (S. Khanna); skuebler@indiana.edu (S. Kübler)

https://github.com/YashHatekar (Y. A. Hatekar); https://github.com/muhsabrys (M. S. Abdo); https://github.com/k3va (S. Khanna); https://cl.indiana.edu/~skuebler/ (S. Kübler)

0000-0003-0885-5436 (S. Kübler)

Table 1
Examples from the EXIST dataset.

Sexist	Call me sexist but it just feels wrong that women are reffing the NBA like go ref the WNBA.
	Esta gringa sigue llorando por el gamergate, que c̈oincidenciaq̈ue tenga pronombres en su perfil
Non-Sexist	Even if you get embarrassed and blush, you can still confront hard things. #KeepMoving
	Los políticos acostumbran a hablarle al pueblo como si fueran una manada de estúpidos pero lamanada no hacemos nada por contradecirlos.

automatic detection and classification of this content into distinct categories have become a critical task to promote gender equality and create a safe online environment for everyone.

Machine learning techniques have proven to be effective in detecting and classifying sexist content. By training machine learning models on large datasets labeled for sexism, algorithms can learn the patterns and features that characterize such content. Both binary classification (i.e., sexist and non-sexist) or a more fine-grained classification, such as implicit and direct sexism exist [6, 7, 8]. Nevertheless, detecting sexist content on online platforms is challenging, especially on Twitter. Tweets are typically short, making it difficult for the models to extract unique patterns and features which discriminate sexist from non-sexist content. Also, because Twitter users have to limit their tweets to a small number of words, they resort to using non-standard language, emojis, and abbreviations, among other ways, to send their messages in the shortest form. Additionally, sarcasm, irony, and vague language make it difficult for the models to perform well [9].

In this paper, we present our IUEXIST team submissions for the EXIST 2023 [10, 11] Shared Task 1. For this task, the system needs to decide whether a given tweet in English or Spanish is sexist or not. As per the task, sexist is chosen if the tweet i) is sexist, ii) describes a sexist situation, iii) criticizes a sexist behavior. Table 1 shows sample sexist and non-sexist tweets in both English and Spanish.

The remainder of the paper is organized as follows. In section 2, we present a review of related work on detecting sexist content on various online platforms. In section 3, we describe our methodology, including the dataset, data pre-processing, and the machine learning algorithms used in our experiments. We then present our experimental results and discuss the implications of our findings in section 4. Finally, in section 5, we conclude the paper by summarizing our contributions, discussing the limitations of our study, and outlining avenues for future research.

2. Related Work

The SemEval-2023 Shared Task 10 [12] aimed to improve the automatic detection of online sexism. Unlike previous studies that focused on the binary classification of sexist content, this task introduces a new hierarchical taxonomy of sexist content that contains granular vectors of sexism. The study uses a dataset of 20k social media comments and aims to create more accurate and explainable models for sexism detection. The taxonomy included four categories

(Threats, Derogation, Animosity, and Prejudiced discussion) and 11 subcategories (e.g., threats of harm, aggressive attacks, gender stereotypes, and supporting mistreatment of women). The data used in the study was compiled from both Reddit and Gab, and sexist content was later annotated by highly-trained female annotators. The shared task involved three main tasks: a) a binary classification (sexist vs non-sexist), b) a four-category classification, and c) an 11-fine-grained-vector classification. The leading system in Task A employed a multi-task DNN structure and performed additional pretraining of DeBERTa-v3 and TwHIN-BERT on the starter kit unlabelled data, as well as an extra dataset. In Task B, the top-performing system utilized an instruction-tuned Pathways Language Model (PaLM) with the model, with a prompt that was parameter-efficient and tuned specifically for the task data. The system used majority voting over six iterations. Lastly, for Task C, the best-performing system conducted further training of DeBERTa-v3 using the starter kit unlabelled data and incorporated a second loss term known as normalized temperature-scaled cross entropy.

Almanea and Poesio [13] created a corpus of Arabic misogynistic tweets, annotated by three annotators, and used it to train a model for classifying tweets for misogyny using AraBERT. They trained a binary classifier for each coder with soft loss functions and a majority vote hard training. The results showed that the model trained using CE soft loss had the highest accuracy (77.79%) and F1-score (77.38), but had a relatively higher cross-entropy (0.586) and JSD (0.244) compared to the other models. The overall agreement between the three annotators was low, and the results suggest that annotator subjectivity has a significant impact on the accuracy of machine learning models for classifying sexist language.

Parikh et ak. [8] developed a semi-supervised multi-task learning neural framework for the multi-label fine-grained sexism classification of accounts of sexism, using sentence representations from word embeddings and pre-trained models. The study used a dataset of 13 023 accounts of sexism that was tagged with 23 different categories of sexism, created by trained annotators who had formal experience with studying gender and/or sexuality. The study explored different baselines and approaches for classifying sexism. With regard to baselines, random labeling and traditional machine learning methods such as SVM, random forest, and logistic regression were explored. The features chosen in these methods include TF-IDF on character n-grams, word unigrams, and bigrams, ELMo embeddings, and a composite set of features. Additionally, they explored various deep learning architectures, including LSTM-based architectures such as biLSTM, biLSTM-Attention, and hierarchical-biLSTM-Attention. They also included sentence embeddings with biLSTM-attention, CNN-based architectures such as CNN-Kim and C-biLSTM, and CNN-biLSTM-Attention. Logistic regression with averaged ELMo embeddings as features was found to perform best among the traditional ML methods with an F1 score of 0.595, and a macro-F of 0.479. Among the deep learning baselines, biLSTM-Attention (F1: 0.728, macro-F: 0.650) and Hierarchical-biLSTM-Attention (F1: 0.725, macro-F: 0.650) were the best models.

3. Methodology

3.1. Data

For the development of the EXIST dataset, over 400 popular expressions and terms that are commonly used to undermine women's roles in society, in English and Spanish, were used

as search terms. Overall, the original training data consists of 3,660 Spanish tweets and 3,260 English tweets. We also used additional data from the EXIST task1 datasets in 2021 and 2022 [4, 14]. The final size of the training dataset is 8 960 tweets in Spanish and English out of which 5 593 were sexist and 3 367 non-Sexist.

Since the tweets were provided with six annotator votes, we use a majority voting scheme to label the tweets as either sexist or not-sexist. For tweets for which there was a tie in the annotations, we consider them sexist to partly address the class imbalance.

3.2. Data Pre-Processing

The data pre-processing step involved five steps. 1) Any URLs were replaced by 'URL'. 2) Retweet 'RT', which is not relevant to the task, was removed. 3) Usernames were replaced with the word 'USER'. 4) Emojis were converted to their corresponding text equivalents using the Python library 'emoji' (https://pypi.org/project/emoji/). 5) All non-alphanumeric characters, except apostrophes and spaces, were removed.

A first attempt to eliminate hashtags showed that they are helpful and should not be deleted. For instance, the tweet *"#Catcalling is #Harassment. It's Not a Compliment. It's never okay. #feminist #feminism #stopstreetharassment https://t.co/g5nJy12sIl'*", is labeled as sexist by the majority of annotators in the training dataset. Removing the hashtags results in the removal of all relevant content.

3.3. Classifiers

We used a range of classifiers: multinomial Naive Bayes and Support Vector Machines using the scikit-learn implementation [15], XGBoost [16], DistilBERT [17], RoBERTa [18], XLM-RoBERTa [19], and TwHIN [20]. We used HuggingFace to fine-tune these transformers for our Twitter dataset.

Since the transformers can only accept input of a predetermined maximum length, we experiment with different vector lengths and found the following lengths optimal: 95 for XML-RoBERTa base and 128 for the remaining transformers.

For the multinomial Naive Bayes and SVM, we used the default parameters. To hyper-parameterize XGBoost, we used GridSearchCV along with a five-fold cross-validation.The best hyperparameters were identified as a maximum depth of 128, a learning rate of 0.1, the number of estimators set to 200, a seed of 47, and the internal evaluation metric set to logloss. We used HuggingFace's Autotrain to fine-tune the transformer models.

3.4. Evaluation

The official score in task 1 is ICM (Information Contrast Measure) [21], thus we report our results using this metric. We also report macro-averaged F1 in the HARD-HARD setting.

After considering former studies and the gold labels provided by the guidelines, we decided to focus on the Hard-Hard evaluation.

Table 2
Official results of the IUEXIST submissions.

		HARD-HARD			SOFT-SOFT	
Language	Model	Rank	ICM	F1	Rank	ICM
All	IUEXIST_1	16	0.5313	0.7734	**9**	**0.7115**
	IUEXIST_2	**15**	**0.5341**	**0.7717**	17	0.6141
	baseline	0	0.9948	1	0	3.1182
English	IUEXIST_1	19	0.5225	0.7509	9	0.6802
	IUEXIST_2	24	0.5059	0.7419	19	0.3893
	baseline	0	0.9798	1	0	3.1141
Spanish	IUEXIST_1	16	0.5294	0.7907	14	0.7076
	IUEXIST_2	13	0.5460	0.7942	12	0.7479
	baseline	0	0.9999	1	0	3.1177

4. Results

4.1. Official Results

We submitted two systems for evaluation. IUEXIST_1 uses XLM-RoBERTa Large trained on the official training set provided by the shared task. IUEXIST_2 uses an ensemble of four transformers, XLM-RoBERTa base and large, and TwHIN base and large. We then train XGBoost on the output of the transformers. All the ensemble models are trained on the combination of the official training set and the additional data (see Section 3.1).

Table 2 provides a summary of the official result or our team's submission. These results show that IUEXIST_2 performs slightly better than IUEXIST_1 in the hard evaluation while IUEXIST_1 performs significantly better in the soft evaluation. The gains of IUEXIST_2 in the hard evaluation are due to gains in Spanish, where the ICM is about 0.02 higher than for IUEXIST_1 (0.5460 for IUEXIST_2 and 0.5294 for IUEXIST_1). These gains are offset by a smaller loss for English. The good performance of IUEXIST_1 in the soft evaluation is due to its performance in English (ICM: 0.7115 vs. 0.6141 for IUEXIST_2).

4.2. Results on the Development Set

In addition to the officially submitted systems, we performed a more extensive evaluation on the development set.

We trained and evaluated the 10 different classifiers and the ensemble described in Section 3.3, the individual classifiers trained on the original training set, and the ensemble on the extended training set.

The results of these experiments are shown in Table 3. Since ICM and the F-scores show the same trends, we will focus on ICM scores in this analysis. The best model performance was achieved by the ensemble without pre-processing with an ICM of 0.5873. The non-neural classifiers generally show lower performance than the transformers. Among the latter, the XLM-RoBERTa base shows the best performance, with an ICM of 0.5716.

When we look at the question of whether pre-processing is useful, we see that some classifiers,

Table 3
Evaluation of different models and pre-processing settings on the development set.

Pre-processing	Classifier	ICM (hard)	F1 (positive)	macro F1
original	multinomial Naive Bayes	0.1354	0.6785	0.6729
	Support Vector Machines	0.2738	0.7108	0.7180
	XGBoost	0.3220	0.7273	0.7332
	DistilBERT	0.3822	0.7427	0.7522
	ROBERTA base	0.4233	0.7479	0.7650
	XLM-RoBERTa base	0.5716	0.8025	0.8120
	XLM-RoBERTa large	0.5547	0.7965	0.8067
	TwHIN base	0.5130	0.7856	0.7934
	TwHIN large	0.5377	0.7884	0.8014
	ensemble	**0.5873**	**0.8054**	**0.8171**
pre-processing	multinomial Naive Bayes	0.1446	0.6811	0.6761
	Support Vector Machines	0.2377	0.7009	0.7067
	XGBoost	0.2638	0.7079	0.7149
	DistilBERT	0.3757	0.7416	0.7502
	RoBERTa base	0.4833	0.7749	0.7841
	XLM-RoBERTa base	0.5167	0.7856	0.7947
	XLM-RoBERTa large	0.5487	0.7996	0.8045
	TwHIN base	0.5206	0.7876	0.7958
	TwHIN large	0.5100	0.7841	0.7925

Table 4
Comparing ensemble variations.

Classifier	Train	ICM (hard)	F1 (positive)	macro F1
XLM-ROBERTA large	2023	0.5547	0.7965	0.8067
ensemble	2023	0.5634	0.7980	0.8095
ensemble	extended	0.5873	0.8054	0.8171

such as the multinomial Naive Bayes and RoBERTa base profit from pre-processing, but for most models, pre-processing is detrimental.

Since our official results leave us with the question of whether the gains of the IUEXIST_2 model over IUEXIST_1 are due to the ensemble approach or to the extended training set, we perform a comparison including an experiment where we use the ensemble with only the current training set. The results are shown in Table 4. They show that the gains are mostly due to the additional training data, the ICM for the ensemble with this year's training data only reaches 0.5634, in comparison to 0.5534 for the XLM-RoBERTa model. Adding the 2021 training data adds a larger gain to 0.5873.

5. Conclusion

We have presented our submissions to the EXIST 2023 shared task 1. We found that an ensemble of four different transformers with XGBoost for voting provides the best results in the HARD-HARD evaluation (rank 15). However, for the SOFT-SOFT evaluation, we found XML-RoBERTa to reach significantly higher results. This system was ranked 9th out of 70 submissions.

As described above, this system was developed as a project in a course on machine learning. For most of us, this was our first time collaborating on a shared NLP task, drawing on experiences and discussions with individuals from various academic backgrounds and cultural perspectives. One of the key challenges we faced was being able to confine ourselves to the definition of Sexism, as multiple examples in the data set seemed to be open to interpretation depending on the context.

Yet another challenge was knowing how to sift through the extensive amount of technical information and resources available to us, and to direct our attention to problem-solving through continuous learning.

For the future, we plan on investigating the effect of using additional training data for the different systems since the results showed that adding more training data was more successful than going from a single transformer to the ensemble. On the one hand, adding training data can help combat data sparsity, but it also adds the risk of distorting the class distribution. We are also interested in a long-term evaluation to see to what degree the temporal distance between the training and test has a negative effect on performance.

References

[1] H. Abburi, P. Parikh, V. Chhaya, Niyati ad Varma, Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach, Data Science and Engineering 6 (2021) 359–379.

[2] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, An annotated corpus for sexism detection in French tweets, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 2020, pp. 1397–1403. URL: https://aclanthology.org/2020.lrec-1.175.

[3] J. K. Swim, R. Mallett, C. Stangor, Understanding subtle sexism: Detection and use of sexist language, Sex Roles 51 (2004) 117–128.

[4] F. Rodríguez-Sánchez, J. C. 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.

[5] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, Computers in Human Behavior 52 (2015) 436–442.

[6] A. Jha, R. Mamidi, When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data, in: Proceedings of the Second Workshop on NLP and Computational Social Science, Vancouver, Canada, 2017, pp. 7–16.

[7] S. Sharifirad, A. Jacovi, Learning and understanding different categories of sexism using

convolutional neural network's filters, in: Proceedings of the 2019 Workshop on Widening NLP, Florence, Italy, 2019, pp. 21—23.
[8] P. Parikh, H. Abburi, P. Badjatiya, R. Krishnan, N. Chhaya, M. Gupta, V. Varma, Multi-label categorization of accounts of sexism using a neural framework, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 1642—165.
[9] S. Butt, N. Ashraf, G. Sidorov, A. Gelbukh, Sexism identification using BERT and data augmentation – EXIST2021, in: IberLEF@ SEPLN, 2021.
[10] 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.
[11] 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.
[12] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, Technical Report arXiv:2303.04222, arXiv, 2023.
[13] D. Almanea, M. Poesio, ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 2022, pp. 2282–2291. URL: https://aclanthology.org/2022.lrec-1.244.
[14] 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.
[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[16] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD), New York, NY, 2016, pp. 785–794.
[17] V. Sanh, L. Debut, J. C. andThomas Wolf, DistilBERT, a distilled version of BERT: Smaller, Faster, Cheaper and Lighter, Technical Report abs/1910.01108, ArXiv, 2019.
[18] 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, Technical Report abs/1907.11692, arXiv, 2019.
[19] 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, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[20] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, , A. El-Kishky, TwHIN-BERT: A Socially-Enriched Pre-trained Language Model for Multilingual Tweet Representations, Technical Report arXiv:2209.07562, arXiv, 2022.

[21] 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.