

# AIT\_FHSTP at EXIST 2023 Benchmark: Sexism Detection by Transfer Learning, Sentiment and Toxicity Embeddings and Hand-Crafted Features<sup>\*</sup>

Notebook for the EXIST2023 Lab at CLEF 2023

Jaqueline Böck<sup>2,\*</sup>, Mina Schütz<sup>1</sup>, Daria Liakhovets<sup>1</sup>, Nathanya Queby Satriani<sup>2</sup>, Andreas Babic<sup>2</sup>, Djordje Slijepčević<sup>2</sup>, Matthias Zeppelzauer<sup>2</sup> and Alexander Schindler<sup>1</sup>

<sup>1</sup>Austrian Institute of Technology, Giefinggasse 4, 1210 Vienna, Austria

<sup>2</sup>St. Pölten University of Applied Sciences, 3100 St. Pölten, Austria

## Abstract

Sexism has become a widespread problem on social media and in online conversations. Therefore, the sEXism Identification in Social neTworks (EXIST) challenge addresses this issue at CLEF in 2023. In this year's version of this international benchmark, the goal is to automatically identify sexism in texts with the help of Natural Language Processing (NLP). The tasks are to determine whether a text is sexist, what the source intention behind it is and which type of sexist category it belongs to. This paper presents the contribution of our team, AIT\_FHSTP, in the EXIST challenge held at CLEF in 2023. We present three approaches to solve the classification tasks of this year's shared task. The baseline for all three approaches is an XLM-RoBERTa model pre-trained with additional datasets and fine-tuned on the EXIST2023 data. For our second and third approach we extracted the fine-tuned embeddings of the model and concatenated them with additional features. On the one hand we added sentiment and toxicity model embeddings and on the other hand we added multiple hand-crafted features and reduced the dimensionality with PCA. Afterwards we used these embeddings as an input for a Random Forest classifier who generated the final predictions. Our approach combining XLM-RoBERTa embeddings with additional crafted features and PCA achieved the 1<sup>st</sup> rank on the soft-soft evaluation of task 2 (source intention) with Spanish content and the 2<sup>nd</sup> rank for English content. For task 3 (sexism multilabel categorization), we achieved the 3<sup>rd</sup> rank in the hard-hard evaluation.

## Keywords

Sexism detection, Sexism identification, Social Media Retrieval, Transformer Models, Natural Language Processing

## 1. Introduction

In recent years - through the rise of social networks and media - discriminatory views and statements have been a common phenomenon, especially against women. This relates to other

---


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

\*Corresponding author.

✉ Jaqueline.Boeck@fhstp.ac.at (J. Böck); Mina.Schuetz@ait.ac.at (M. Schütz); Daria.Liakhovets@ait.ac.at (D. Liakhovets); Nathanya.Satriani@fhstp.ac.at (N. Q. Satriani); andreas.babic@fhstp.ac.at (A. Babic); Djordje.Slijepcevic@fhstp.ac.at (D. Slijepčević); Matthias.Zeppelzauer@fhstp.ac.at (M. Zeppelzauer); Alexander.Schindler@ait.ac.at (A. Schindler)



© 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)

domains, such as hate speech [1] and disinformation [2]. Therefore, the shared task on sEXism Identification in Social neTworks (EXIST) at CLEF 2023 [3, 4] tackles the issue of automatic sexism detection in social media texts. Task 1 did not change in the third iteration of the EXIST challenge and is still a binary classification task where each text is annotated either as *sexist* or *not-sexist*. The second task is dedicated to source intention. This specifies whether a sexist comment was *direct*, *reported*, or *judgmental*. “Direct” describes whether the comment is simply sexist or tries to be. The label “reported” states, whether a sexist situation is reported in the text and is not sexist by itself. Lastly, “judgmental” means that: “[...] intention was to judge, since the tweet describes sexist situations or behaviours with the aim of condemning them.” (see: <http://nlp.uned.es/exist2023/>). The third task refers to a multi-label classification into different types of sexist content: *ideological-inequality*, *objectification*, *stereotyping-dominance*, *misogyny-non-sexual-violence*, *sexual-violence*, *non-sexist*.

This paper presents our contribution to the benchmark, describes our approach, and summarizes the obtained results for all three tasks, i.e., the binary sexism identification task (task 1), the source intention task (task 2), and sexism categorization (task 3). To account for the bilingual dataset we employ the multilingual model XLM-RoBERTa [5] as a baseline representation. The XLM-RoBERTa model was pre-trained in an unsupervised manner on 10 million tweets and additional sexism related datasets and fine-tuned on the EXIST2023 data. The methodical approach includes generating additional embeddings for task 1 and 2 using our custom pre-trained and fine-tuned XLM-RoBERTa model. Furthermore, we utilized pre-existing, task-specific models from HuggingFace [6] to derive additional sentiment and toxicity embeddings. Besides a simple baseline approach in which we fine-tuned this pre-trained XLM-RoBERTa model on the EXIST2023 data, we also investigated approaches in which we used the XLM-RoBERTa embeddings and combined them with the sentiment and toxicity embeddings as well as with various hand-crafted features by utilizing a Random Forest [7] as classifier. Experiments with and without dimensionality reduction via Principal Component Analysis (PCA) [8] have been performed.

Our paper is structured as follows: Section 2 describes our methodological approach with a focus on the employed datasets and models. The presentation of the results is presented in Section 3), which is followed by the the discussion and final conclusions (Section 4).

## 2. Methodology and Evaluation

In this paper, the term “pre-training” refers to the unsupervised re-training of a model. The term “fine-tuning” refers to the supervised training on the downstream tasks of the challenge. Our methodological approach is based on the EXIST2023 dataset generating embeddings for three different use cases. We generated embeddings with an own pre-trained and fine-tuned XLM-RoBERTa (XLM-R) model [5] for the downstream tasks of this year’s competition. Additionally, already pre- and fine-tuned models from HuggingFace [6] have been utilized to generate additional sentiment embeddings and toxicity embeddings from the original data. The model architecture behind these are also XLM-RoBERTa models trained on the respective tasks. In this paper, the sentiment XLM-RoBERTa model is referred as XLM-R-SENT [9] and the toxicity model as XLM-R-TOXI [10]. Furthermore, we employed hand-crafted features including the

number of extracted hashtags and links per text, the word and emoji count, punctuation-, exclamation-, and question marks, as well as the ratio of those. The generated embeddings and features were then used as an input to train a Random Forest [7] classifier. Optionally, dimensionality reduction of the input embeddings with Principal Component Analysis (PCA) [8] was performed.

Following the strategy of the past two years of contributing in the challenge [11] [12], we also trained a simple XLM-R model with the original data for generating a baseline. More detailed information on the exact training strategies can be found in the following.

## 2.1. EXIST2023 Data

In an attempt to first explore the given dataset, we investigated the *EXIST2023* data for potential duplicates, compared it further with the previous years' data, and searched for possible relations between the hashtags and the labels of the tweets. Based on the inspection, we then determined that common data preprocessing techniques would suffice for the tweets to be used as inputs for the models.

The dataset includes postings from social media platforms such as Twitter and Gab, as well as annotations for different categories of sexism which were then split into training, development, and test partitions. The training set consists of 6,920 instances in English (3,260) and Spanish (3,660) while the development set includes 1,038 samples and the test set contains 2,076 samples. Each data instance is assigned a binary label (for task 1) indicating whether it is sexist (*yes*) or non-sexist (*no*). In addition, a ternary classification assignment is provided for task 2: *direct*, *reported*, *judgmental*, and a multi-label categorization is the target of task 3, i.e., *ideological-inequality*, *objectification*, *stereotyping-dominance*, *misogyny-non-sexual-violence*, *sexual-violence*, *non-sexist*.

## 2.2. External Data

To further pre-train the chosen transformer model (XLM-R) we used additional datasets - as in our last year's approach [12] for the EXIST2022 shared task [13]. Apart from utilizing the EXIST2022 dataset, we incorporated additional datasets specifically created for analogous classification tasks (refer to the following list for more details). We utilized additionally unlabeled tweets during the pre-training process as we did in 2022. By doing so, our aim was to establish a degree of comparability between our approach for EXIST2023 and the previous iteration. The unlabeled tweets were extracted via the official Twitter API - specifically from the full COVID-19 stream, which was made openly accessible in 2020 due to the pandemic (<https://developer.twitter.com/en/docs/twitter-api/tweets/covid-19-stream/overview>). We filtered the Twitter stream with hashtags - that contain sexism related content - present in the data of the EXIST2022 challenge. This resulted in around 40 million tweets in English and Spanish, which was randomly sampled to a total amount of 10,475,215 for pre-training. The following list describes the external datasets we used for pre-training:

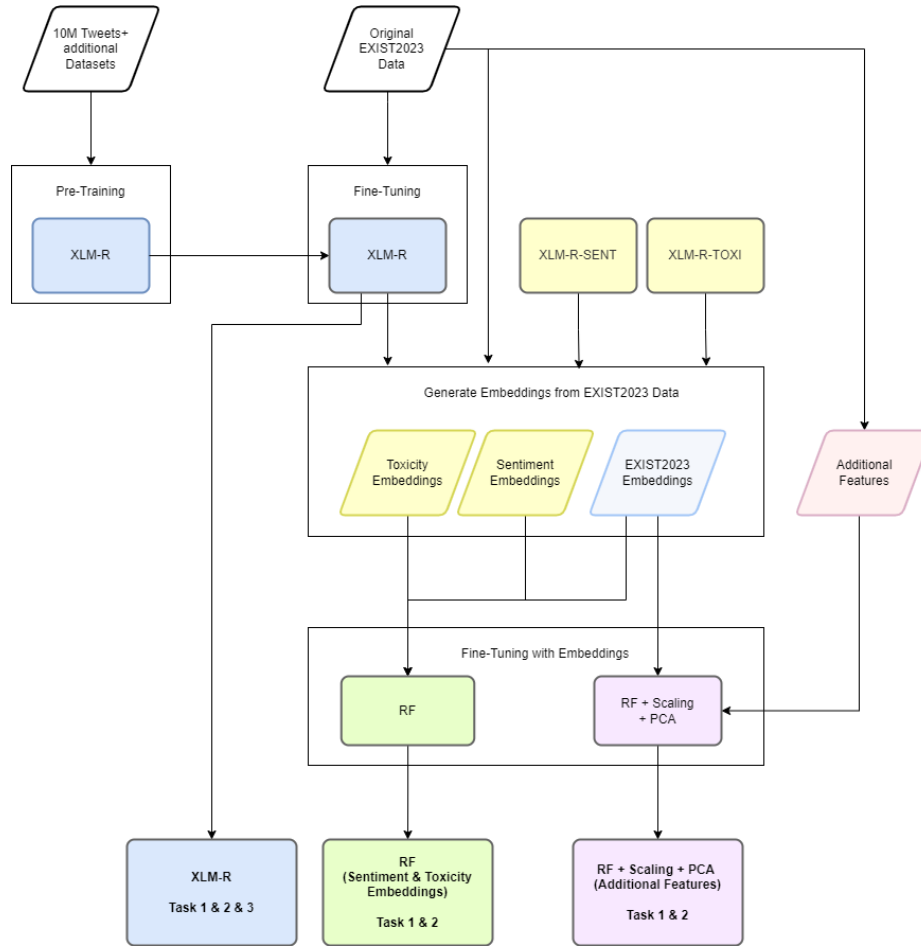
- **SOT** [14] is a dataset that contains Spanish comments from the Twitch network. The comments were extracted from user profiles of female Twitch streamers, filtered based on topic and amount of followers. Each comment was first classified into *innocuous*

or *inappropriate*, and afterwards into *love-stuck* and *strongly sexist*. Approximately 300 comments have been classified as inappropriate, which consist of 50% *love-stuck* and 50% *strongly-sexist*. The other 3,000 comments were classified as innocuous.

- **CMSB [15]** is an English dataset that contains a total amount of 13,634 texts combining social media posts (tweets), psychological survey items and synthetic adversarial modifications of both. The social media posts were aggregated from three published datasets: *the hostile sexism dataset* [16], *the benevolent sexism dataset* [17] and the *call me sexism dataset* [18].
- **SCB [19]** is a Spanish dataset about violent and misogynistic content retrieved from Twitter. It contains a total of 5,520 instances.
- **MeTwo [20]** is a Spanish dataset containing Twitter comments regarding sexist innuendo, behaviors, and expressions labeled with the following terms: *SEXIST*, *NON\_SEXIST* and *DOUBTFUL*. The dataset consists of 3,600 comments in total.
- **SSC [21]** is an English dataset with a total of 6,238 comments retrieved from Instagram and annotated with the labels *sexist* and *non-sexist*. The authors filtered comments based on hashtags, such as: "bloodymen", "boys", "everydaysexism", "girls", "guys", "manspaling", "metoo", "sexism", "sexist" and "slutshaming".
- **ISEP [22]** is an English dataset containing workplace-related sexist comments. The authors differentiate between unambiguous sexist comments and ambiguous or neutral cases. The dataset contains 1,100 comments in total.
- **MTM [23]** is an English dataset that contains definitions from Urban dictionary. The - in total 2,285 - definitions were labeled as *misogynistic* and *non-misogynistic*.
- **HatEval2019 [24]** is an English and Spanish dataset containing 13,000 English and 6,000 Spanish tweets. The main theme of the comments are hate speech against women and immigrants. The tweets were annotated into three categories: 1) *Hate Speech* (against women or immigrants), 2) *Target Range* (against a generic group or individual), and 3) *Aggressiveness*.

### 2.3. Overview of Approaches

In Figure 1 a graphical overview of our approaches including the different training strategies are displayed. The focus of our setup is on generating embeddings based on the EXIST2023 data using different XLM-RoBERTa models and training a Random Forest on these embeddings. This includes embeddings generated by an pre-trained and fine-tuned XLM-R model, a sentiment detection XLM-R-SENT model [9], a toxicity detection XLM-R-TOXI model [10], and additional hand-crafted features.



**Figure 1:** An overview of the various strategies implemented for the three distinct approaches (highlighted in blue, green, and violet). All approaches are built upon the XML-R model that was pre-trained in an unsupervised manner on external datasets and fine-tuned on the EXIST 2023 data. The first approach (blue) leverages the predictions of XML-R. The second approach (green) incorporates sentiment and toxicity embeddings in addition to the XML-R embedding. The third approach (violet) employs hand-crafted features along with the XML-R embeddings.

## 2.4. Unsupervised Pre-Training of XLM-R

We trained the XLM-R model on the Masked Language Modeling (MLM) task with a probability of 0.15. This means that 15% of the text gets masked randomly during training. The masked out tokens have to be predicted during pre-training. The model was trained for five epochs with a batch size of 16 and a learning rate of  $2e^{-5}$ . The model was trained on the EXIST2022 data and the additional datasets (CMSB, HateEval, ISEP, MeTwo, MTM, SCB, SOT, and SSC), but also on a large dataset of tweets we sampled from the Covid-19 stream. To reduce the amount of training time, we only used 10 million randomly selected samples. This resulted in an overall training time of about seven days.

## 2.5. Supervised Fine-Tuning of XLM-R

For training the baselines for task 1, 2 and 3, we fine-tuned our pre-trained XLM-R on the original EXIST 2023 dataset for three epochs using a learning rate of  $3e^{-5}$ , linear scheduler with 500 warm-up steps, weight decay of 0.01 and a batch size of eight. We trained a binary classification model for task 1, multi-class classification model for task 2 and multi-label classification model for task 3. We obtained hard labels with a predicted probability threshold of 0.5. We used the CLS-token from the last network layers to extract text embeddings for the combined approaches.

## 2.6. Sentiment and Toxicity Features

To enrich the baseline representation (see above), we incorporated learned embeddings from sentiment [9] and toxicity classification models [10, 25] in our second approach. First, we ran a few experiments on different pre-trained and fine-tuned sentiment and toxicity classification models from HuggingFace. Then, the embeddings of the best-performing models were computed for the training, development, as well as the test set. These embeddings were then concatenated along with the XLM-R embeddings for the final submissions of our second approach.

## 2.7. Additional Hand-Crafted Features

For our third approach we extracted additional hand-crafted features and followed a similar approach as in [26]. The authors concatenated embeddings extracted from fine-tuned language models with additional hand-crafted features and used a Multi-Layer Perceptron as classifier. Although this approach was originally targeted at the detection of toxic content, we transferred it to sexism detection, since we believe those domains are similar. The hand-crafted features include:

- *Hashtags*: Number of extracted hashtags per text.
- *Links*: Number of extracted links per text.
- *Word Count*: Total number of words.
- *Punctuation marks*: Total number of punctuation marks.
- *Exclamation marks*: Total number of exclamation marks.
- *Question marks*: Total number of question marks.
- *Word Punctuation Ratio*: The ratio of punctuation in relation to the number of words.
- *Word Exclamation Ratio*: The ratio of exclamation marks in relation to the number of words.
- *Word Question Ratio*: The ratio of question marks in relation to the number of words.
- *Emoji Count*: The number of emojis found in each text. The emojis were extracted via the emojis library.
- *Emoji Ratio*: The number of emojis in relation to the number words.

## 2.8. Supervised Training of Random Forests

For our second approach (XLM-R\_senttox), we trained the Random Forest on the embeddings generated by our pre- and fine-tuned XLM-R model as well as with the additional sentiment and toxicity embeddings described in section 2.6.

For our third approach (XLM-R\_craft), instead of utilizing sentiment and toxicity embeddings, we trained the Random Forest model using the additional hand-crafted features described in Section 2.7. Here, the embeddings and the additional features were compressed with PCA [8]. The selection of the number of components was based on a threshold of 95% (preserved variance), ensuring that the cumulative explained variance surpasses the specified threshold. We performed a grid-search to determine the best values for the number and the maximum depth of the trees. For the final predictions we defined the RFs with 500 trees with a maximum depth of 15 in both approaches.

### 3. Results

This section documents every result for all tasks and evaluation types for the proposed approaches on this year’s EXIST shared task. The shared task describes multiple classification levels. After categorizing each text as sexist or not (task 1) the sexist comments have several sub-levels (task 2 and 3). For task 2 it can only be one of the three labels and for task 3 the labels are not mutually exclusive. The standard evaluation metrics for classification tasks would be accuracy, precision, recall and the f1-score. However, for this year’s challenge the authors introduced new evaluation metrics. Furthermore there are so-called "hard" and "soft" labels. In the “hard” ground truth the final annotations are made via a gold standard such as majority voting. In the “soft” ground truth the variability of annotations by different annotators is taken into account. In the hard setting for each text a final label has to be predicted, in comparison to the soft setting, where the output has to be a probability for each label. Therefore, the evaluation combinations by the authors are defined as: hard-hard, hard-soft, and soft-soft. The first part of the terms relate to the system output; the second to the ground truth, e.g., for hard-soft the hard predictions are evaluated with the soft ground truth. The proposed metrics by the organizers are ICM (Information Contrast Measure) [27] and ICM-soft. The ICM measures the similarity of the predicted labels to the original ground truth categories. The ICM-soft is an extension of the ICM measure by the EXIST 2023 organizers to help with the hierarchical multilabel classification issue for soft outputs and ground truths [4]. For hard-hard the official metric is ICM, for hard-soft the official metric is ICM-soft, and for soft-soft the official metric is also ICM-soft. This leads to three approaches submitted to the shared task:

1. **Approach 1: XLM-R\_only (task 1, task 2, task 3):** We fine-tuned the XLM-R on the validation and dev data from the EXIST 2023 in a supervised end-to-end approach for the binary classification task predicting the class labels (i.e., one output node for each class). This approach obtained only hard predictions.
2. **Approach 2: XLM-R\_senttox (task 1, task 2):** We used the embeddings from the model trained in XLM-R\_only and additionally we used embeddings from a sentiment analysis and a toxicity analysis model and trained a random forest on these embeddings. This approach obtained hard and soft predictions. For the soft predictions we used the prediction probabilities of the Random Forest.
3. **Approach 3: XLM-R\_craft (task 1, task 2):** We used the embeddings from the model trained in XLM-R\_only, performed a PCA on these embeddings and trained a Random

**Table 1**

Results of our approach for task 1 (sexism detection). Each sub-task is shown for both languages and split by language. Cross-entropy (Cross Ent.) was not provided for hard-soft evaluation (noted with an "X"). Runs without results are noted with "-".

Task	Lang	Run	Approach	ICM-Soft	ICM-Soft Norm	Cross Ent.	Rank
1 Soft-Soft	ALL	1	XLM-R_senttox	0.5648	0.5875	1.1491	20 <sup>th</sup>
1 Soft-Soft	ALL	2	XLM-R_only	-	-	-	-
1 Soft-Soft	ALL	3	XLM-R_craft	0.5955	0.5875	0.9392	19 <sup>th</sup>
1 Soft-Soft	ES	1	XLM-R_senttox	0.6512	0.5667	1.1177	19 <sup>th</sup>
1 Soft-Soft	ES	2	XLM-R_only	-	-	-	-
1 Soft-Soft	ES	3	XLM-R_craft	0.6446	0.5655	0.9178	20 <sup>th</sup>
1 Soft-Soft	EN	1	XLM-R_senttox	0.4045	0.609	1.1844	17 <sup>th</sup>
1 Soft-Soft	EN	2	XLM-R_only	-	-	-	-
<b>1 Soft-Soft</b>	<b>EN</b>	<b>3</b>	<b>XLM-R_craft</b>	<b>0.4887</b>	<b>0.6211</b>	<b>0.9632</b>	<b>14<sup>th</sup></b>
1 Hard-Hard	ALL	1	XLM-R_senttox	0.485	0.6751	0.755	33 <sup>rd</sup>
<b>1 Hard-Hard</b>	<b>ALL</b>	<b>2</b>	<b>XLM-R_only</b>	<b>0.5086</b>	<b>0.6901</b>	<b>0.7571</b>	<b>23<sup>rd</sup></b>
1 Hard-Hard	ALL	3	XLM-R_craft	0.4832	0.6739	0.7544	34 <sup>th</sup>
1 Hard-Hard	ES	1	XLM-R_senttox	0.4801	0.6559	0.771	31 <sup>st</sup>
1 Hard-Hard	ES	2	XLM-R_only	0.5015	0.67	0.7709	25 <sup>th</sup>
1 Hard-Hard	ES	3	XLM-R_craft	0.4829	0.6577	0.7721	30 <sup>th</sup>
1 Hard-Hard	EN	1	XLM-R_senttox	0.4769	0.6942	0.7338	33 <sup>rd</sup>
1 Hard-Hard	EN	2	XLM-R_only	0.5033	0.7102	0.7387	27 <sup>th</sup>
1 Hard-Hard	EN	3	XLM-R_craft	0.4695	0.6897	0.7338	35 <sup>th</sup>
1 Hard-Soft	ALL	1	XLM-R_senttox	0.10.14	0.5126	X	33 <sup>rd</sup>
1 Hard-Soft	ALL	2	XLM-R_only	0.1411	0.519	X	27 <sup>th</sup>
1 Hard-Soft	ALL	3	XLM-R_craft	0.0932	0.5113	X	34 <sup>th</sup>
1 Hard-Soft	ES	1	XLM-R_senttox	0.2478	0.4958	X	29 <sup>th</sup>
<b>1 Hard-Soft</b>	<b>ES</b>	<b>2</b>	<b>XLM-R_only</b>	<b>0.277</b>	<b>0.5009</b>	<b>X</b>	<b>25<sup>th</sup></b>
1 Hard-Soft	ES	3	XLM-R_craft	0.2388	0.4942	X	30 <sup>th</sup>
1 Hard-Soft	EN	1	XLM-R_senttox	-0.15	0.529	X	32 <sup>th</sup>
1 Hard-Soft	EN	2	XLM-R_only	-0.0984	0.5364	X	27 <sup>th</sup>
1 Hard-Soft	EN	3	XLM-R_craft	-0.1565	0.528	X	34 <sup>th</sup>

Forest. This approach obtained hard and soft predictions. For the soft predictions we used the prediction probabilities of the Random Forest.

In general in our experiments we found that our models performed best on task 2, the multiclass classification, as well as the Spanish texts overall. For the soft-soft versions our XLM-R\_craft model scored the best results. For the hard-hard evaluation our best model was - for task 1 as well as task 2 - the XLM-R\_only. For task 3 we only submitted the XLM-R\_only model which scored 3<sup>rd</sup> place for the ALL hard-hard evaluation.

In Table 1 the results for task 1 are shown, where our models performed best for soft-soft predictions and much worse for hard-hard and hard-soft predictions. However, for the soft-soft evaluation, the XLM-R\_craft model performs best with ranking 17<sup>th</sup>. On the other hand, in task 2 (Table 2) we performed significantly better. Results show that our models had problems



**Table 2**

Results of our approach for task 2 (sexism detection). Each sub-task is shown for both languages and split by language. Cross-entropy (Cross Ent.) was not provided for hard-soft evaluation (noted with an "X"). Runs without results are noted with "-".

Task	Lang	Run	Approach	ICM-Soft	ICM-Soft Norm	Cross Ent.	Rank
2 Soft-Soft	ALL	1	XLM-R_only	-	-	-	-
2 Soft-Soft	ALL	2	XLM-R_craft	-1.435	0.8049	1.6486	2 <sup>nd</sup>
2 Soft-Soft	ALL	3	XLM-R_senttox	-2.1619	0.7863	2.0897	7 <sup>th</sup>
2 Soft-Soft	ES	1	XLM-R_only	-	-	-	-
<b>2 Soft-Soft</b>	<b>ES</b>	<b>2</b>	<b>XLM-R_craft</b>	<b>-1.2317</b>	<b>0.7861</b>	<b>1.6415</b>	1 <sup>st</sup>
2 Soft-Soft	ES	3	XLM-R_senttox	-1.8631	0.781	2.061	7 <sup>th</sup>
2 Soft-Soft	EN	1	XLM-R_only	-	-	-	-
2 Soft-Soft	EN	2	XLM-R_craft	-1.7985	0.8264	1.6566	2 <sup>nd</sup>
2 Soft-Soft	EN	3	XLM-R_senttox	-2.7485	0.8056	2.1219	8 <sup>th</sup>
2 Hard-Hard	ALL	1	XLM-R_only	0.2229	0.7198	0.5029	6 <sup>th</sup>
2 Hard-Hard	ALL	2	XLM-R_craft	0.1475	0.7037	0.4759	13 <sup>th</sup>
2 Hard-Hard	ALL	3	XLM-R_senttox	0.1662	0.7077	0.4911	12 <sup>th</sup>
<b>2 Hard-Hard</b>	<b>ES</b>	<b>1</b>	<b>XLM-R_only</b>	<b>0.2948</b>	<b>0.7123</b>	<b>0.5414</b>	5 <sup>th</sup>
2 Hard-Hard	ES	2	XLM-R_craft	0.1647	0.6837	0.5007	13 <sup>th</sup>
2 Hard-Hard	ES	3	XLM-R_senttox	0.1944	0.6902	0.5148	11 <sup>th</sup>
2 Hard-Hard	EN	1	XLM-R_only	0.1206	0.7307	0.4488	9 <sup>th</sup>
2 Hard-Hard	EN	2	XLM-R_craft	0.1148	0.7295	0.4412	10 <sup>th</sup>
2 Hard-Hard	EN	3	XLM-R_senttox	0.1158	0.7297	0.4568	9 <sup>th</sup>
2 Hard-Soft	ALL	1	XLM-R_only	-6.8143	0.6675	X	15 <sup>th</sup>
2 Hard-Soft	ALL	2	XLM-R_craft	-6.3494	0.6794	X	11 <sup>th</sup>
2 Hard-Soft	ALL	3	XLM-R_senttox	-6.735	0.6696	X	14 <sup>th</sup>
2 Hard-Soft	ES	1	XLM-R_only	-6.161	0.6451	X	15 <sup>th</sup>
<b>2 Hard-Soft</b>	<b>ES</b>	<b>2</b>	<b>XLM-R_craft</b>	<b>-5.6617</b>	<b>0.6594</b>	<b>X</b>	9 <sup>th</sup>
2 Hard-Soft	ES	3	XLM-R_senttox	-6.0034	0.6496	X	13 <sup>th</sup>
2 Hard-Soft	EN	1	XLM-R_only	-8.1348	0.6875	X	15 <sup>th</sup>
2 Hard-Soft	EN	2	XLM-R_craft	-7.6509	0.6981	X	12 <sup>th</sup>
2 Hard-Soft	EN	3	XLM-R_senttox	-8.0912	0.6885	X	14 <sup>th</sup>

in classifying the non-sexist samples from task 1. For task 2 the models performed especially well for the Spanish texts in combination with the additional hand-crafted features. In Table 3 the results for task 3 are shown. For task 3 we only submitted the XLM-R\_only model which scored 3<sup>rd</sup> place for the English hard-hard evaluation.

## 4. Discussion & Conclusion

In this paper, we provided the details on our submission to the EXIST 2023 benchmark, which consists of three tasks on the classification of sexist content. We presented three approaches using an XLM-R model as a baseline. We pre-trained the already available XLM-R model with 10 million tweets containing hashtags from last year's shared task data (EXIST 2022) as well

**Table 3**

Results of our approach for task 3 (sexism detection). Each sub-task is shown for both languages and split by language. Cross-entropy (Cross Ent.) was not provided for hard-soft evaluation (noted with an "X"). Runs without results are noted with "-".

Task	Lang	Run	Approach	ICM-Soft	ICM-Soft Norm	Cross Ent.	Rank
3 Soft-Soft	ALL	1	XLM-R_only	-	-	-	-
3 Soft-Soft	ES	1	XLM-R_only	-	-	-	-
3 Soft-Soft	EN	1	XLM-R_only	-	-	-	-
3 Hard-Hard	ALL	1	XLM-R_only	0.2366	0.6372	0.5842	3 <sup>rd</sup>
<b>3 Hard-Hard</b>	<b>ES</b>	<b>1</b>	<b>XLM-R_only</b>	<b>0.2681</b>	<b>0.6454</b>	<b>0.5995</b>	3 <sup>rd</sup>
3 Hard-Hard	EN	1	XLM-R_only	0.1837	0.6263	0.5609	4 <sup>th</sup>
3 Hard-Soft	ALL	1	XLM-R_only	-13.6923	0.5833	X	16 <sup>th</sup>
<b>3 Hard-Soft</b>	<b>ES</b>	<b>1</b>	<b>XLM-R_only</b>	<b>-12.4109</b>	<b>0.5999</b>	<b>X</b>	16 <sup>th</sup>
3 Hard-Soft	EN	1	XLM-R_only	-15.2332	0.5656	X	20 <sup>th</sup>

as with additional annotated datasets related to the topic. We then fine-tuned this XLM-R model on the EXIST 2023 data to achieve baselines for tasks 1-3. To enhance our baseline we integrated several additional features for our second and third approach (targeted embeddings for sentiment and toxicity representation as well as hand-crafted features) and trained a Random Forest on top of them to obtain final hard and soft labels.

Results show that our approaches performed best on task 2, in particular for the Spanish content. For the soft-soft evaluation the XLM-R model with hand-crafted features scored best. For Spanish content we achieved the best results among all participants. For the hard-hard evaluation the best approach was the XLM-R\_only for task 1 as well as for task 2. For task 3 we only submitted the XLM-R baseline which scored 3<sup>rd</sup> across all participants. In the future, we intend to investigate the combination of our findings from EXIST 2022, which involved data augmentation, using translations, and additional datasets for fine-tuning, with the approach employed in the current year.

## Acknowledgments

This work is enhanced by the Austrian Institute of Technology GmbH (AIT) funded by the FFG project "RAIDAR" (grant no. 886364, Austrian security research program KIRAS of the Federal Ministry of Finance (BMF)). This project further received funding by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20016] at St. Pölten University of Applied Sciences.

## References

- [1] C. Demus, J. Pitz, M. Schütz, N. Probol, M. Siegel, D. Labudde, Detox: A comprehensive dataset for german offensive language and conversation analysis, in: Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022), Association for Computational Linguistics, Online, 2022, pp. 54–61.

- [2] M. Schütz, A. Schindler, M. Siegel, K. Nazemi, Automatic fake news detection with pre-trained transformer models, in: A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges*, Springer International Publishing, Cham, 2021, pp. 627–641.
- [3] 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, in: A. Arampatzis, E. Kanoulas, T. Tsirikika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, CLEF Association, Thessaloniki, Greece, 2023.
- [4] 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.
- [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, *CoRR abs/1911.02116* (2019). URL: <http://arxiv.org/abs/1911.02116>. arXiv:1911.02116.
- [6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [7] L. Breiman, Random forests, *Machine Learning* 45 (2001) 5–32. doi:10.1023/A:1010950718922.
- [8] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), *Computers & Geosciences* 19 (1993) 303–342. URL: <https://www.sciencedirect.com/science/article/pii/009830049390090R>. doi:[https://doi.org/10.1016/0098-3004\(93\)90090-R](https://doi.org/10.1016/0098-3004(93)90090-R).
- [9] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, in: *Proceedings of the Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: <https://aclanthology.org/2022.lrec-1.27>.
- [10] F. Barbieri, L. A. Espinosa, J. Camacho-Collados, Detoxify, Github. <https://github.com/unitaryai/detoxify>, 2020.
- [11] M. Schütz, J. Boeck, D. Liakhovets, D. Slijepcevic, A. Kirchknopf, M. Hecht, J. Bogensperger, S. Schlarb, A. Schindler, M. Zeppelzauer, Automatic sexism detection with multilingual transformer models, *CoRR abs/2106.04908* (2021). URL: <https://arxiv.org/abs/2106.04908>. arXiv:2106.04908.
- [12] D. Liakhovets, M. Schütz, J. Böck, M. Andresel, A. Kirchknopf, A. Babic, D. Slijpcevic, J. Lampert, A. Schindler, M. Zeppelzauer, Transfer learning for automatic sexism detection with multilingual transformer models, in: M. Montes-y Gómez, J. Gonzalo, F. Rangel, M. Casavantes, M. Álvarez-Carmona, G. Bel-Enguix, H. Escalante, L. Freitas, A. Miranda-Escalada, F. Rodríguez-Sánchez, A. Rosá, M. Sobrevilla-Cabezudo, M. Taulé, R. Valencia-García (Eds.), *Proceedings of the Iberian Languages Evaluation Forum (IberLef 2022)*,

volume 3202, CEUR-WS, 2022, pp. 1–13. EXIST2022 ; Conference date: 20-09-2022.

- [13] F. Rodríguez-Sánchez, J. C. 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).
- [14] D. G. Ibáñez, R. V. Puig, Sexism bot classifier on twitch: Moderating sexist comments on twitch streamings., 2020. URL: <https://github.com/VPRamon/SexismOnTwitch>, accessed: 2021-05-04.
- [15] M. Samory, I. Sen, J. Kohne, F. Floeck, C. Wagner, The 'call me sexist but' dataset (cmsb), 2021. URL: <https://doi.org/10.7802/2251>, accessed: 2022-03-09.
- [16] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: <https://aclanthology.org/N16-2013>. doi:10.18653/v1/N16-2013.
- [17] A. Jha, R. Mamidi, When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: *Proceedings of the Second Workshop on NLP and Computational Social Science*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 7–16. URL: <https://aclanthology.org/W17-2902>. doi:10.18653/v1/W17-2902.
- [18] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "unsex me here": Revisiting sexism detection using psychological scales and adversarial samples, *CoRR abs/2004.12764* (2020). URL: <https://arxiv.org/abs/2004.12764>. arXiv:2004.12764.
- [19] R. I. Medina, Sexismo en código binario: Violencia digital y política contra las mujeres en México, 2021. URL: [https://github.com/RMedina19/sexismo\\_codigo\\_binario](https://github.com/RMedina19/sexismo_codigo_binario), accessed: 2022-03-09.
- [20] F. Rodríguez-Sánchez, J. Carrillo-de-Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, *IEEE Access* 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.
- [21] A. Debnath, S. S, N. Bhakt, K. Garg, P. Parikh, Sexist stereotype classification on instagram data, 2020. URL: [https://github.com/djinn-anthrope/Sexist\\_Stereotype\\_Classification](https://github.com/djinn-anthrope/Sexist_Stereotype_Classification), accessed: 2021-05-04.
- [22] D. Grosz, P. C. Céspedes, Automatic detection of sexist statements commonly used at the workplace, *CoRR abs/2007.04181* (2020). URL: <https://arxiv.org/abs/2007.04181>. arXiv:2007.04181, <https://www.kaggle.com/datasets/dgrosz/sexist-workplace-statements>, accessed: 2022-03-09.
- [23] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, D. . Ging, Urban dictionary definitions dataset for misogyny speech detection, 2019. URL: <https://data.mendeley.com/datasets/3jfwsdkryy/3>. doi:10.17632/3jfwsdkryy.3, accessed: 2021-05-04.
- [24] 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.
- [25] T. Davidson, D. Warmesley, M. W. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, *CoRR abs/1703.04009* (2017). URL: <http://arxiv.org/abs/1703.04009>. arXiv:1703.04009.

- [26] M. Schütz, C. Demus, J. Pitz, N. Probol, M. Siegel, D. Labudde, DeTox at GermEval 2021: Toxic comment classification, in: Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments, Association for Computational Linguistics, Duesseldorf, Germany, 2021, pp. 54–61. URL: <https://aclanthology.org/2021.germeval-1.8>.
- [27] 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.