iimasGIL_NLP@EXIST2023: Unveiling Sexism on Twitter with Fine-tuned Transformers

Notebook for the EXIST Lab at CLEF 2023

Andrea Sanchez-Urbina[1,*], Helena Gómez-Adorno[2], Gemma Bel-Enguix[3,6], Vianey Rodríguez-Figueroa[4] and Angela Monge-Barrera[5]

[1]*Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autonoma de Mexico, Circuito Escolar 3000, Copilco Universidad, Coyoacán, 04510 Ciudad de México, CDMX*

[2]*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autonoma de Mexico, Circuito Escolar 3000, Copilco Universidad, Coyoacán, 04510 Ciudad de México, CDMX*

[3]*Instituto de Ingeniería, Universidad Nacional Autonoma de Mexico, Circuito Escolar S/N, Copilco Universidad, Coyoacán, 04510 Ciudad de México, CDMX*

[4]*Facultad de Ciencias, Universidad Nacional Autonoma de Mexico, Investigación Científica, C.U., Coyoacán, 04510 Ciudad de México, CDMX*

[5]*Facultad de Filosofía y Letras, Universidad Nacional Autonoma de Mexico, Circuito Interior, C.U., Coyoacán, 04510 Ciudad de México, CDMX*

[6]*Universitat de Barcelona. Gran Via de les Corts Catalanes, 585, 08007, Barcelona*

Abstract

This paper presents the iimasGIL_NLP approach for classifying tweets in the context of the sexism identification in social networks task (EXIST) at CLEF 2023. Identifying sexism in social media is a problem related to natural language processing. It can be approached as a binary classification problem from the machine learning perspective. Other subtasks presented in EXIST 2023 include categorizing the author's intention and identifying the category of the sexist message (in English and Spanish). These tasks can be approached as multiclass and multilabel classification problems. For the binary classification task, we evaluate many linguistic patterns combined with bag-of-words and n-gram features as input in classical machine learning algorithms.Additionally, we fine-tuned transformer models. We utilized embeddings from these fine-tuned transformer models in both the English and Spanish datasets for the categorization of intentions and categories of sexist messages. Our classification models in English obtained higher scores than those developed in Spanish. In the Hard-Soft evaluation of Task 3 (type of sexism) for English, we achieved the highest scores among all participating teams.

Keywords

Sexism detection, Twitter, Machine Learning, Transformer

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

*Corresponding author.

andreasanchezurbina@comunidad.unam.mx (A. Sanchez-Urbina); helena.gomez@iimas.unam.mx (H. Gómez-Adorno); gbele@iingen.unam.mx (G. Bel-Enguix); vianey_rodriguezf@ciencias.unam.mx (V. Rodríguez-Figueroa); angelamb580@gmail.com (A. Monge-Barrera)

0009-0006-2445-9280 (A. Sanchez-Urbina); 0000-0002-6966-9912 (H. Gómez-Adorno); 0000-0002-1411-5736 (G. Bel-Enguix); 0009-0005-6202-5208 (V. Rodríguez-Figueroa); 0009-0005-9530-8503 (A. Monge-Barrera)

1. Introduction

According to the Encyclopedia Britannica [1] sexism is the "prejudice or discrimination based on sex or gender, especially against women and girls". The term was coined in the sixties, modeled in the framework of the civil rights movement in comparison with racism. The problem has been identified by UNESCO [2, 3], and extensively studied by sociology and feminism, as it can be seen in classical reviews and compilations [4] and theoretical approaches [5].

Sexism is present in everyday life, in areas like work [6] and academy [7]. Language is often a vehicle to convey sexism [8, 9]. In the last decades, sexism has found a new context of development in social platforms like Twitter [10, 11] and Reddit [12].

The NLP community has treated the topic as a specific type of hate speech detection. SemEval 2019 and 2020 held the tasks OffensEval 2019 "Identifying and Categorizing Offensive Language in Social Media" and OffensEval 2020 "Multilingual Offensive Language Identification in Social Media". HatEval 2019 [13] was organized as part of SemEval and focused on hate speech against immigrants and/or women. It consisted of two tasks. The first task involved determining whether an English or Spanish tweet contained hate speech, and the second task was to decide if that speech was aggressive.

MeOffendEs@IBERLEF 2021 aimed to develop systems to detect offensive and discriminatory language in Spanish in online forums. It was organized within the IBERLEF workshop in 2021. The comments were divided into three categories: sexist comments, racist comments, and homophobic comments (Plaza-del-Arco et al., 2021) [14]

DETOXIS 2021 [15] was a shared task within the framework of the IberLEF 2021 workshop, aimed at developing systems to detect the toxic and offensive language in the context of the COVID-19 pandemic in Spanish. Participants had to develop systems that could classify the comments into toxic and non-toxic categories. Recipients of the offensive language were people attacked, threatened, insulted, offended, or denigrated "on the basis of race, ethnicity, nationality, political ideology, religion, gender, and sexual orientation" [15].

Another workshop that has tackled the same problem is EDOS (Explainable Detection of Online Sexism) [16]. The tweets in the task were annotated increasingly granularly, from the first binary classification to the final taxonomy that included eleven types of sexism.

EXIST was held in 2021 and 2022 at the Iberlef evaluation campaign, aiming to develop systems to detect sexism in both Spanish and English tweets [17, 18]. Participants had to create strategies to identify sexist posts from the dataset and classify them as "non-sexist" or "sexist". In 2023, the task is organized as a part of CLEF [19], combining English and Spanish tweets [20]. In this edition, the task incorporates the learning from disagreements paradigm, which makes it a more challenging problem.

Section 2 provides detailed information about the dataset provided by the EXIST 2023 organizers. Subsequently, in Section 3, we outline our methodology for addressing sexism detection in all tasks, covering aspects such as pre-processing, feature extraction, machine learning algorithms, and transformer algorithms. In Section 4, we present and discuss the results obtained from our in-house experiments and the EXISTS 2023 test. Finally,

in Section 5, we draw conclusions based on our findings and discuss the implications of the results.

2. Task description and data

The EXIST 2023 dataset is a collection of over 10,000 labeled tweets covering both English and Spanish languages. The dataset has been meticulously curated to include a balanced distribution of tweets in each language. Specifically, the training set comprises 6,920 tweets. The development set comprises 1,038 tweets, and the test set encompasses 2,076 tweets. This diverse dataset serves as a valuable resource for language-related research and analysis. To incorporate learning with disagreement, EXIST provides the labels proposed by each annotator, as well as some characteristics of the annotators, such as age (18-22 y.o./23-45 y.o./+46 y.o) and gender (MALE/FEMALE). Six different annotators annotate each tweet in the dataset.

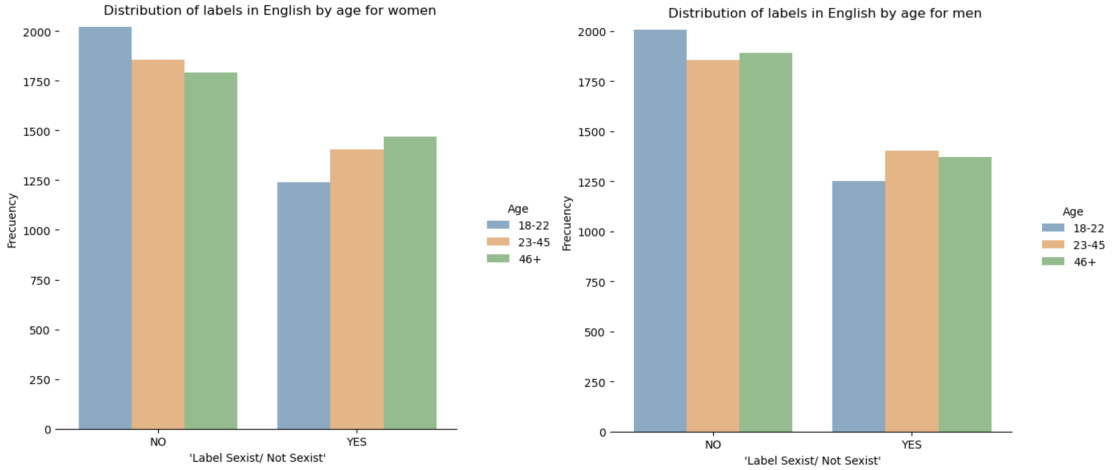
2.1. Task 1

The first task consists of a binary classification with labels "sexist" and "non-sexist". It is important to highlight that in the tweets with English labels (Figure 1a and 1b), the votes provided by the annotators for non-sexism (58%) are significantly higher than the votes for sexism (42%). There are also differences in the labels associated with sexism based on age and gender. Older women tend to propose the "sexism" label more frequently than younger women, while among men, those in the age group between 23 and 45 tend to use the "sexism" label more frequently. In the case of Spanish labels (Figure 1c and 1d), we observe a more balanced distribution of sexist and non-sexist labels for both genders. Furthermore, it is worth noting that the aforementioned patterns of behaviour related to age and gender persist.

To incorporate the disagreement paradigm and assign a single label to each tweet, our criteria for labeling a tweet as sexist was based on the majority agreement among the annotators. In cases of a tie, we further examined the gender of those who labeled the tweet as sexist. We classified the tweet as sexist if at least two women provided affirmative labels. This label was subsequently used for training the models.

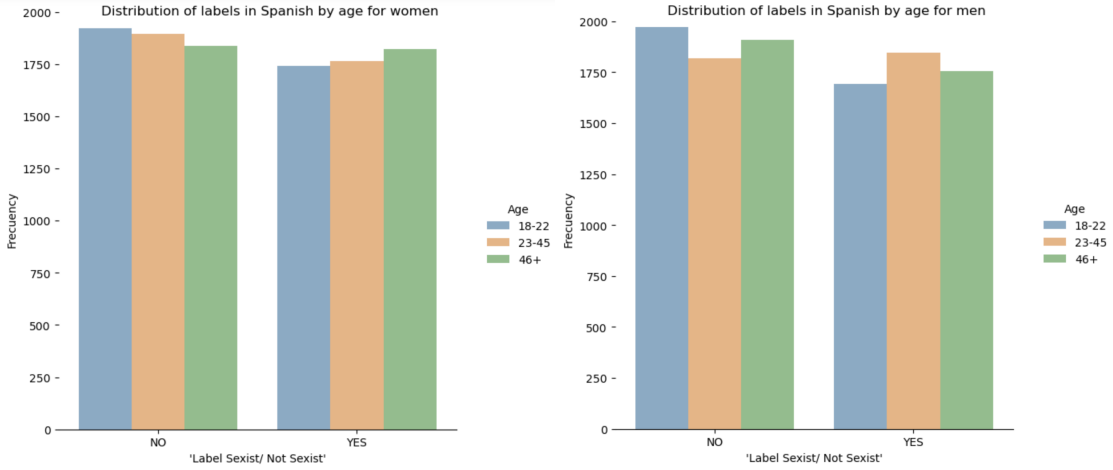
2.2. Task 2

For the tweets classified as sexist, the second task focuses on categorizing each tweet based on the author's intention. Each tweet can be assigned to three categories: Direct, Reported, and Judgemental. We can notice that the classes are imbalanced, particularly encountering a higher number of tweets labeled as direct sexism content. We also observe greater differences between the labels provided by women (Figure 2a and 2c) and those labeled by men (Figure 2b and 2d).



(a) English tweets labeled by women.

(b) English tweets labeled by men.



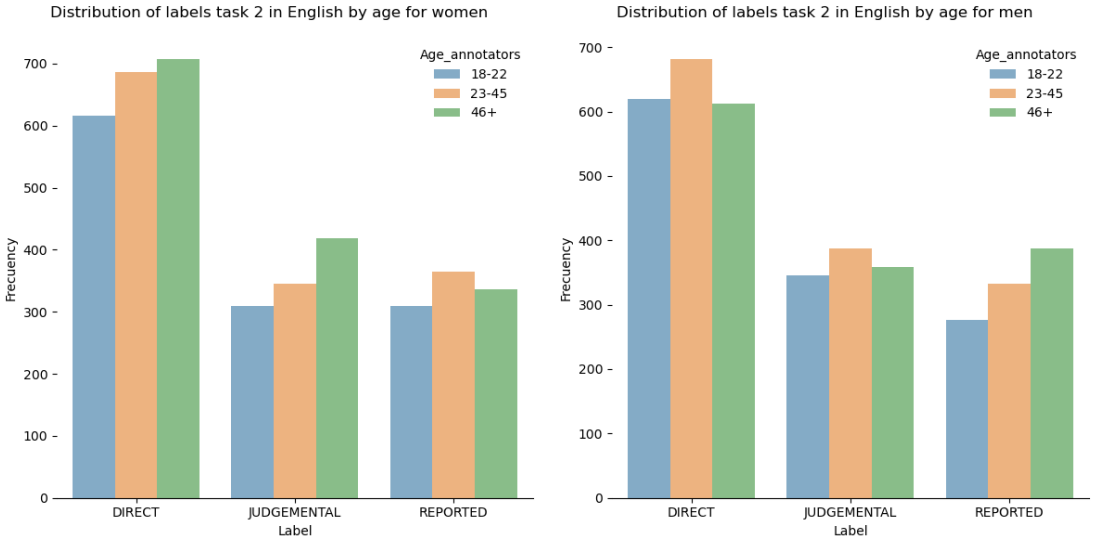
(c) Spanish tweets labeled by women.

(d) Spanish tweets labeled by men.

Figure 1: Distribution of labels for Sexism Identification (Task 1) based on language and gender

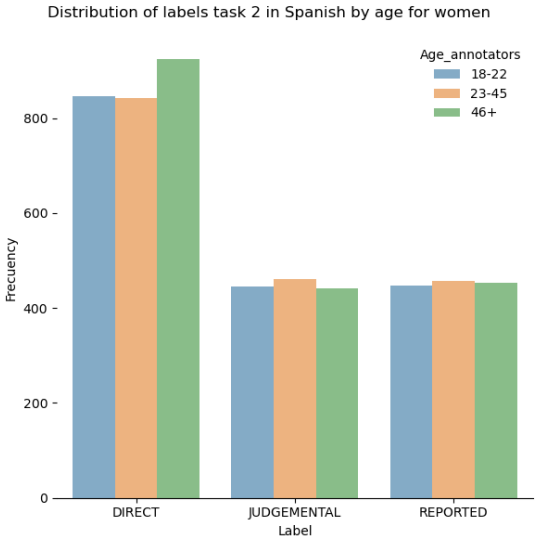
2.3. Task 3

The last task focuses on categorizing the tweets classified as sexist based on the type of sexism exhibited. EXIST 2023 proposes a five-class classification task, where multiple labels can be assigned to each tweet as a multi-label task. The following labels may be assigned to each tweet: Ideological and Inequality, Stereotypes and Dominance, Objectivation, Sexual Violence, Misogyny, and Non-Sexual Violence.

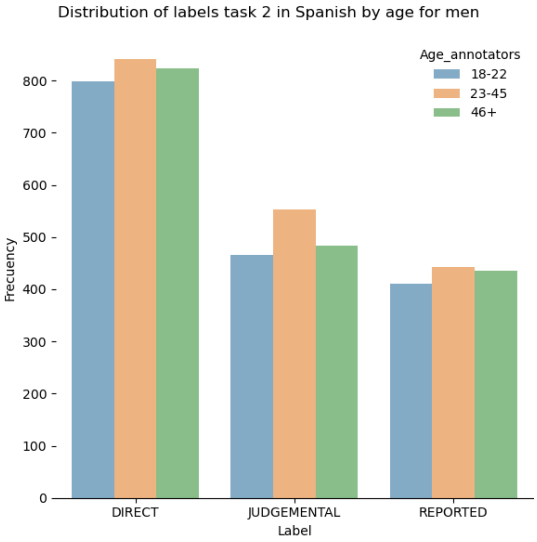


(a) English tweets labeled by women.

(b) English tweets labeled by men.



(c) Spanish tweets labeled by women.



(d) Spanish tweets labeled by men.

Figure 2: Distribution of labels for source intention (Task 2) based on language and gender

3. Approaches

3.1. Task 1 - Traditional Machine Learning Approach

Conventional natural language processing techniques focused on linguistic features, preprocessing, and traditional machine learning. Since the dataset is bilingual, it was divided and processed separately for each language to capture the essence of each one. The

variations considered in the text pre-processing and linguistic variants were as follows:

- Bag of words / Frequency
- Bag of words / Tfidf
- Remove stopwords
- Word n-grams
- Minimum frequency of the word in the corpus (min-df)
- Hashtag removal
- User removal
- Emoji removal
- Emoji count
- Count of emojis associated with sexist tweets
- Count of sexist expressions
- Addition of in-group index [21], [22]
- Addition of categorical index [23], [22]
- Default embeddings (word2vec, fasttext)

Python's scikit-learn library [24] was utilized to employ the base versions of machine learning classifiers. :

- Naive Bayes Classifier
- SVM (Support Vector Machine)
- Logistic Regression
- Random Forest

Each model was evaluated using stratified 10-fold cross-validation, and the reported metric is F1, as it has been used in previous EXIST editions. Combinations of features were made both in text pre-processing and feature extraction.

3.1.1. Sexist expressions

The linguistic analysis of sexist sentences consisted of a search for syntactic structures and semantic areas. For the English data, we obtained the frequency of words appearing in sexist and non-sexist sentences. Then, we selected those words that appeared more frequently and all the sexist sentences with that word in their structure. The words we analyzed are: like, women, woman, men, man.

The sexist sentences with the word "like" were generally accompanied by words: women, girl, feminist or female. In the non-sexist sentences, these words did not appear with the word "like". However, "like" did not appear as its verb form "to like". In most cases "like" appeared as a conjunction and introduced adverbial circumstantial sentences of manner. In sexist and non-sexist sentences, the frequency of the word "like" as a conjunction was higher.

The words "men" and "man" were the words with the highest predictive power for detecting sexist sentences. In the non-sexist sentences, the "men" word appeared 39 times, while in the sexist sentences, "men" appeared more than 100 times. These words

could appear alone, but in sexist sentences, they were usually accompanied by the word "women".

Spanish data were analyzed using the gender inflection of nouns. With this data we focused on insults such as *zorra*, *perra*, *estúpida*, *puta*. However, some of the words like *puta* qualified another noun in sentences like: *ni puta idea* or *hijo de puta*. These sentences were not marked as sexist even though they contained a sexist word.

Subsequently, we looked for the positions in the sentence where the word *mujer* appeared. This led us to find two frequent structures in sexist sentences: *mujer florero* and *se nota que es mujer*. We also found the insult of *lagartona* and *rubia tonta*. These sentences showed us that sexist expressions in Spanish operate from metaphors and double meanings.

3.2. Task 1 - Transformer approach

Following the success of the first BERT [25] model, multiple versions were subsequently developed, including Bertweet [26], which was the first publicly available large-scale pre-trained language model specifically designed for English Tweets. Bertweet was built upon the architecture of BERTbase and trained as RoBERTa [27]. The DistilBERT [28] model is a distilled version of BERT. It has 40% fewer parameters than the original BERT model. Both models were trained in English, and adjustments have been made to perform multiple natural language processing tasks, particularly classification. For the Spanish language, there are fewer models available. One notable model is "roberta-base-bne," a transformer-based masked language model specifically designed for Spanish [29]. It is built upon the RoBERTa base model and is pre-trained using a vast Spanish corpus obtained from web crawlings conducted by the National Library of Spain (Biblioteca Nacional de España).

For the purpose of this project, the previous models were revisited, and the *Hugging Face* library was used to perform a fine-tuning process. Pre-trained models shared by other users of the library were employed. Specifically, for the English language, models that had been pre-trained using sexist content classification on Twitter were identified, making them suitable for our purpose. In contrast, although multiple models were available for the Spanish language, many of them did not align with the required binary classification task.

The fine-tuning process entailed performing a grid search on the following parameters:

- Learning rate (lr)
- Weight decay (wd)
- Number of training epochs (epochs)

We used the models *NLP-LTU/bertweet-large-sexism-detector* [30] and *NLP-LTU/distilbert-sexism-detector* [31] for the English language, *hackathon-somos-nlp-2023/roberta-base-bne-finetuned-suicide-es* [32] and *edu-munozsala/roberta_bne_sentiment_analysis_es* [33] for the Spanish language.

3.3. Task 2 - Task 3

Based on the models generated using the transformer approach, we identified models that effectively captured sexist behaviors and provided accurate scores. Leveraging the knowledge gained from these models, we proceeded to extract their embeddings and incorporate them as features in traditional machine learning classifiers. Specifically, an SVM was employed for the source intention (Task 2), and a Random Forest was utilized for the sexism categorization (Task 3). Both the SVM and Random Forest models were implemented using Python's scikit-learn library [24] with default parameters.

4. Results

In this section, we assess the effectiveness of our approaches. Section 4.1 presents the results of our experiments on an internal evaluation dataset, while Section 4.2 showcases our findings on the EXIST 2023 test set.

4.1. Internal results

The purpose of doing a preliminary evaluation is to observe the performance of different models and ensembles using an in-house data partition. The aim was to determine the best models and configurations for both our traditional machine learning models and *transformer*. Subsequently, we decided to prioritize the *transformer* BERT based models in the evaluation as they exhibited better metrics in-house experiments.

4.1.1. Task 1

For the first approach of binary classification, we present the top 5 models based on the F1 metric. It is important to highlight that in the Spanish language , better values were obtained compared to English (Table 1), although none of the models achieved a metric higher than 0.7. Furthermore, it is very interesting to observe that, regardless of the language, the inclusion of features such as the addition of counts of sexist expressions, counts of emojis, and the use of pre-trained fasttext embeddings improved the obtained F1 scores.

For the second approach, it is worth noting that regardless of the language, the F1 score was improved by more than 10 percent with respect to the classical machine learning approach after fine-tuning. However, the English language appears to have more consistent metrics across the models.

Referring to the previous results, the following models were considered for evaluation in EXIST 2023 for Task 1:

- iimasGIL_NLP_1, Spanish Id model : m11 and English Id model : m1.
- iimasGIL_NLP_2, Spanish Id model : m11 and English Id model : m2.
- iimasGIL_NLP_3, Spanish Id model : m11 and English Id model : m3.

Table 1
Internal results for binary classification (Task 1)

English			
Id model	Model	Parameters	F1
Transformer			
m1	bertweet-large-sexism-detector	lr=2e-05, wd=0.01,epochs=3	0.78
m2	distilbert-sexism-detector	lr=3e-06,wd=0.01,epochs=3	0.73
m3	bertweet-large-sexism-detector	lr=2e-05, wd=0.0001,epochs=3	0.69
m4	distilbert-sexism-detector	without fine tunning	0.48
m5	bertweet-large-sexism-detector	without fine tunning	0.47
Traditional ML			
m6	SVM	features = fasttext embeddings	0.62
m7	Random Forest	features = BoW and emojis counts,min_df =3	0.62
m8	Random Forest	features = TF-IDF, emojis counts and ingroup index, min_df =3	0.62
m9	Random Forest	features = BoW, emojis counts,min_df =3	0.62
m10	Random Forest	features = BoW, ig_ind,min_df =3	0.62
Spanish			
Id model	Model	Parameters	F1
Transformer			
m11	roberta-base-bne-finetuned-suicide-es	lr=1e-05,wd=0.0001, epochs=3	0.80
m12	edumunozsalaroberta_bne_sentiment_analysis_es	lr=3e-05,wd=0.0001, epochs=3	0.77
m13	edumunozsalaroberta_bne_sentiment_analysis_es	without fine tunning	0.20
m14	roberta-base-bne-finetuned-suicide-es	without fine tunning	0.13
Traditional ML			
m15	SVM	features = fasttext	0.68
m16	SVM	features = fasttext and sexist exp.	0.68
m17	Logistic Regression	features = fasttext	0.65
m18	Random Forest	features = BoW, sexist exp., min_df=2	0.65
m19	Naive Bayes	features = BoW, sexist exp., min_df=3	0.64

4.1.2. Task 2

Considering the modeling for task 2 that was previously described, we can notice that there are broad areas of opportunity. There are categories that are difficult to classify, particularly the 'JUDGMENTAL' category. The development dataset results are presented in Table 2, which includes the weighted average (calculated by averaging the support-weighted mean per label). Task two involves labeling that is performed after detecting sexism in the tweet, which results in a smaller training set. Additionally, there is a class imbalance in the training data, which further adds complexity to this task.

4.1.3. Task 3

In order to model the type of sexism represented by each tweet, we utilized the extracted embeddings from each model and performed a random forest analysis. The development dataset results are presented in Table 3, which includes the weighted average (calculated by averaging the support-weighted mean per label) and sample average. Our findings

Table 2
Internal results for source intention classification (Task 2)

English			
Id model	Model	Parameters	Weighted average F1
m20	SVM	features = embeddings from model m1	0.35
m21	SVM	features = embeddings from model m2	0.43
m22	SVM	features = embeddings from model m3	0.39
Spanish			
Id model	Model	Parameters	Weighted average F1
m23	SVM	features = embeddings from model m11	0.37

indicate that this methodology yields greater effectiveness compared to the previous task in particular for the Spanish tweets.

Table 3
Internal results on multilabel classification of sexism types (Task 3)

English			
Id model	Model	Parameters	Weighted average F1
m24	Random Forest	features = embeddings from model m1	0.4
m25	Random Forest	features = embeddings from model m2	0.23
m26	Random Forest	features = embeddings from model m3	0.39
Spanish			
Id model	Model	Parameters	Weighted average F1
m27	Random Forest	features = embeddings from model m11	0.6

4.2. EXIST 2023 results

In the following sections, we present the results obtained in the EXIST 2023 task. To provide a meaningful context, we include the rankings of the baselines for the same task, the top performer's result, and our own achieved results in each evaluation.

4.2.1. Task 1

The results obtained in the task (Table 4) can be divided into three points:

- The results obtained when considering the complete model in English and Spanish were not favorable. In both evaluations (Soft/Hard), our results are at the lower end of the table.

- When considering only the English model, the results were satisfactory, as one of our outputs ranks fifth among the top results.
- When considering the Spanish model only, the results are considerably low, highlighting the need for improvement in several areas.

Table 4
EXIST 2023 results for binary classification (Task 1)

Evaluation type	Lang.	Run	Rank	ICM (S/H)	ICM norm	Cross Ent.	F1
Soft vs Soft	ALL	iimasGIL_NLP_3	44	-1.55	0.18	1.59	-
Soft vs Soft	ALL	iimasGIL_NLP_2	45	-1.55	0.18	1.59	-
Soft vs Soft	ALL	iimasGIL_NLP_1	46	-1.55	0.18	1.59	-
Hard vs Hard	ALL	iimasGIL_NLP_3	30	0.50	0.69	-	0.76
Hard vs Hard	ALL	iimasGIL_NLP_1	36	0.48	0.67	-	0.75
Hard vs Hard	ALL	iimasGIL_NLP_2	39	0.46	0.66	-	0.75
Hard vs Soft	ALL	iimasGIL_NLP_3	34	0.10	0.51	-	-
Hard vs Soft	ALL	iimasGIL_NLP_1	37	0.06	0.51	-	-
Hard vs Soft	ALL	iimasGIL_NLP_2	40	-0.03	0.49	-	-
Soft vs Soft	ES	iimasGIL_NLP_1	44	-1.55	0.18	1.59	-
Soft vs Soft	ES	iimasGIL_NLP_2	45	-1.55	0.18	1.59	-
Soft vs Soft	ES	iimasGIL_NLP_3	46	-1.55	0.18	1.59	-
Hard vs Hard	ES	iimasGIL_NLP_1	37	0.434	0.63	-	0.76
Hard vs Hard	ES	iimasGIL_NLP_2	38	0.434	0.63	-	0.76
Hard vs Hard	ES	iimasGIL_NLP_3	39	0.434	0.63	-	0.76
Hard vs Soft	ES	iimasGIL_NLP_1	38	0.08	0.47	-	-
Hard vs Soft	ES	iimasGIL_NLP_2	39	0.08	0.47	-	-
Hard vs Soft	ES	iimasGIL_NLP_3	40	0.08	0.47	-	-
Soft vs Soft	EN	iimasGIL_NLP_3	45	-2.44	0.20	2.16	-
Soft vs Soft	EN	iimasGIL_NLP_2	46	-2.50	0.19	2.06	-
Soft vs Soft	EN	iimasGIL_NLP_1	51	-3.02	0.11	3.00	-
Hard vs Hard	**EN**	**iimasGIL_NLP_3**	**5**	**0.57**	**0.75**	-	**0.75**
Hard vs Hard	EN	iimasGIL_NLP_1	25	0.51	0.71	-	0.73
Hard vs Hard	EN	iimasGIL_NLP_2	31	0.48	0.70	-	0.73
Hard vs Soft	EN	iimasGIL_NLP_3	6	0.05	0.56	-	-
Hard vs Soft	EN	iimasGIL_NLP_1	22	-0.05	0.54	-	-
Hard vs Soft	EN	iimasGIL_NLP_2	41	-0.23	0.52	-	-

4.2.2. Task 2

In regards to the results obtained in EXIST 2023 (Table 5), our strongest performance can be observed in the soft evaluations. In this assessment, our scores surpass the average obtained by all participants.

Table 5
EXIST 2023 results for source intention classification (Task 2)

Evaluation type	Lang.	Run	Rank	ICM (S/H)	ICM norm	Cross Ent.	F1
Soft vs Soft	**ALL**	**iimasGIL_NLP_3**	**13**	**-3.51**	**0.75**	**1.89**	-
Soft vs Soft	ALL	iimasGIL_NLP_1	15	-3.56	0.75	1.91	-
Soft vs Soft	ALL	iimasGIL_NLP_2	16	-3.64	0.75	1.91	-
Hard vs Hard	ALL	iimasGIL_NLP_3	31	-0.99	0.46	-	0.29
Hard vs Hard	ALL	iimasGIL_NLP_1	32	-1.06	0.45	-	0.25
Hard vs Hard	ALL	iimasGIL_NLP_2	33	-1.08	0.44	-	0.26
Hard vs Soft	ALL	iimasGIL_NLP_1	28	-10.15	0.58	-	-
Hard vs Soft	ALL	iimasGIL_NLP_2	29	-10.60	0.57	-	-
Hard vs Soft	ALL	iimasGIL_NLP_3	30	-10.74	0.57	-	-
Soft vs Soft	ES	iimasGIL_NLP_1	13	-3.17	0.73	1.83	-
Soft vs Soft	ES	iimasGIL_NLP_2	14	-3.17	0.73	1.83	-
Soft vs Soft	ES	iimasGIL_NLP_3	16	-3.28	0.73	1.84	-
Hard vs Hard	ES	iimasGIL_NLP_3	25	-0.5978	0.52	-	0.37
Hard vs Hard	ES	iimasGIL_NLP_1	28	-0.8216	0.47	-	0.32
Hard vs Hard	ES	iimasGIL_NLP_2	29	-0.8216	0.47	-	0.32
Hard vs Soft	ES	iimasGIL_NLP_3	26	-8.22	0.59	-	-
Hard vs Soft	ES	iimasGIL_NLP_1	27	-8.39	0.58	-	-
Hard vs Soft	ES	iimasGIL_NLP_2	28	-8.39	0.58	-	-
Soft vs Soft	EN	iimasGIL_NLP_3	14	-3.86	0.78	1.94	-
Soft vs Soft	EN	iimasGIL_NLP_1	15	-4.14	0.78	1.99	-
Soft vs Soft	EN	iimasGIL_NLP_2	16	-4.34	0.77	2.01	-
Hard vs Hard	EN	iimasGIL_NLP_1	31	-1.39	0.42	-	0.14
Hard vs Hard	EN	iimasGIL_NLP_2	32	-1.42	0.42	-	0.20
Hard vs Hard	EN	iimasGIL_NLP_3	33	-1.49	0.40	-	0.20
Hard vs Soft	EN	iimasGIL_NLP_1	30	-13.78	0.56	-	-
Hard vs Soft	EN	iimasGIL_NLP_2	31	-14.59	0.55	-	-
Hard vs Soft	EN	iimasGIL_NLP_3	32	-15.22	0.53	-	-

4.2.3. Task 3

Lastly, in task 3, we observed the effectiveness of using embeddings as variables, which is also reflected in the evaluation of EXIST 2023 (Table 6). In general, we achieved higher positions in the English language, and notably, we obtained the highest score in the Hard-Soft evaluation for this language.

Table 6
EXIST 2023 results on multilabel classification of sexism types (Task 3)

Evaluation type	Lang	Run	Rank	ICM (S/H)	ICM norm	F1
Soft vs Soft	ALL	iimasGIL_NLP_3	11	-7.77	0.69	-
Soft vs Soft	ALL	iimasGIL_NLP_2	12	-7.81	0.69	-
Soft vs Soft	ALL	iimasGIL_NLP_1	13	-7.89	0.69	-
Hard vs Hard	ALL	iimasGIL_NLP_3	19	-0.65	0.47	0.45
Hard vs Hard	ALL	iimasGIL_NLP_1	21	-0.69	0.46	0.44
Hard vs Hard	ALL	iimasGIL_NLP_2	22	-0.78	0.45	0.43
Hard vs Soft	ALL	iimasGIL_NLP_2	19	-14.34	0.57	0.1069
Hard vs Soft	ALL	iimasGIL_NLP_3	20	-14.69	0.57	-
Hard vs Soft	ALL	iimasGIL_NLP_1	21	-14.96	0.56	-
Soft vs Soft	ES	iimasGIL_NLP_2	16	-11.37	0.62	-
Soft vs Soft	ES	iimasGIL_NLP_3	17	-11.50	0.62	-
Soft vs Soft	ES	iimasGIL_NLP_1	18	-11.65	0.61	-
Hard vs Hard	ES	iimasGIL_NLP_2	20	-0.7932	0.45	0.46
Hard vs Hard	ES	iimasGIL_NLP_3	21	-0.801	0.45	0.46
Hard vs Hard	ES	iimasGIL_NLP_1	22	-0.84	0.45	0.45
Hard vs Soft	ES	iimasGIL_NLP_3	21	-17.95	0.50	-
Hard vs Soft	ES	iimasGIL_NLP_2	22	-18.09	0.50	-
Hard vs Soft	ES	iimasGIL_NLP_1	23	-18.57	0.49	-
Soft vs Soft	EN	iimasGIL_NLP_3	5	-3.14	0.78	-
Soft vs Soft	EN	iimasGIL_NLP_1	6	-3.21	0.78	-
Soft vs Soft	EN	iimasGIL_NLP_2	7	-3.36	0.78	-
Hard vs Hard	EN	iimasGIL_NLP_3	20	-0.54	0.48	0.38
Hard vs Hard	EN	iimasGIL_NLP_1	21	-0.58	0.47	0.37
Hard vs Hard	EN	iimasGIL_NLP_2	24	-0.83	0.42	0.31
Hard vs Soft	**EN**	**iimasGIL_NLP_2**	**2**	**-8.91**	**0.68**	-
Hard vs Soft	EN	iimasGIL_NLP_1	4	-9.71	0.66	-
Hard vs Soft	EN	iimasGIL_NLP_3	5	-9.88	0.66	-

5. Conclusions

The paper presents our approach to detecting sexism in social media within the framework of the EXIST 2023 Lab at CLEF. We focus on data exploration to generate linguistic patterns for characterizing sexist messages in social networks. Additionally, we evaluate models based on fine-tuned transformers using the Exist 2023 test corpus.

Although our work on extracting natural language regularities and patterns in sexist texts has mainly been devoted to Spanish, our best results have been obtained in English sub-tasks. This could mean that our BERT resources are optimized to deal with the English language.

Beyond the results, we are interested in the relation between the socio-demographic features of the tweets' authors and the classification annotators have given. Also, we think that the analysis of the language of sexist tweets can help to understand how this

type of discrimination works in social networks and how the annotators agree or disagree with certain expressions or attitudes of the authors.

Finally training based on disagreement proved to be challenging due to the absence of a traditional guideline, leading to potential bias when evaluating a model. Moreover, limited training data amplifies the potential for noise introduced by label changes. This was evident in the results across all teams, as there was significant variability in the ratings.

Acknowledgments

This work has been carried out with the support of DGAPA-UNAM PAPIIT project numbers TA400121 and TA101722. CONACYT Scholarship CVU 1233219. GBE is supported by a grant from the Ministry of Universities of the Government of Spain, financed by the European Union, NextGeneration EU (María Zambrano program). The authors also thank CONACYT for the computing resources provided through the Deep Learning Platform for Language Technologies of the INAOE Supercomputing Laboratory. We also want to thank Eng. Roman Osorio for supporting the student administration of the project.

References

[1] G. Masequesmay, Sexism, accessed May 30, 2023. URL: https://www.britannica.com/topic/sexism.

[2] A. Michel, Down with stereotypes! Eliminating sexism from children's literature and school textbooks, UNESCO, 1986.

[3] D. April, B. Barakat, M. Barry, N. B. et al., Global education monitoring report, 2020, Latin America and the Caribbean: inclusion and education: all means all, UNESCO, 2020.

[4] A. J. Kingston, T. Lovelace, Sexism and reading: A critical review of the literature, Reading Research Quarterly 13 (1977) 133–161. URL: http://www.jstor.org/stable/747592.

[5] S. Miller, Language and Sexism, Cambridge University Press, 2009.

[6] J. Stout, N. Dasgupta, When he doesn't mean you: Gender-Exclusive language as ostracism, Pers. Soc. Psychol. Bull. (2011).

[7] L. Davis, M. Reynolds, Gendered language and the educational gender gap, Econ. Lett. (2018).

[8] H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, S. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. Seligman, L. Ungar, Personality, gender, and age in the language of social media: The Open-Vocabulary approach, PLoS One (2013).

[9] V. Gay, D. Hicks, E. Santacreu-Vasut, A. Shoham, Decomposing culture: an analysis of gender, language, and labor supply in the household, Review of Economics of the Household (2018). doi:10.1007/s11150-017-9369-x.

[10] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23, Springer, 2018, pp. 57–64.

[11] J. Bartlett, R. Norrie, S. Patel, R. Rumpel, S. Wibberley, Misogyny on twitter (2014).

[12] T. Farrell, M. Fernandez, J. Novotny, H. Alani, Exploring misogyny across the manosphere in reddit, in: Proceedings of the 10th ACM Conference on Web Science, WebSci'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 87–96. URL: https://doi.org/10.1145/3292522.3326045. doi:10.1145/3292522.3326045.

[13] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[14] F. M. Plaza-del Arco, M. Casavantes, H. J. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes, H. Jarquín-Vásquez, L. Villaseñor-Pineda, et al., Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants, Procesamiento del Lenguaje Natural 67 (2021) 183–194.

[15] M. Taulé, A. Ariza, M. Nofre, E. Amigó, P. Rosso, Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish, Procesamiento del Lenguaje Natural 67 (2021) 209–221.

[16] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, 2023. arXiv:2303.04222.

[17] 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.

[18] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remon, 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.

[19] 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. Experimental IR Meets Multilinguality, Multimodality, and Interaction., in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, , N. Ferro (Eds.), Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), CLEF'23, Thessaloniki, Greece, 2023.

[20] 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, CLEF '23, Thessaloniki, Greece, 2023.
[21] B. Chulvi, A. Toselli, P. Rosso, Fake news and hate speech: Language in common, arXiv preprint arXiv:2212.02352 (2022).
[22] B. Chulvi, M. Mariangeles Molpeceres, A. Toselli, F. Rodrigo, P. Rosso, Politicization of immigration and language use in political elites: A study of spanish parliamentary speeches, Journal of Language and Social Psychology (2023).
[23] S. Corbara, B. Chulvi, P. Rosso, A. Moreo, Rhythmic and psycholinguistic features for authorship tasks in the spanish parliament: Evaluation and analysis, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer, 2022, pp. 79–92.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[26] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2020).
[27] 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, arXiv preprint arXiv:1907.11692 (2019).
[28] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[29] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.
[30] NLP-LTU, bertweet-large-sexism-detector, https://huggingface.co/NLP-LTU/bertweet-large-sexism-detector, 2023.
[31] NLP-LTU, Nlp-ltu/distilbert-sexism-detector, https://huggingface.co/NLP-LTU/distilbert-sexism-detector, 2023.
[32] hackathon-somos-nlp 2023, hackathon-somos-nlp-2023/roberta-base- bne-finetuned-suicide-es, https://huggingface.co/hackathon-somos-nlp-2023/roberta-base-bne-finetuned-suicide-es, 2023.
[33] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.