# *DeepCapture*: *Sexism Detection on Social Media Platforms with Transformer Based Deep Learning Methods*

Atharv Abhijeet Bagde[1] and, Riley Halloran[2]

Indiana University Bloomington, Bloomington IN 47402, USA
{athbagde,rifhall}@iu.edu

**Abstract.** This paper discusses the methodology and results for the sEXism Identification in Social neTwork (EXIST) 2023 challenge. This shared task contains three tasks: sexism identification, source intention, and sexism categorization. In this paper we tackle the 1st challenge and use the mixed Spanish/English sexism identification dataset. Our approach makes use of a transformer-based models to detect sexism. The models used were taken from the Hugging Face library. In this paper we try to examine the effect of gender of annotators, and the tweets language on the final model performance. To that end, we created partitions in the data for both languages and for the male and female annotators. We used these data partitions to train various models and see how they performed on the EXIST shared task.

**Keywords:** Sexism Detection · Transformers · Linguistics. Machine Learning. Deep Learning · Natural Language Processing.

## 1 Introduction

Sexism on social media refers to comments that propagate negative stereotypes about women or gender-based biases. These remarks can take many forms, from subtle discrimination to overt rhetoric meant to demean, objectify, or intimidate women. To create a safe and welcoming environment for all users, it's important to identify and analyze sexism and its underlying intentions on social media. This helps to prevent the spread of harmful attitudes, minimize the psychological impact on people, and gain insights into patterns and trends in gender-based discrimination. Recognizing sexism on social media can also help us understand social prejudices and attitudes better. By examining the frequency and types of sexist comments, we can uncover societal problems such as discrimination, gender inequity, and gender-based violence. The resulting information can inform policy choices, education and awareness-raising campaigns, and other initiatives aimed at advancing gender equality and social justice. With the motivation of facilitating better sexism detection on social media and other platforms, we take part in the EXIST shared task 1

The EXIST shared tasks has 3 main tasks: 1. Sexism Identification, 2. Source Intention and 3. Sexism Categorization. Task1 of Sexism Identification is a binary classifier task, meaning that our model will take in the provided data, and return a labeling for each tweet. Our submission will enter the database for EXIST and be compared with others around the globe. With the amount of content generated each day the need for automated detection algorithms has only risen. To that end EXIST and those participating are pushing the boundaries of what is possible in the space

Our approach to this task was to use a BERT based transformer models. The models were pretrained with a masked language modeling (MLM) objective on a large corpus of English data. When trained this way, the model has 15% of its words masked over so the model can not take them in as an input. From there the data that is used gets better represented and the model learns a closer representation of the English language. This meant that we could set up the model very efficiently and have a working project up and running quickly. A large amount of emphasis was placed on tuning our model using the dev split of the data. We split the data based on annotator age and gender to examine how potential biases in the annotators can influence the efficacy of the model

## 2 Literature Review

In this literature review, we will explore the most significant works that we have referred to to finalize our approach for completing the EXIST 2023 Task 1.The EXIST 2021 overview paper by Sanchez et. al [1] discusses the different classification approaches employed for Task 1 (sexism detection) — transformer-based models, other deep learning methods like LSTMs and traditional ML approaches of SVM, Logistic Regression, Random Forest. Transformer based model architecture is preferred by the majority of participating teams and all of the top-10 performing teams. The trend continues in the EXIST 2022 overview paper by Sanchez et. al [2] where with the exception of one team all other teams employ transformer-based models for both Task 1 and Task2 of sexism detection and classification. The traditional ML methods even with the extraction of additional features didn't produce comparable results to the transformer based models. We can see performance comparisons of different transformer models in the paper by Álvarez et al. [3], in which we see that the RoBERTa model is performing well on English datasets but lacks on the Spanish datasets. This is evident from [1] and [2] as well where for English datasets either RoBERTa or BERT based model are used, but for Spanish dataset only BERT is used by the top 10 teams across the tasks.

For exploring data preparation techniques we refer to Schutz et al. [4] where they compare the model performance on sexism detection between external data pre-training and data augmentation. The results showed that performance is enhanced by pre-training on external data but it deteriorates by the use of data augmentation.

The decrease in model performance by using data augmentation is also shown in Bedmar et al. [5] where they compare model performance of BERT, RoBERTa, XLNet, DistilBERT with and without augmented data. For all the models the non-augmented data gives better results.
The performance enhancement by pre-training on external data is also stated in [2] where the authors claim that transformer-based models benefit from training with data from the same source. One more unique data processing approach was explored in Paula et al. [6], where they first translate the foreign-language data into English and then train a single language model for sexism detection and classification. This framework of single language model yielded better results than ensemble transformer models approach
From the above review we considered to start with non-augmented data input using RoBERTa for English and BERT for Spanish datasets. We further train different BERT variants like DistilBERT, RoBERTuito, BETO and XLM-RoBERTa on different sets of dev dataset to try and identify any trends with respect to the gender and age of the annotators.

## 3 Data

With rising use of social media platforms in the past decade, there have been multiple challenges with respect to the content being posted of them. Usage of offensive language and hate speech has become quite prevalent throughout all the platforms. In particular we concentrate on the presence of sexism which can be observed in various forms like stereotyping, ideological biases, and sexual violence. In an attempt to detect, analyze and model these sexist content, EXIST dataset has been created which is a collation of sexist tweets in both English and Spanish language. It includes over 400 popular expressions and terms that are commonly used to undermine women's roles in society. The dataset is balanced with almost equal number of Spanish (3660) and English (3260) tweets present in it.
Specific care is taken while building this dataset to avoid different biases. They do so by annotating the tweet by a total of six annotators -- 3 male and 3 female. Each gender of the annotators is also equally distributed across the age groups with one from each of the following : 18-22, 23-45 and 46+. Additionally in an attempt to avoid temporal bias the tweets were selected from different time periods. The entire dataset is divided into 3 sets. There is a training set of 6920 tweets for training the models.
For evaluating a model, validation or dev set is given. There is also a test set provided to share the final predictions on. Only training and dev sets have labels provided, while test set just has the unannotated tweets present. The dataset also comes with the labels for Task 2 and Task 3 of the shared tasks. Only the positive labels in Task 1 have corresponding Task 2 and Task 3 labels which correspond to source intention and sexism categorization. For the scope of this paper, we just focus on the Shared Task 1 data and labels.

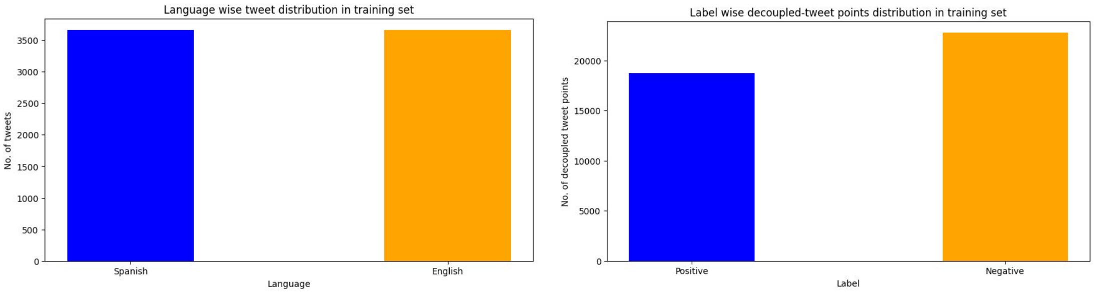


***Fig1**. Bar graphs depicting the distribution of tweets in training set based on : a) language and b) classification label*

### 3.1 Data Pre-Processing

From [5] we know that data augmentation and other preprocessing steps on data doesn't help in improving the transformer based model performance. In fact it deteriorates the model score significantly in some cases as well. This can be attributed to the fact that the pre-trained BERT based models take unfiltered and unprocessed text as inputs. No standard processing steps like punctuation-removal, emoji-removal, small-casing, stemming or lemmatization are applied to it. The BERT based tokenizer thus expects the fine-tuning data in similar format as well. However we do divide the dataset into multiple sub-parts.

Firstly, to enable the training of English and Spanish tweets on different models we create two sets based on the language. Further we also aim to analyze any potential pattern in sexism annotation based on the gender of the annotator as well. Accordingly we also divide the training sets to get male-only and female-only annotated tweets. Luckily since there are equal number of male and female annotators are present, there is no need to worry about disproportionate amount of data being assigned to a specific gender. In order to create separate male and female annotators dataset, we need to decouple a single tweet into 6 -- one each for each annotator. Distribution of both --language and label-- of data in the training set can be viewed in the Fig1. With almost equal distribution in both cases, the need for data augmentation is further eliminated.

## 4 Methodology

Once the data-splits and the data-subsets are generated, we select the model architectures to train our data on. The main focus of the paper would be to test out different variants of BERT based transformation model. We start by training a simple BERT base-model using the full dataset and then compare the performance of other models specifiably trained on the other data subsets. Lets take a glance at different training setups.

### 4.1 Base model

We choose a RoBERTa based architecture for training our base model which is available as a pre-trained model from HuggingFace. In order to fine-tune the model on our data and use-case we add fully connected layers using Huggingface's auto-classification model. We import the corresponding RoBERTa-base tokenizer for the model from Hugging face as well. The training parameters are as follows:

1. Optimizer: AdamW (ep =1e-8)
2. Learning rate: 2 e-5
3. Epochs: 5
4. Batch size: 16
5. Weight Decay: 0.01

### 4.2 Other models

The training parameters for the other models remain the same, only the architecture used and tokenizer changes as per the data subset used for the training. First we start by training a separate model each for the 2 languages. For the English language model we use the RoBERTa-large variant. For training on the Spanish data we make use of RoBERTuito variant which is specifically pre-trained on Spanish language data. From [3] we get to know that RoBERTa based models tend to not perform well on Spanish datasets. Thus for the Spanish subset we train two models : BERT based BETO and RoBERTa based RoBERTuito in order to compare their performance.

We also have created gender based subsets of the training sets. Since the gender based subset have mixed-language tweets, we use XLM-RoBERTa as the base pre-trained architecture. XLM-RoBERTa is a multilingual model which is trained on a corpus of 100 languages data. We fine-tune it for English and Spanish datasets by attaching the Auto Classification model as mentioned in the above process.

### 4.3 Evaluation Strategy

For evaluating the trained models we make use of the eval dataset provided to us. The metrics used for evaluation would be F1 (Positive class), Macro F1 and the custom metrics provided by the EXIST: ICM.

F1 score is defined as the harmonic mean of precision and recall

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

The macro-averaged F1 score (or macro F1 score) is computed using the arithmetic mean of all the per-class F1 scores.

$$\text{Macro F1 Score} = \frac{\sum_{i=1}^{n} \text{F1 Score}_i}{n}$$

ICM metric provided by EXIST has been defined by them as :

$$\mathbf{ICM(A, B) = PMI(A, B) - log\ P(B|A)}$$

where IC(A) is the Information Content of the item represented by the set of features A, etc.

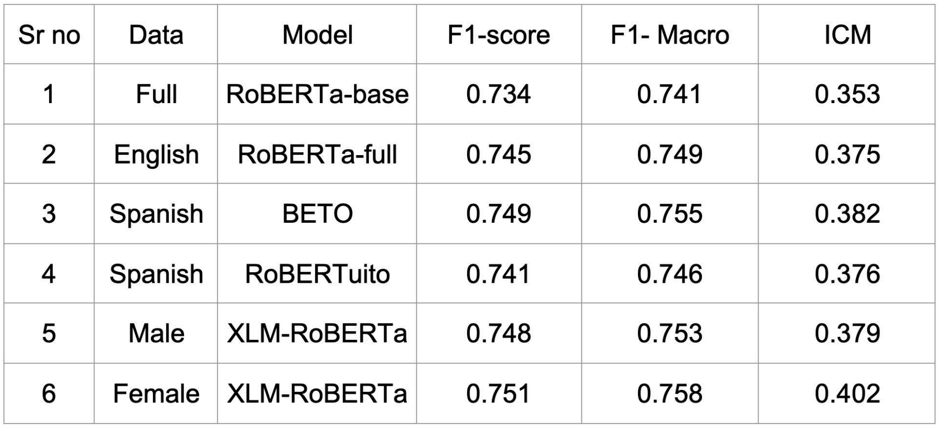
In the training phase we trained individual models based on the different data subsets created as mentioned above. This gives us multiple models to make predictions. On the test set, we need to decide a strategy that helps us choose the correct label from the multiple predictions we get from all the models. For this purpose we came up with the idea of using soft scores. Model prediction with the highest soft score or confidence score will be selected as the final prediction.

## 5 Results

The results on the evaluation set are as mentioned in the table below -- Table1. The base RoBERTa model has the lowest score among all the models. The models trained ands evaluated individually only on single language seem to gave a better performance than the single RoBERTa base model used for predictions of data belonging to both English and Spanish. This result is as expected since a single RoBERTa [8] model which is primarily trained on English data corpus is bound to perform worse on Spanish data than the model specifically trained on the Spanish data. Among the two Spanish models : RoBERTuito [10] and BETO [9], BETO gives us better performance across all the three metrics. This is in line with the case observed phenomenon in [3],[1] and [2] where the BERT model gave better results for Spanish data in place of RoBERTa model.

Coming to the results of male-only and female-only trained models, we observe that they perform significantly better than the RoBERTa baseline model. It shows that depending on the gender of the annotator the classification of a content, as sexist or not, varies. Also since the tweets in the gender divided data subsets contain both Spanish and English tweets, XLM-RoBERTa [7] is used as the base pre-trained model to fine tune further.

The ICM, F1 and F1 macro scores of all the models are generally in positive co-relation i.e. one score increases as the other increases and decreases as the other decreases too. However at certain instances ICM score does waver of from the trend a bit and increases despite the decrease in F1 score. This can be seen while comparing the RoBERTa models for English and Spanish language.



| Sr no | Data | Model | F1-score | F1- Macro | ICM |
|---|---|---|---|---|---|
| 1 | Full | RoBERTa-base | 0.734 | 0.741 | 0.353 |
| 2 | English | RoBERTa-full | 0.745 | 0.749 | 0.375 |
| 3 | Spanish | BETO | 0.749 | 0.755 | 0.382 |
| 4 | Spanish | RoBERTuito | 0.741 | 0.746 | 0.376 |
| 5 | Male | XLM-RoBERTa | 0.748 | 0.753 | 0.379 |
| 6 | Female | XLM-RoBERTa | 0.751 | 0.758 | 0.402 |

*Table1. This table depicts the final evaluation score of models trained on the different data subsets.*

## 6 Discussion and Conclusion

The aim of this paper was to provide an approach to enhance the sexism detection abilities of Transformer based deep learning model. To achieve this we proposed training of individual models in data subsets based on language and annotator's gender. While individual models for language have already been proposed, training individual models based on annotator's gender is a novel concept that hasn't been explored.

From the above results its apparent that training individual models based on annotator's gender gives a better performance than the baseline model. This means that some features of the data get lost while generalizing for both the genders. Although tested on a limited set, this experiment showcased that existence of sexism in a content can be perceived differently based on the gender of the content consumer. This finding can help tailor content and give appropriate warnings while consumption on a more granular level.

With respect to the future scope of the experiment, we can try and train the gender based models on more external data to bolster the findings and also create a new BERT or RoBERTa variant trained on gender specific content. One other factor that can affect the perception of sexism is age. We can also split the data on the basis of annotator's age and train model on each age group. This models performance can be similarly compared to baseline models to see if some features are lost again in generalization and if model performance can be improved .

## References

1. Rodríguez-Sánchez, Francisco, Jorge Carrillo-de-Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet and Trinidad Donoso. “Overview of EXIST 2021: sEXism Identification in Social neTworks.” Proces. del Leng. Natural (2021).

2. Rodríguez-Sánchez, Francisco & Carrillo-de-Albornoz, Jorge & Plaza, Laura & Mendieta-Aragón, Adrián & Marco-Remón, Guillermo & Makeienko, Maryna & Plaza, María & Gonzalo, Julio & Spina, Damiano & Rosso, Paolo. (2022). Overview of EXIST 2022: sEXism Identification in Social neTworks. Procesamiento de Lenguaje Natural. 69. 229-240.

3. Álvarez, Victoria Pachón, Jacinto Mata Vázquez, Wissam Chibane, Juan Luis Domínguez Olmedo4.“Automatic Sexism Identification Using an Ensemble of Pretrained Transformers” Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)

4. Schütz, Mina & Böck, Jaqueline & Liakhovets, Daria & Slijepcevic, Djordje & Armin, Kirchknopf & Hecht, Manuel & Bogensperger, Johannes & Schlarb, Sven & Schindler, Alexander & Zeppelzauer, Matthias. (2021). Automatic Sexism Detection with Multilingual Transformer Models.

5. Segura-Bedmar, Isabel. (2023). HULAT at SemEval-2023 Task 10: Data augmentation for pre-trained transformers applied to the detection of sexism in social media. 10.48550/arXiv.2302.12840.

6. Alejandro Vaca-Serrano .”Detecting and Classifying Sexism by Ensembling Transformers Models” Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)

7. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm’an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. “Unsupervised Cross-lingual Representation Learning at Scale." CoRR, vol. abs/1911.02116, 2019. arXiv:1911.02116. URL: http://arxiv.org/abs/1911.02116.

8. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692, 2019. URL: http://arxiv.org/abs/1907.11692.

9. https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

10. Pérez, Juan & Furman, Damián & Alemany, Laura & Luque, Franco. (2021). RoBERTuito: a pre-trained language model for social media text in Spanish.