

# Tlatlamiztli: Fine-Tuned RoBERTuito for Sexism Detection

Hardik Asnani<sup>1</sup>, Andrew Davis<sup>1</sup>, Aaryana Rajanala<sup>1</sup> and Sandra Kübler<sup>1</sup>

<sup>1</sup> Indiana University, Bloomington, IN, USA

## Abstract

We present an approach for detecting sexism in tweets containing English, Spanish, and a mix of both languages. We participated in task 1 of the EXIST 2023 shared task. Our research question focuses on evaluating the performance of a transformer-based model fine-tuned on a Spanish tweet dataset generated by translating tweets containing English, Spanish, and mixed English-Spanish text into Spanish. We use a RoBERTuito model fine-tuned on the EXIST2021 dataset. Our results show that our model placed 31st from 70 evaluated models in the HARD-HARD evaluation, but placed 10th in the SOFT-SOFT evaluation. When evaluating only the Spanish data, we ranked 12th for HARD-HARD and 4th for SOFT-SOFT, showing that our approach works well for the Spanish data.

## Keywords

Sexism Detection, machine translation, RoBERTuito

## 1. Introduction

Hate speech detection has emerged as a critical area of focus within sentiment analysis, particularly in light of the widespread use of social media. While occurring in various forms, sexism is a specific type of hate speech that occurs rampantly on social media platforms and in a multitude of languages “in many forms in social networks, includes a wide range of behaviours (such as stereotyping, ideological issues, sexual violence, etc.)” [1]. The EXIST shared task, which started in 2021, is the first of its kind to specifically address sexism detection in tweets written in both English and Spanish.

EXIST2023 [2, 3] has three sexism detection tasks, we focused on Task 1. This task is defined as a binary classification to determine whether a given tweet includes expressions or behaviors that are sexist. This includes instances where the tweet itself is sexist, it describes a situation that involves sexism, or it criticizes sexist behavior.

The dataset consisted of tweets written in English, tweets written in Spanish, and tweets that contained a mixture of both languages. As a result, it was important for us to address the tweets that had a combination of English and Spanish content. We aimed to develop a method that could effectively handle this mixed language scenario. We investigate whether translating


---


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

✉ hasnani@iu.edu (H. Asnani); ad7@indiana.edu (A. Davis); aarajana@iu.edu (A. Rajanala); 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)

all tweets to Spanish results in a more robust model. Our approach was developed during a course on machine learning in NLP.

The remainder of this paper is structured as follows: Section 2 provides an overview of the prior research on this topic, while Section 3 describes our system architecture. In Section 4, we describe the experiments we conducted to evaluate our approach, and in Section 5, we analyze our results. Finally, we conclude our work in Section 6.

## 2. Related Work

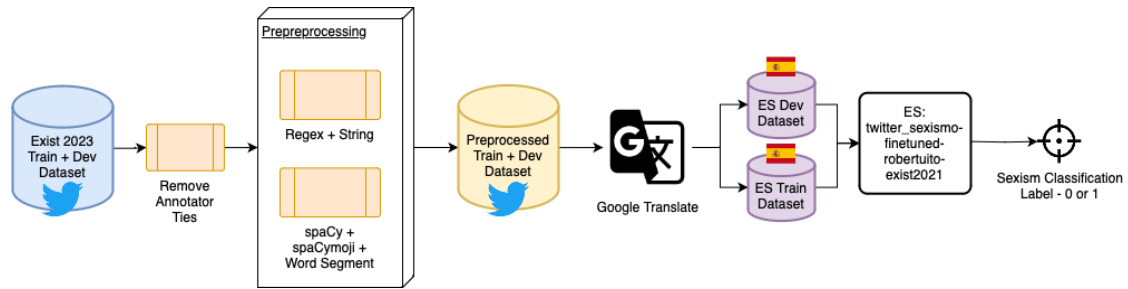
The EXIST (sEXism Identification in Social neTworks) at IberLEF 2021 was the first shared task in this series. It encouraged the progress and development of automatic identification of sexism in a broad sense. We focus on publications from the 2021 and 2022 EXIST Task 1, a binary classification task. The curated dataset for the task in 2021 consisted of 6 977 tweets for training and 3 386 tweets for testing. In 2022, the training set was based on the EXIST 2021 dataset. The EXIST 2022 dataset contained the same number of tweets for training and testing as the 2021 set. Additionally, the dataset contained 982 gabs, 492 in English and 490 in Spanish. The Gab information was labeled following the same process as the tweets in the EXIST 2022 dataset [1].

We examined the pre-processing techniques of teams from previous years and the results they obtained. López-López et al. [4] filtered punctuation and stop words and performed tokenization and lemmatization. They ranked 32nd with an F1 score of 0.7237 for Spanish and English using a majority vote between SDG, XGBoost, RoBERTa, and BERT.

De Paula et al. [5] conducted a study on sexism detection and explored the efficacy of utilizing Google Translate to convert data into both English and Spanish, thereby increasing the availability of training data. The authors categorized their models into three types. The first model was a multilingual classifier applied to the dataset without any translation. The second and third models were monolingual classifiers (English and Spanish) applied to the dataset without any translation and the dataset translated to the matching language of the classifier, respectively.

Their findings revealed that the multilingual model exhibited inferior performance compared to the two types of monolingual models. Additionally, the models incorporating translation outperformed the models without translation. Notably, for the Spanish language, the translated model achieved the same F1-score as the monolingual model and a slightly higher accuracy for Task 1.

Plaza-del Arco et al. [6], ranking fourth with a macro-averaged F1-score of 0.7841 in 2022, unpacked hashtags and split them into their constituent words and converted emojis to their alias. Both Plaza-del Arco et al. [6] and Talavera et al. [7] used BETO for Spanish [7]. BETO, short for Bidirectional Encoder Representations from Transformers for Spanish, is a pre-trained model that captures the contextual information and semantic representations of Spanish text, enabling more accurate natural language processing tasks [8]. Plaza-del Arco et al. [6] used BERT whereas Talavera et al. [7] used RoBERTa for English. Vaca-Serrano [9], the best performing approach from the 2022, used different combinations of BERT ensembles, including RoBERTa and BERTweet, and found that domain specific models produced the best results. The most



**Figure 1:** System Architecture of our Model. From left to right, the EXIST2023 dataset has annotator ties removed, is pre-processed, translated to Spanish, and classified via a fine-tuned transformer-based model.

"tweet": "Regálenme un llavero que diga “fuck the patriarchy” 🙄🙄🙄🙄🙄 "

**Figure 2:** Tweet from EXIST2023 Data that contains a mix of English and Spanish text as well as emojis.

successful teams also used ensembles and variants of BERT, particularly those trained on data from Twitter [1].

### 3. System Overview

In our system, we employ a transformer-based model that has been fine-tuned on a large dataset of Spanish text. To ensure the versatility of our model in processing English and Spanish data, we adopt a preprocessing step wherein we transform the entire dataset, comprising tweets in English, Spanish, and a mixture of both languages, into Spanish. This process involves leveraging the googletrans python library for machine translation, automatically converting English tweets and any mixed-language tweets into Spanish. By unifying the dataset in Spanish, we create a homogeneous corpus that consists of tweets in the same language.

The transformed dataset, now entirely in Spanish, is used to fine-tune the transformer model. This enables it to capture the linguistic patterns and nuances specific to the Spanish language, while still maintaining its ability to understand and process English text effectively. An overview of the complete pipeline is shown in Figure 1.

### 4. Methodology

#### 4.1. Data

The data is split into a training and development set with the former containing 6 920 tweets and the latter 1 038. The test set contains 2 076 tweets in English, Spanish, and a mix of both languages. Figure 2 shows an example from the data where the tweet contains both English and Spanish text as well as emojis.

The shared task focuses on learning under uncertainty. Thus, there are no gold labels per

se. Instead each tweet is annotated by six annotators, with an equal split of three female and three male annotators. As a result of having an even number of annotators, there were instances where certain tweets resulted in tied classifications. Specifically, there were 856 ties in the training set and 104 ties in the development set. For the HARD-HARD evaluation, the shared task considers the majority vote as the hard label, and tweets with ties are ignored. We followed this procedure and removed all the tweets with annotator ties from our training and development set.

## 4.2. Pre-Processing

Our pre-processing consists of deleting and filtering aspects of the text and then translating the text to prepare it for our experiments. First, we lowercased the text and then deleted URLs and user mentions. The next step was to convert emojis to their textual representation using spaCymoji (<https://spacy.io/universe/project/spacymoji>) as well as segment hashtags with wordsegment (<https://pypi.org/project/wordsegment/>) Finally, we deleted punctuation and extra white spaces between words. Non-Latin characters as well as spelling errors were left unchanged.

The entire pre-processed dataset was then converted to Spanish using the Google Translate API (<https://pypi.org/project/googletrans/>) to address the issue of tweets containing English, Spanish, and mixed English-Spanish textual data [10].

## 4.3. Model

As domain-specific models were shown to produce the best results in the previous EXIST Task, we opted to employ a version of RoBERTuito, namely the model developed by de Paula et al. [5] ([https://huggingface.co/hackathon-pln-es/twitter\\_sexismo-finetuned-robertuito-exist2021](https://huggingface.co/hackathon-pln-es/twitter_sexismo-finetuned-robertuito-exist2021)). The resource can be accessed through HuggingFace and is fine-tuned on the EXIST2021 dataset [11].

We then fine-tuned the model. For the official submission to the shared task, we trained on this year’s shared task training data and the development set. For the internal experiments, we trained on the training data and tested on the development set.

We tokenized all tweets using the tokenizer and pad or truncate the sequences to a maximum length of 128 tokens. We employed the ‘pysentimiento/robertuito-base-uncased’ tokenizer, which is a variant of the RoBERTa tokenizer designed for Spanish text processing. This tokenizer is specifically trained on a large corpus of Spanish text data and utilizes the uncased strategy, treating all text as lowercase to facilitate better generalization. By utilizing this tokenizer, we converted each tweet into a sequence of tokens, allowing us to process the text at the token level.

We set the hyperparameters for the training process, such as the number of epochs, the evaluation strategy, and learning rate after optimizing them through a series of 11 experiments. We found that our model was overfitting quickly and that it performed best with just one training epoch. We found that the default learning rate, 5E-5 performed best. With a single training epoch and a default learning rate of 5E-5, our macro average F1 score, precision, recall, and F2 score were all consistently high, with values of 86.71%, 86.74%, 86.69%, and 85.80%

**Table 1**

Official results of our system.

Language	HARD-HARD			SOFT-SOFT	
	Rank	ICM	F1	Rank	ICM
All	31	0.5013	0.7535	10	0.6879
English	40	0.4314	0.7090	22	0.3663
Spanish	12	0.5482	0.7867	4	0.9021

respectively, showcasing the robustness and overall effectiveness of our approach in addressing the task at hand.

We retrieve the hard label and soft labels (probabilities of each class) from the classifier. The hard label is a binary label (YES or NO) obtained by choosing the label with the highest probability, and the soft label distribution consists of the probabilities per class.

#### 4.4. Evaluation

The shared task results are evaluated in three different scenarios, we focus on HARD-HARD and SOFT-SOFT. For the HARD-HARD evaluation, the gold label consists of the class annotated by the majority of annotators. Any items with no majority class are removed from the evaluation. Here, the model produces a single label, which is compared to the gold label. In the SOFT-SOFT evaluation, the system provides probabilities for each class, and this distribution is evaluated against the distribution of the annotator decisions.

The official metric for the shared task is the Information Contrast Measure (ICM) [12]. For the official evaluation, we report our rank, ICM, and F1 for the HARD-HARD evaluation, and our rank and ICM for the SOFT-SOFT evaluation.

## 5. Results

### 5.1. Official Results

As described above, our system consists of a Spanish transformer that was originally fine-tuned on the 2021 training data, and that we further fine-tuned on this year’s training data. The official results are shown in Table 1. Since ICM and F1 show the same trends, we focus on ICM in our analysis. The results show that overall our system ranks 31st out of 70 submissions, with an ICM of 0.5013 for the HARD-HARD evaluation. For the SOFT-SOFT evaluation, however, our system ranks 10th overall (ICM: 0.6879) and 4th for Spanish (ICM: 0.9021). This shows on the one hand that our system may not be ideal for a hard evaluation, but it is suitable for modeling the distribution of opinions of the different annotators. On the other hand, our system reaches high results for Spanish but much lower results for English, showing that unsurprisingly, sexism is more difficult to detect in translated tweets.

**Table 2**

Results of our experiments on using different language models, evaluated on the development set.

Model	translated	ICM (HARD-HARD)	ICM (SOFT-SOFT)	macro F1
ours	yes	0.6007	0.9127	86.71
	no	0.5722	0.8776	85.63
RoBERTuito	yes	0.5776	0.9040	85.94
	no	0.5694	0.8652	85.96

## 5.2. Results on the Development Set

We also decided to have a closer look at how the language model influences the results and thus compared our model, which is based on RoBERTuito, with that original model. The results of these experiments are shown in Table 2. The results show that our model outperforms RoBERTuito by a small margin, showing that the first fine-tuning on the 2021 data was helpful, even though that dataset will have a different distribution to this year’s data.

## 6. Conclusion and Future Work

In our contribution to the shared task, we addressed the problem of detecting sexism in tweets that contain both English and Spanish text. We fine-tuned a RoBERTuito-based transformer model using a Spanish tweet dataset that was generated by translating tweets containing English, Spanish, and mixed English-Spanish text into Spanish. Our ranking, especially in the SOFT-SOFT setting, shows that our approach is effective in detecting sexism in tweets. Our experiments also demonstrate the importance of using existing data in a double fine-tuning setting, first on the 2021 training data, then on the current training set.

As described above, this system was developed as a project in a course on machine learning. Participating in an evaluation campaign of this magnitude offered us invaluable practical experience in real-world NLP tasks. It allowed us to apply our course knowledge and techniques to address the intricate problem of detecting sexism in social media.

The dataset’s inherent diversity posed a significant challenge as we grappled with tweets composed in both English and Spanish, as well as a combination of the two. This multifaceted language composition presented intricate complexities in our language processing endeavors. We dedicated substantial time to devising effective strategies for handling language complexities and maintaining high performance across the dataset, accommodating language variations and switching within tweets.

In conclusion, our involvement in the CLEF evaluation campaign as NLP course students has been an enriching experience. It deepened our understanding of the challenges associated with real-world NLP tasks, enhanced our technical skills, and offered a glimpse into the dynamic and evolving landscape of NLP research.

We are planning to further investigate the differences in our approach between the HARD-HARD and SOFT-SOFT evaluation. We are also planning to investigate an architecture where we translate each training tweet into the other language, and then classify each tweet in an ensemble of four classifiers, to see if we can improve the English results.

## References

- [1] 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 de Lenguaje Natural* 69 (2022) 229–240.
- [2] 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. Tsirikla, 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.
- [3] 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.
- [4] D. López-López, J. C. de Albornoz, L. Plaza, Combining transformer-based models with traditional machine learning approaches for sexism identification in social networks at EXIST 2021, in: *IberLEF@SEPLN*, 2021.
- [5] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, 2021. [arXiv:2111.04551](https://arxiv.org/abs/2111.04551).
- [6] F. M. Plaza-del Arco, M.-D. Molina-González, L. A. Ureña-López, M.-T. Martín-Valdivia, Exploring the use of different linguistic phenomena for sexism identification in social networks, in: *IberLEF@SEPLN*, 2022.
- [7] I. Talavera, D. C. Fidalgo, D. Vila-Suero, System description for EXIST shared task at IberLEF 2021: Automatic misogyny identification using pretrained transformers, in: *IberLEF@SEPLN*, 2021.
- [8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, *Pml4dc at iclr 2020* (2020) 1–10.
- [9] A. Vaca-Serrano, Detecting and classifying sexism by ensembling transformers models, in: *IberLEF@SEPLN*, 2022.
- [10] G. García Subies, EXIST2021: Detecting sexism with transformers and translation-augmented data, in: *IberLEF@SEPLN*, 2021.
- [11] M. Grandury, R. del Campo, *hackathon-pln-es/twitter\_sexismo-finetuned-robertuito-exist2021*, 2021. URL: [https://huggingface.co/hackathon-pln-es/twitter\\_sexismo-finetuned-robertuito-exist2021](https://huggingface.co/hackathon-pln-es/twitter_sexismo-finetuned-robertuito-exist2021).
- [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.