ROH_NEIL@EXIST2023: Detecting Sexism in Tweets using Multilingual Language Models

Rohit Kooniredddy[1,*,†], Niloofar Adel[1,‡]

[1]University of Zurich, 8006 Zurich, Switzerland

Abstract

This paper describes a submission to the EXIST 2023 challenge for binary, multi-class, and multi-label classification tasks, targeting the detection of sexism in English and Spanish tweets. Our approach employs a cross-lingual transformer model for these tasks, with a specific emphasis on "discrete" (hard-label) classification. Through a comparative analysis with other transformer-based models, our final model demonstrates effectiveness of the cross-lingual transformer models in achieving competitive performance. Notably, our model also achieves favorable rankings in hard-labeling for all the tasks. These results underscore the potential of cross-lingual models in accurately classifying English and Spanish text data without relying on language-specific models.

Keywords

Transformers, Cross-lingual Language Models, Multilingual Language Models, Sexism Detection, Sexism Categorization

1. Introduction

Gender-based discrimination in the form of sexism remains a pervasive issue that continues to affect digital interactions, posing significant challenges in the creation of inclusive and respectful online spaces. The rapid growth of social media platforms has exacerbated the dissemination of sexist content, emphasizing the urgent need for automated approaches to detect and classify such content.

The objective of this paper is to identify sexism in tweets by utilizing pre-trained transformer models and leveraging the data provided by the EXIST 2023 challenge. This study aims to harness the capabilities of pre-trained transformer models, specifically tailored to excel in natural language processing (NLP), to develop a robust and meticulous classification system capable of distinguishing tweets containing sexist content. The EXIST 2023 challenge provides a carefully curated dataset to detect sexism in both English and Spanish languages' tweets, serving as an optimal foundation for training and evaluating our models. The implications of sexism classification research extend beyond academia and hold the potential to make a profound impact on combating sexism within online spaces.

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

*Main and Corresponding author.

†The source code for the findings of this study can be found here: GitHub for EXIST2023 roh-neil

rohit.kooniredddy@uzh.ch (R. Kooniredddy); niloofar.adel@uzh.ch (N. Adel)

https://github.com/rkooniredddy (R. Kooniredddy)

In the following sections, we will delve into pertinent existing literature, describe the dataset used for this challenge, detail the methodology centered on pre-trained transformer models, present experimental results, draw conclusions, and outline future directions for this research.

2. Related Works

Detecting sexism and offensive language in social media conversations and general text has garnered significant attention, leading to the development of various models and approaches. Lexical and linguistic analysis, leveraging word embeddings and semantic similarity, has proven effective in identifying instances of gender bias and potential sexism [1]. Until the recent past, machine learning techniques, particularly Recurrent Neural Networks (RNNs), have been widely adopted for the classification of social media posts as sexist or non-sexist by capturing sequential dependencies in the text [2].

With the advent of transformers since the seminal work of "Attention is All You Need" by Vaswani et al. [3], the utilization of transformers for classification tasks has garnered significant attention. Transformers have demonstrated superior accuracy compared to their predecessors. Notably, XLM models have recently achieved state-of-the-art performance in various benchmark tasks [4, 5]. XLM transformers have emerged as a powerful tool for text classification tasks, particularly in the domain of cross-lingual language modeling. XLM, which stands for Cross-lingual Language Model, is specifically designed to effectively handle multiple languages [4]. By leveraging large-scale pre-training on diverse multilingual corpora, XLM transformers capture comprehensive language representations that can be transferred across different languages [5]. This cross-lingual capability enables XLM transformers to generalize well to languages with limited training data, facilitating effective knowledge transfer from high-resource to low-resource languages [6]. Furthermore, XLM transformers incorporate cross-lingual alignment techniques, enabling the alignment of word and sentence embeddings across multiple languages. This facilitates cross-lingual transfer learning and leads to enhanced performance in multilingual text classification tasks [5]. Domain-specific models tailored for sentiment analysis on Twitter, such as the XLM-T models, have also demonstrated exceptional performance [7].

This is the third edition of the EXIST Shared Task [8]. In the editions of the past two years, plenty of methods and models have been used to solve the challenge. In the first edition in 2021, ensembles of language-specific models tailored for English (RoBERTa) and Spanish (BETO) achieved the best results. Similarly, in the second edition in 2022, ensembles of different language-specific transformer models, including BERTweet-large [9], RoBERTa, DeBERTa v3 [10] for English, and BETO, BERTIN [11], MarIA-base [12], and RoBERTuito for Spanish, achieved the best results.

Given the enduring success of transformer-based models, we initially employed language-specific models to solve the tasks and subsequently explored the use of superior Multilingual Language Models.

3. Dataset and Tasks Description

EXIST 2023 edition presents a hierarchical classification task. First, each tweet is labeled manually by 6 different annotators either as "SEXIST" (indicated with "YES") or as "NON-SEXIST" (indicated with "NO"). The percentage of annotators categorizing a tweet as "SEXIST" and "NON-SEXIST" are given as a soft-labels. The label with larger percentage is considered the hard-label (SEXIST% > NON-SEXIST% => "YES", else if "NO", else "-"). Tweets are labeled with "-" in tasks 2 and 3 when an annotator classifies them as "NON-SEXIST" or "NO" in the first task. Additionally, the label "UNKNOWN" is assigned to tweets for which no clear label was provided by the majority of annotators [8]. We safely ignore these instances of "-" and "unknown" as these labels wont affect the final classification.. Once the tweet is labeled as either, annotators further label them in tasks 2 and 3. We observed (almost) even distribution of tweets across languages, labels, annotators, and age groups, as shown in Figure 1. More details about the labelling process can be found in EXIST2023 overview[8].

Table 1 shows the details about the given "hard-labels" training and development data sets for EXIST 2023. Tables 8, 9, 10 present examples of hard-labels and soft-labels for each task.

The test data was a mix of English (978) and Spanish(1098) tweets accounting for a total of 2076 tweets.

Table 1
EXIST 2023 Dataset distribution for hard-labels, EXIST2023

	Twitter				
	Training		Development		
	Spanish	English	Spanish	English	Total
	Task 1				
Sexist	1560	1137	261	194	3152
Non-sexist	1634	1733	229	250	3846
	Task 2				
Direct	749	545	117	87	1498
Reported	265	194	40	35	534
Judgemental	228	148	55	28	459
	Task 3				
Ideological-inequality Misogyny-non-sexual-violence Objectification Sexual-violence Stereotyping-dominance	Values are not shown since each tweet was given multiple labels. A representation of data can be seen in Table 10.				

EXIST 2023 demonstrates a higher level of evolution and complexity in comparison to its previous two editions. The dataset used for classification purposes is smaller in size than the datasets employed in previous instances. Furthermore, the number of tasks has increased from two to three, and soft-labels are now incorporated. Hence, in an attempt to use larger corpus, we use data from EXIST 2021 for one run of task 1 hard-label classification. Some more information about the datasets used in EXIST 2021 and EXIST 2022 can be found in the Table 11.

3.1. Tasks in EXIST 2023

All the three tasks in the hierarchy are classification tasks, two in the area of classification (binary and multi-class classification) and the last one is multi-label classification (categorization). Compared to the discrete label classification tasks from EXIST 2021 and EXIST 2022, this edition requires participants to publish probability values associated with each prediction which are called soft-labels.

3.1.1. Task 1- Binary Classification

The first task was to carry out a binary classification in which the models had to classify the tweets into two classes, namely: "SEXIST" tweets and "NON-SEXIST" tweets. It is important to note that this task is used as an initial step for the other two tasks, meaning that tweets classified as not sexist in the first task were not considered in the second and third tasks. Some examples of tweets in the dataset with the predicted labels are shown in Table 8[13].

3.1.2. Task 2 - Multi-class Classification

The second task is a ternary classification task. The tweets that were classified as Sexist (with the label YES) in the previous task, must be classified into 3 different classes in accordance with their creator's intention, thus revealing the role that social networks play in generating and disseminating sexist posts. Table 9 demonstrates some examples of labeled tweets in this task. The classes are[13]:

1. DIRECT
2. REPORTED
3. JUDGEMENTAL

3.1.3. Task 3 - Multi-label Classification

Task 3 is a multi-label classification task. Here, in contrast to the previous tasks, each annotator could label the tweet with a list of labels and as a result, a list of hard-labels are expected instead of just one. Table 9 demonstrates some examples of labeled tweets in this task. The classes are [13]:

1. IDEOLOGICAL AND INEQUALITY
2. STEREOTYPING AND DOMINANCE
3. OBJECTIFICATION
4. SEXUAL VIOLENCE
5. MISOGYNY AND NON-SEXUAL VIOLENCE

4. Experimental Setup

We decided to exclusively utilize pre-trained Transformer models based on their impressive performance in previous editions of EXIST. To ensure both robustness and ease of implementation, we selected HuggingFace's [14] Trainer API and Optuna [15] as our helping architectures,

minimizing the need for custom code development. Our experimentation involved a comprehensive evaluation of existing General Language and Tweet-specific language (NLP) models. For the initial model evaluation, we focused on Task 1, specifically the prediction of hard-labels. Various parameter configurations were tested for each transformer model, as shown in the Table 2. To perform the initial model evaluation, the training data and development data were merged into a single file and then split into an 80:10:10 ratio for training, validation, and testing folds. Two distinct sets of parameters were selected to assess the performance of each pre-trained language model.

Once the best model(s) is chosen from this initial selection stage, Optuna [15] was used to search for the best hyperparameters for each task again for both hard-label and soft-label predictions. Details about the hyperparameters tuning are available in the GitHub repository [16]. Once the best hyperparameters were found for each task and sub-task, the merged data is split into 85:15 ratio for training and validation sets. The final models are used to predict the official test data items and corresponding labels were submitted according to the organizers' requirements [8]. For ease of understanding, the flowchart 2 illustrates the steps involved to complete this study.

4.1. Initial Model Selection

In the 2021 edition of EXIST, a majority of those who submitted the results, approached the tasks using transformer-based models. Key architectures used were BERT[17] (or multilingual BERT - mBERT), Spanish version of BERT called BETO[18], RoBERTa[19] and a multilingual version of RoBERTa called XLM-R[4]. In the 2022 edition, a majority of those who submitted used transformer models. DeBERTa[10] and RoBERTuito[20] was also used along with transformer models employed in 2021.

Based on the performance of these models in the past two years [21, 22, 23, 24] and an exploration of available HuggingFacemodels, we carefully selected a group of large language models pre-trained for classification, as presented in Table 2. (Please note that this is not a comprehensive list but presents the outlook of the scale of exploration. The models are chosen such that are equipped with both general language and twitter data understanding.) In this table 2, under the column "Data Type" three labels: "xlm" , "es" and "en" can be identified which represent the type of the data used for model training and evaluation. "xlm" represents combined data of Spanish and English tweets, as given for each task. "en" represents only English tweets where the original Spanish tweets were translated to their English versions using google translate[25]. "es" represents only Spanish language tweets where English tweets are translated into the Spanish language.

Once the datasets are prepared as required, each model is tested with varying hyperparameters including choosing an embedding length. All the pre-trained transformer models used in the task are "Sequence Classification" models as required by the tasks in this challenge.

Upon inspecting the models' performance 2, we observed that the cross-lingual model pre-trained on twitter data called "sdadas/xlm-roberta-large-twitter" provides the best test scores. In the rest of the paper, this model is called "XLM-T-10-L" model. This is a XLM-RoBERTa-large model tuned on a corpus of over 156 million tweets in ten languages: English, Spanish, Italian, Portuguese, French, Chinese, Hindi, Arabic, Dutch and Korean. The model has been trained

from the original XLM-RoBERTA-large checkpoint for 2 epochs with a batch size of 1024. The model has 560 million parameters. [26].

4.2. Preprocessing and Data Augmentation

In our study, we adopt a straightforward preprocessing approach, which aligns with the recommended practices outlined in the usage guide of the XLM-T-10-L model. The preprocessing method focuses on handling tags and URLs present in the tweets. Any user handle encountered in the tweets is replaced with "@USER", while URLs are replaced with "#HTTPURL". We refrain from implementing additional preprocessing steps in order to retain the intricacies and unique characteristics inherent in the tweet data. By avoiding excessive preprocessing, we aim to preserve the originality and nuances of the tweet content, allowing the models to capture the genuine nature of the tweets during the subsequent analysis and classification stages.

We also utilized the data from the EXIST 2021 edition to augment the provided training data for one run of the first task. This additional data was also translated into full Spanish and full English versions using googletrans[25]. This data was also used to examine the performance of language-specific models such as BERTweet-large [9] RoberTuito [20] on the English and Spanish data respectively. These models performed well, but not as well as the XLM-RoBERTa-Large[26] based models as seen in the Table 2.

4.3. Hyperparameter Search

It is imperative to test hyperparameters in machine learning models to determine performance, particularly in transfer learning tasks and also when training data is limited, as in this challenge. Since hyperparameter tuning is known to play an important role in improving the performance and generalization of the model, an open-source framework that allows hyperparameters search, called Optuna[15] together with HuggingFace Transformers and TrainerAPI is used. The Optuna solution offers a versatile and effective way to identify optimal hyperparameters automatically, which reduces the need for manual tuning and boosts the development speed of your application, thus reducing the reliance on manual tuning.

Optuna works by defining an objective function that represents the evaluation metric or loss function of the model. This objective function takes an arbitrary set of hyperparameters and their search space as input and returns a score or value that represents the performance of the model with those hyperparameters. This function could be a loss function to minimize or a metric to maximize. Optuna's goal is to find the set of hyperparameters that optimize this objective function through Optuna's method called create_study[15].

We used Optuna determine the "Learning Rate", "Weight Decay" and "Number of Epochs" for each task and its subsequent hard and soft-label predictions as shown in Table 3. More details can be found here.

4.4. Training Model

Once we found desired best hyperparameters using Optuna, we trained models for each task using HuggingFace trainer API as shown in the part of source code.. We ran each task (except task 2) three times and the results were published as runs 1, 2, 3. During training, as described

Table 2
Results for testing models using Task1 hard-label predictions

Example Parameters (one of the two parameter sets used)
TrainingArguments(output_dir="./output",
num_train_epochs=3,
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
learning_rate=2e-5,
weight_decay=0.01,
logging_dir="./logs",
evaluation_strategy="epoch",
save_strategy="epoch",
logging_strategy="epoch")

Model	**Data Type**	**Embed Length**	**Validation Set**		**Test Set**	
			Accuracy	**F1-Score**	**Accuracy**	**F1-Score**
amberoad/bert-multilingual-passage-reranking-msmarco	xlm	512	0.793	0.775	0.793	0.775
cardiffnlp/twitter-roberta-base-sentiment	xlm	512	0.814	0.797	0.809	0.782
microsoft/Multilingual-MiniLM-L12-H384	xlm	512	0.540	0.000	0.551	0.000
papluca/xlm-roberta-base-language-detection	xlm	512	0.540	0.000	0.551	0.000
symanto/xlm-roberta-base-snli-mnli-anli-xnli	xlm	512	0.710	0.633	0.701	0.600
cross-encoder/mmarco-mMiniLMv2-L12-H384-v1	xlm	128	0.793	0.775	0.793	0.775
xlm-roberta-large	xlm	256	0.540	0.000	0.551	0.000
jhu-clsp/bernice	xlm	128	0.846	0.833	0.851	0.835
sdadas/xlm-roberta-large-twitter	**xlm**	**256**	**0.844**	**0.830**	**0.856**	**0.835**
sdadas/xlm-roberta-large-twitter	**xlm**	**128**	**0.851**	**0.839**	***0.870***	***0.856***
xlm-roberta-base	xlm	128	0.827	0.820	0.829	0.811
Twitter/twhin-bert-base	xlm	128	0.850	0.839	0.844	0.828
Josue/BETO-espanhol-Squad2	es	512	0.737	0.679	0.701	0.623
cardiffnlp/twitter-roberta-base-sentiment	es	512	0.540	0.000	0.551	0.000
finiteautomata/beto-sentiment-analysis	es	256	0.827	0.807	0.830	0.811
dccuchile/bert-base-spanish-wwm-cased	es	256	0.836	0.821	0.847	0.829
mariav/bert-base-spanish-wwm-cased-finetuned-tweets	es	256	0.834	0.818	0.840	0.819
bertin-project/bertin-roberta-base-spanish	es	256	0.827	0.809	0.787	0.764
pysentimiento/robertuito-base-cased	es	128	0.844	0.829	0.840	0.823
JosePezantes/finetuned-robertuito-base-cased-V-P-G	es	128	0.840	0.826	0.814	0.801
hackathon-pln-es/paraphrase-spanish-distilroberta	es	128	0.831	0.817	0.829	0.808
mrm8488/bert-spanish-cased-finetuned-ner	es	128	0.836	0.824	0.823	0.807
Hate-speech-CNERG/dehatebert-mono-spanish	es	256	0.800	0.783	0.773	0.747
MMG/xlm-roberta-base-sa-spanish	es	128	0.829	0.821	0.820	0.804
JonatanGk/roberta-base-bne-finetuned-cyberbullying-spanish	es	128	0.800	0.778	0.779	0.741
dccuchile/albert-base-spanish-finetuned-xnli	es	128	0.799	0.783	0.799	0.771
cardiffnlp/roberta-base-tweet-sentiment-en	en	256	0.816	0.806	0.797	0.779
cardiffnlp/twitter-roberta-base-sentiment-latest	en	256	0.851	0.839	0.821	0.803
NLP-LTU/bertweet-large-sexism-detector	en	256	0.540	0.000	0.551	0.000

Table 3
Best hyperparameters for each task determined by Optuna

Task	Learning Rate	Weight Decay	Epochs
Task 1 (Hard)	3.3951×10^{-5}	0.0049	3
Task 2 (Hard)	0.0004	0.0054	4
Task 3 (Hard)	1.5592×10^{-5}	0.0002	2
Task 1 (Soft)	2.5174×10^{-5}	0.0096	3
Task 2 (Soft)	0.0006	6.2090×10^{-5}	2
Task 3 (Soft)	1.1213×10^{-6}	0.0005	2

earlier, given training data and development data are combined, shuffled, and then split into 85:15 training and validation datasets with which training takes place. Each run of the task was submitted to achieve the best scores in both hard-labeling and soft-labeling sub-tasks.

For Run 1 – Task 1 and Task 2- The model is trained with best hyperparameters to predict hard-labels. Once hard-labels are predicted, their corresponding logits from the model are converted to probabilities using the Softmax function. The predicted hard-labels and soft-labels are submitted together as Run 1. Binary Cross Entropy loss (BCELoss), and Categorical Cross Entropy loss (CCELoss) are used here.

For Run 1 - Task 3 - Instead of the Softmax function, Sigmoid is applied on logits to calculate the probabilities of hard-labels. BCELoss for each label is applied separately for the third task.

For Run 2 - Task 1, Task 2, Task 3 - The model is trained with best hyperparameters to predict the given soft-labels (probabilities for each label) and the output predictions of soft-labels are combined with output predictions of hard-labels from Run 1. PyTorch's existing functionality of CCELoss[27] which handles soft-labels is used for all the tasks as shown here.

For Run3 - Task 1 - Large augmented dataset which combined 2021 editions' twitter tweets data along with current data is used to train the model for hard-label predictions. Predictions and corresponding Softmax on logits are submitted.

For Run3 - Task 3 - Soft-label predictions from Run 2 and its corresponding "argmax" hard-labels are submitted.

4.5. Metrics

A variety of evaluation metrics, such as F1 Score, Cross Entropy, ICM-Soft, ICM-Soft Norm, ICM-Hard, and ICM-Hard Norm, are used by the organizers to assess model performance. Particularly, the Information Contrast Measure (ICM) is employed as a similarity function for evaluating hierarchical classification outputs, providing insights into class relationships and similarities (higher the better)[28].

In the evaluation, three types of assessment criteria are employed: hard-hard, hard-soft, and soft-soft. Hard-hard evaluation compares the system's output with the majority class determined by annotators' labels. Hard-soft evaluation compares the system's categories with the probabilities assigned to each category in the ground truth. Soft-soft evaluation compares the system's probabilities with those assigned by human annotators. These details can be observed in the Section 5 tables .

Table 4
Description and Direction of Evaluation Metrics

Metric	Description and Direction
ICM[28]	Information Contrast Measure (ICM) is a similarity function used to evaluate the outputs of classification systems in hierarchical classification tasks. It generalizes Pointwise Mutual Information (PMI) and measures the resemblance between the system's output and the ground truth labels. Higher values of ICM indicate better performance.
ICM-Soft	ICM-Soft is an evaluation metric that compares the categories assigned by the system with the probabilities assigned to each category in the ground truth. It considers the distribution of labels and the number of annotators assigned to each instance to determine the probability of the classes. Higher values of ICM-Soft indicate better performance.
ICM-Hard	ICM-Hard evaluation involves comparing the system's "hard" output with the hard ground truth labels. A probabilistic threshold is employed to extract the hard-labels from the ground truth, considering the approval of multiple annotators for each task. Only the most popularly labeled classes are included in this evaluation. Higher values of ICM-Hard indicate better performance.
ICM-Soft Norm	ICM-Soft Norm is a normalized version of ICM-Soft that takes into account the number of annotators assigned to each instance and adjusts the probabilities accordingly. It handles instances labeled as "UNKNOWN" by reducing the number of annotators considered based on the count of "UNKNOWN" labels associated with them. Higher values of ICM-Soft Norm indicate better performance.
ICM-Hard Norm	ICM-Hard Norm is a normalized version of ICM-Hard that adjusts the hard-labels based on the number of annotators assigned to each instance. It considers instances labeled as "UNKNOWN" and adjusts the threshold for label extraction accordingly. Higher values of ICM-Hard Norm indicate better performance.
F1 Score	F1 Score is a commonly used evaluation metric in classification tasks. It is the harmonic mean of precision and recall, weighted by the same values. The F1 score treats false positives and false negatives equally, assuming that both types of errors have the same consequences. Higher values of F1 Score indicate better performance.
Cross Entropy	Cross Entropy is a metric used to measure the difference between the predicted probabilities and the true probabilities. It quantifies the average amount of information needed to identify the true class given the predicted probabilities. Lower values of Cross Entropy indicate better model performance.

5. Results

The overview of the final results of this study's submissions can be found in the Tables 5, 6,7 with the prefix "roh-niel" and with a prefix 1 or 2 or 3 depending on the run that achieves the shown results (Information about each run is explained in previous sections). For comparison, metrics of the gold-labels (as provided by the organizers) and the best ranked results are provided for each task and criterion.

This study achieves favourable rankings in hard-labelling for Task 1, Task 2 and Task 3, , using pre-trained language model based on "XLM-T-10-L"[26], as shown below. These results highlight the potential of cross-lingual models to accurately classify English and Spanish text data without the need for language-specific model.

Table 5
Results for Task 1

Task 1

soft-soft all				
Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2023_test_gold_soft	0	3.1182	1	0.5472
SINAI_3	1	0.903	0.6421	0.796
roh-neil_2	49	-2.8848	0.0302	1.5472
hard-hard all				
Run	Rank	ICM-Hard	ICM-Hard Norm	F1 Score
EXIST2023_test_gold_hard	0	0.9948	1	1
Mario_3	1	0.6575	0.785	0.8109
roh-neil_1	4	0.5795	0.7353	0.784
hard-soft all				
Run	Rank	ICM-Soft	ICM-Soft Norm	-
EXIST2023_test_gold_soft	0	3.1182	1	-
EXIST2023_oracle_most_voted	1	1.1977	0.6897	-
Mario_3	2	0.4719	0.5725	-
roh-neil_1	5	0.3111	0.5465	-

Table 6
Results for Task 2

Task 2

soft-soft all				
Run	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2023_test_gold_soft	0	6.2057	1	0.9128
DRIM_1	1	-1.3443	0.8072	1.7833
roh-neil_1	23	-5.759	0.6945	3.7519
hard-hard all				
Run	Rank	ICM-Hard	ICM-Hard Norm	F1 Score
EXIST2023_test_gold_hard	0	1.5378	1	1
Mario_2	1	0.4887	0.7764	0.5715
roh-neil_1	2	0.3883	0.755	0.548
hard-soft all				
Run	Rank	ICM-Soft	ICM-Soft Norm	-
EXIST2023_test_gold_soft	0	6.2057	1	-
EXIST2023_oracle_most_voted	1	-2.3974	0.7803	-
UMUTeam_2	2	-5.12	0.7108	-
roh-neil_1	12	-6.522	0.675	-

5.1. Why "XLM-T-10-L" model performs well?

First, the XLM-T-10-L is purposefully designed and pre-trained on a comprehensive corpus of Twitter data, allowing it to capture the distinctive characteristics and intricate patterns inherent in tweets. This specialized training equips the model with the ability to excel in our specific context.

Table 7
Results for Task 3

Task 3

soft-soft all				
Run	Rank	ICM-Soft	ICM-Soft Norm	-
EXIST2023_test_gold_soft	0	9.4686	1	-
AI-UPV_3	1	-2.3183	0.7879	-
roh-neil_1	7	-6.6622	0.7098	-
hard-hard all				
Run	Rank	ICM-Hard	ICM-Hard Norm	F1 Score
EXIST2023_test_gold_hard	0	2.1533	1	1
roh-neil_1	1	0.4433	0.6763	0.6296
hard-soft all				
Run	Rank	ICM-Soft	ICM-Soft Norm	-
EXIST2023_test_gold_soft	0	9.4686	1	-
EXIST2023_oracle_most_voted	1	-8.3816	0.6788	-
EXIST2023_test_majority_class	2	-8.7089	0.6729	-
Mario_1	3	-9.1398	0.6652	-
roh-neil_1	13	-12.195	0.6102	-

Second, the model harnesses the transformer architecture renowned for its capacity to capture extensive contextual information and long-range dependencies in text. Despite the inherent brevity of tweet texts, the transformer architecture empowers the model to proficiently discern the underlying semantics and contextual nuances, taking into account the intricate interplay between words. Through the acquisition of comprehensive representations, the model adeptly extracts meaningful features and intricate patterns from tweet text.

Third, the model undergoes training on a diverse array of tweets spanning ten distinct languages, thereby demonstrating its aptitude for handling multilingual data. By leveraging the principles of cross-lingual transfer learning, the model effectively generalizes across languages, even in scenarios characterized by limited training data. This innate cross-lingual capability confers significant advantages in the domain of tweet classification, given the multilingual nature of tweets and the potential occurrence of transliteration phenomena.

In conclusion, the exemplary performance of the XLM-T-10-L can be attributed to its meticulous training on Twitter data, meticulous fine-tuning tailored for tweet classification tasks, the inherent effectiveness of the transformer architecture for representation learning, and the valuable cross-lingual capabilities it possesses. The synergistic interplay of these factors enables the model to achieve accurate understanding and proficient classification of tweets.

6. Conclusion and Future Work

In this study, our best solutions for tasks 1,2 and 3 of EXIST2023 are discussed and presented. A detailed study about the performance of existing transformer-based Language Models[14] is conducted. To arrive at the final results of the task, a two-level approach is taken. First, a study to find the best models using Task 1 hard-label classification is conducted. Second, the

best model is fine-tuned for each task and sub-task requirements using Optuna. Fine-tuned is trained on more data and then predictions are acquired and published.

This study's performance is sub-optimal for the soft-label classifications for each task. Poor performance for soft-labels may be majorly accounted to these 2 reasons:

- Utilizing incorrect loss function:
  While it is clear that using the logistic loss function is the appropriate approach for solving the soft-label task, we initially attempted to implement custom loss functions for each task based on matching distributions of the soft-labels. Specifically, we explored the adoption of the Kullback-Liebler Divergence loss and utilized the "sum reduction" variant of Cross Entropy Loss for each task. However, our analysis and findings suggest that a straightforward implementation of the logistic loss function will likely achieve superior soft-label metrics.

- Relying on task 1 hard-labels classification as the basis for model selection and subsequently applying it to soft-label tasks:
  In our methodology, we initially relied on task 1 hard-labels classification as the foundation for selecting the most suitable model. Subsequently, we utilized the same model for the soft-label tasks. This way of approaching the task may also have resulted in poor performance in the soft-label tasks.

It is certain that using the correct loss functions with the same models will better this study's results for all the tasks reflecting those of hard-label scores. The next steps would be to fine-tune models with standard logistic loss functions to achieve better soft-label predictions.

Acknowledgments

We would like to extend our sincere gratitude to the University of Zurich, Department of Computational Linguistics, for their invaluable support. Our sincere thanks go to Mr. Andrianos Michail and Dr. Simon Clematide for their invaluable guidance and assistance.

Additionally, we would like to express our deep appreciation to the organizers of EXIST2023 for providing us with the opportunity to participate in this esteemed event. We are truly grateful for the chance to present our work and contribute to the progress of our field.

We would like to once again acknowledge and appreciate the support and encouragement received from all individuals involved, as their contributions have been instrumental in our accomplishments.

References

[1] M. H. Ribeiro, T. Wu, K. P. Gummadi, J. Kleinberg, Word embeddings, bias in ai, and semantically unfair comparisons, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 64–74.

[2] P. Fortuna, C. Nunes, Recurrent neural network models for sexist content detection in social media, in: 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), IEEE, 2018, pp. 1–6.

[3] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017.
[4] G. Lample, A. Conneau, Cross-lingual language model pretraining (2019). arXiv:1901.07291.
[5] 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, 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.
[6] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4623–4637. URL: https://aclanthology.org/2020.acl-main.421. doi:10.18653/v1/2020.acl-main.421.
[7] F. Barbieri, L. E. Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, in: International Conference on Language Resources and Evaluation, 2021.
[8] L. Plaza, J. C. 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.
[9] D. Q. Nguyen, T. Vu, D. Q. Nguyen, Bertweet: A pre-trained language model for english tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), 2020.
[10] H. Pengcheng, G. Jianfeng, C. Weizhu, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, CoRR abs/2111.09543 (2021). URL: https://arxiv.org/abs/2111.09543.
[11] J. L. F. De la Rosa, G. A. A. Molina, R. M. T. SÁnchez, M. R. Pascual, J. C. G. Ramírez, Bertin: A spanish bert-based model for natural language processing tasks, in: International Conference on Computational Science, 2022, pp. 303–314.
[12] Y. A. Gutiérrez-Fandiño, J. C. Cardona, E. A. Montoya, S. A. García, Maria at iberlef 2021: Transfer learning and ensemble techniques for hate speech detection in spanish, in: Iberian Languages Evaluation Forum, 2021, pp. 278–288.
[13] Universidad Nacional de Educación a Distancia (UNED), EXIST 2023, Website, 2023. URL: http://nlp.uned.es/exist2023/, accessed: July 10, 2023.
[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: http://arxiv.org/abs/1910.03771. arXiv:1910.03771.
[15] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework (2019). arXiv:1907.10902.
[16] rkoonireddy, exist_2023, https://github.com/rkoonireddy/exist_2023, 2023.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[18] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[19] 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 (2019). arXiv:1907.11692.

[20] D. R. Pérez, A. F. Alayón, J. R. Salas, Robertuito: A spanish roberta model for sentiment analysis, in: International Conference on Natural Language Processing, 2021, pp. 424–438.

[21] 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, in: Proces. del Leng. Natural, 2021.

[22] 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.

[23] M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Álvarez-Carmona, E. Álvarez Mellado, J. Carrillo-de Albornoz, L. Chiruzzo, L. Freitas, H. Gómez Adorno, Y. Gutiérrez, S. M. Jiménez Zafra, S. Lima-López, F. M. Plaza-de Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR-WS.org, MÁlaga, Spain, 2021. URL: https://ceur-ws.org/Vol-2943/, xXXVII International Conference of the Spanish Society for Natural Language Processing.

[24] M. Montes-y Gómez, J. Gonzalo, F. Rangel, M. Casavantes, M. Á. Álvarez-Carmona, G. Bel-Enguix, H. J. Escalante, L. Freitas, A. Miranda-Escalada, F. Rodríguez-SÁnchez, A. RosÁ, M. A. Sobrevilla-Cabezudo, M. Taulé, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR-WS.org, A Coruña, Spain, 2022. URL: https://ceur-ws.org/Vol-3202/, xXXVIII International Conference of the Spanish Society for Natural Language Processing.

[25] S. Han, Googletrans, https://github.com/ssut/py-googletrans, 2021.

[26] sdadas, xlm-roberta-large-twitter, https://huggingface.co/sdadas/xlm-roberta-large-twitter, 2023.

[27] PyTorch, torch.nn.crossentropyloss, PyTorch Documentation, Accessed July 6, 2023. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html.

[28] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

A. Online Resources

The source code and the final submission files can be accessed through the following official GitHub repository for EXIST2023:

- GitHub for EXIST2023 roh-neil

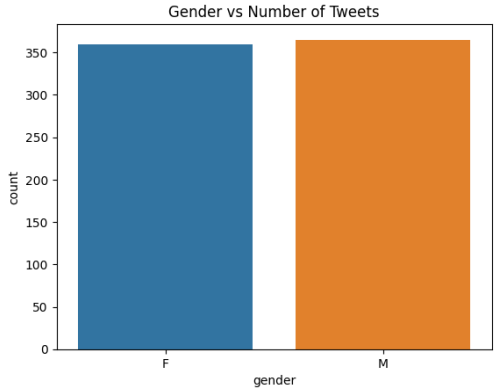
B. Additional Information and Illustrations

Table 8
Examples of task 1 classification

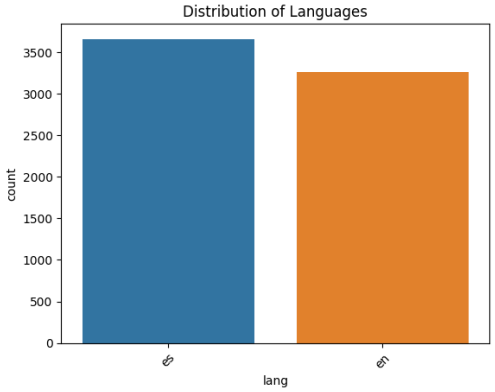
Tweet	soft-labels	hard-label
"Mientras Scooby disfruta unas vacaciones en Lamanada yo cuido en casa a Merlot, el russian bleu de mi hija @victoria_pujals https://t.co/K1ozE7iCwA"	'NO': 1.0, 'YES': 0.0	NO
"Lo de Lopetegui en rueda de prensa diciendo no se que de la minifalda y tal... Era por los de #LaManada... No?"	'YES': 0.83, 'NO': 0.17	YES

Table 9
Examples of task 2 classification

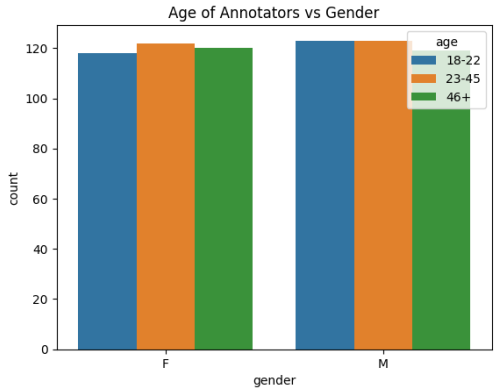
Tweet	soft-labels	Hard-labels
"@Pretorian_MGTOW No se que le pasó a ésta mujer? Está confundida o le pagaron en sobre?"	'OBJECTIFICATION': 0.33 , 'SEXUAL-VIOLENCE': 0.17 , 'STEREOTYPING-DOMINANCE': 0.33 , 'MISOGYNY-NON-SEXUAL-VIOLENCE': 0.33 , 'IDEOLOGICAL-INEQUALITY': 0.0, 'NO': 0.0	DIRECT
@Dan_Brocklebank Wonder if he was able to make the time to get that slapper round the back of the building and give her a ride ?? @MattHancock"	'OBJECTIFICATION': 0.5, 'NO': 0.17, 'MISOGYNY-NON-SEXUAL-VIOLENCE': 0.17, 'STEREOTYPING-DOMINANCE': 0.17, 'IDEOLOGICAL-INEQUALITY': 0.0, 'SEXUAL-VIOLENCE': 0.0	JUDGEMENTAL



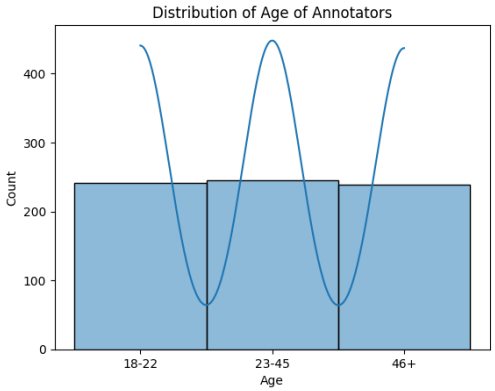
(a) Gender distribution



(b) Language distribution



(c) Age and Gender distribution



(d) Annotators' age distribution

Figure 1: Exploratory Data Analysis of Training Data

Table 10
Examples of task 3 classification

Tweet	soft-labels	hard-labels
"@Pretorian_MGTOW No se que le pasó a ésta mujer? Está confundida o le pagaron en sobre?"	'OBJECTIFICATION': 0.33, 'SEXUAL-VIOLENCE': 0.17, 'STEREOTYPING-DOMINANCE': 0.33, 'MISOGYNY-NON-SEXUAL-VIOLENCE': 0.33, 'IDEOLOGICAL-INEQUALITY': 0.0, 'NO': 0.0	['OBJECTIFICATION', 'STEREOTYPING-DOMINANCE', 'MISOGYNY-NON-SEXUAL-VIOLENCE']
"@LibertadSurja Es mucho peor, los veganos rechazan los productos animales por moral, por no hacer un da o a seres que consideran, que hay que protege. Los MGTOW rechazan a las mujeres por resentimiento, odio y paranoia."	'IDEOLOGICAL-INEQUALITY': 0.17, 'OBJECTIFICATION': 0.33, 'MISOGYNY-NON-SEXUAL-VIOLENCE': 0.67, 'NO': 0.33, 'SEXUAL-VIOLENCE': 0.17, 'STEREOTYPING-DOMINANCE': 0.0	['OBJECTIFICATION', 'MISOGYNY-NON-SEXUAL-VIOLENCE']

Table 11
EXIST previous Datasets distribution

	Training								Testing					
	Twitter				Gab				Twitter					
	Spanish		English		Spanish		English		Spanish		English		Total	
	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022	2021	2022
Sexist	1741	2599	1636	2494	858	265	265	300	858	254	300	215	5658	6127
Non-sexist	1800	2612	1800	2658	812	225	225	192	858	271	192	305	5687	6263
Ideological-inequality	480	695	386	619	215	73	73	100	233	97	100	64	1487	1648
Misogyny-non-sexual-violence	401	600	284	436	199	58	58	63	152	32	63	25	1157	1214
Objectification	244	368	256	377	124	50	50	29	121	18	29	21	824	863
Sexual-violence	173	304	344	494	131	71	71	48	150	44	48	43	917	1004
Stereotyping-dominance	443	632	366	568	189	13	13	60	202	60	60	55	1273	1388

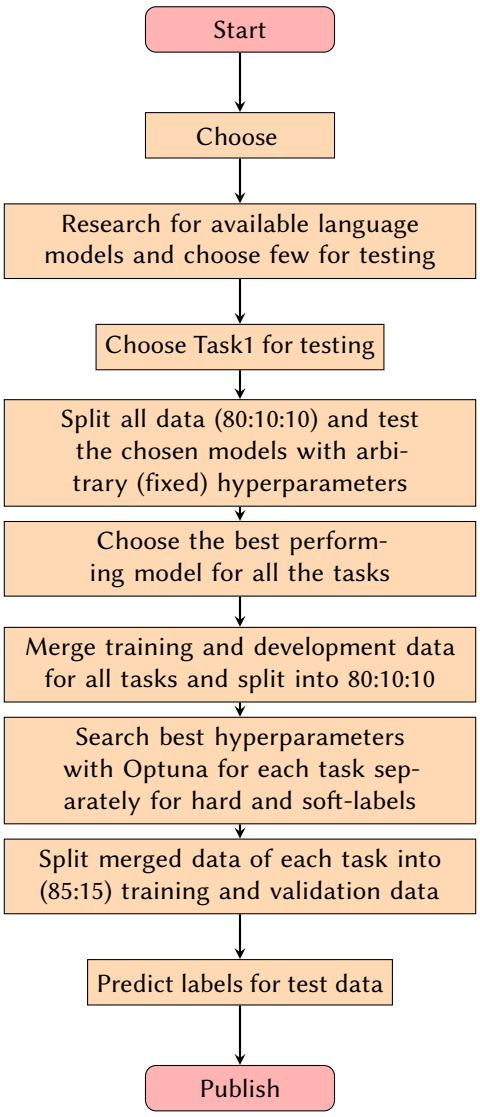


Figure 2: Flowchart for Experiment Setup