RoBEXedda: Sexism Detection in Tweets

Giacomo Aru[1,†], Nicola Emmolo[1,†], Simone Marzeddu[1,†], Andrea Piras[1,†], Jacopo Raffi[1,†] and Lucia C. Passaro[1]

[1] *University of Pisa (Università di Pisa), Largo Bruno Pontecorvo 3, 56127 Pisa PI, Italy*

Abstract

Sexism remains a pervasive issue, significantly hindering women's progress in various aspects of life. This paper focuses on online misogyny, where women face high levels of abuse and threats. The "EXIST 2024" challenge aims to detect and classify sexist content on social media. In particular, in this paper, we address the "Sexism Categorization in Tweets" task, which involves identifying sexist tweets and categorizing them into predefined categories. A dataset comprising over 10,000 tweets in English and Spanish was exploited to train Transformer-based systems with "Binary Relevance" and "Classifier Chain" architectures. This report presents an analysis of the performance of our three candidate models in relation to the EXIST 2024 challenge. It includes a detailed examination of the results obtained and a comparison with the official ranking of the challenge. As team "Medusa", we achieved second place in the competition, with three runs submitted in the soft-soft ranking. The models runs, designated "RoBEXedda", attained the fourth, fifth, and sixth positions in the "Task 3 Soft-Soft ALL" ranking.

Keywords

Sexism Characterization, EXIST 2024, CLEF 2024, Transformer, Binary Relevance, Classifier Chain

1. Introduction

Nowadays sexism, characterized by discrimination against women, has become a pervasive issue, creating substantial obstacles for women in numerous aspects of their lives, including work, family life, and personal development. This discrimination acts as a significant barrier to their progress [1]. This paper focuses on the growing concern of online misogyny. Research indicates that the online environment has long been challenging for women, as they experience higher levels of bullying, abuse, hateful language, and threats compared to men [2].

EXIST [3, 4] is a series of scientific events and shared tasks that aim to capture sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviours [5]. In fact, many facets of a woman's life may be the focus of sexist attitudes, including domestic and parenting roles, career opportunities, sexual image, and life expectations, to name a few. In EXIST 2024, the fourth edition of the sEXism Identification in Social neTworks challenge at CLEF 2024 [6], the proposed tasks were focused on detecting and classifying sexist textual messages and image memes.

Overall, the shared task comprises 5 different sub-tasks. Among them, we focus solely on the third one, "Sexism Categorization in Tweets" [5]. In this task, each tweet must be categorized into one or more of the six categories spanning from ideological inequality to sexual violence.

The Sub-task dataset, consisting of more than 10,000 tweets in English and Spanish, was used to train neural networks based on the Transformer architecture [7]. To face the task, we exploited two different architectures: "Binary Relevance"[8], which treats each label separately, and "Classifier Chain"[9], which links classifiers to improve predictions. The rationale behind this choice is twofold. On the one hand, Binary Relevance allows for a straightforward approach to multi-label classification by handling each label as an independent binary classification problem. This simplicity can lead to efficient computation and ease of implementation, making it suitable for scenarios where labels are largely uncorrelated.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[†] These authors contributed equally.

g.aru@studenti.unipi.it (G. Aru); n.emmolo@studenti.unipi.it (N. Emmolo); s.marzeddu@studenti.unipi.it (S. Marzeddu); a.piras18@studenti.unipi.it (A. Piras); j.raffi@studenti.unipi.it (J. Raffi); lucia.passaro@unipi.it (L. C. Passaro)

https://github.com/GiacomoAru (G. Aru); https://github.com/nicolaemmolo (N. Emmolo); https://github.com/SimoneMarzeddu (S. Marzeddu); https://github.com/aprs3 (A. Piras); https://github.com/JacopoRaffi/ (J. Raffi); https://github.com/luciacpassaro (L. C. Passaro)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3740/paper-88.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

On the other hand, the Classifier Chain method may enhance predictive performance by considering label dependencies. By sequentially linking classifiers, each subsequent classifier in the chain incorporates the predictions of previous classifiers as additional features. This approach captures the interdependencies among labels, which can significantly improve prediction accuracy, especially in datasets where labels exhibit strong correlations.

As for the model, we decided to start from the XML-RoBERTa models family [10], with the aim of leveraging its robust pre-training on a diverse range of languages and textual contexts. Our decision to use XML-RoBERTa, and not larger models[11], was influenced by constraints in terms of computational power. Larger models, while potentially offering higher accuracy and better performance due to their increased capacity and deeper architectures, require significantly more computational resources for training and inference. This includes the need for more powerful hardware, increased memory, and longer training times, which were beyond the scope of our available resources. By choosing XML-RoBERTa, we aimed to balance model complexity and resource efficiency. We named our model family RoBEXedda, which is derived from RoBERta, adding "EX" for EXIST, and "edda" which is a suffix in the Sardinian language meaning "tiny". The model selection process in the development of RoBEXedda models involved an initial search for the optimal pretrained transformer from the XML-RoBERTa family and identifying other hyperparameters, guiding the research using Bayesian optimisation [12].

The remainder of the paper is as follows: Section 2 presents previous works on the topic. in Section 3 are outlined the goals of our task. Section 4 explains the dataset's structure and the preprocessing techniques employed during the development process. Section 5 reports on the baselines taken into account during development. Section 6 summarises the computational resources employed during the production process. Section 7 presents an in-depth analysis of our system, highlighting the state-of-the-art approaches considered in the processes of training and model selection. The results of the development are discussed in Section 8. Finally, Section 9 is left for the conclusion and future expansions of our work.

2. Related Work

Sexism, defined as prejudice or discrimination based on gender, is a pervasive issue amplified by online platforms. Researchers have made significant progress in developing automated systems for sexism detection. These systems employ various techniques, ranging from rule-based approaches to advanced machine learning. Notably, recent work and competitions have begun exploring visual and multimodal aspects of sexism detection as well. Since the very first edition of the EXIST challenge[13], several methods have been proposed to face the task. For instance, the authors of [14] propose a system leveraging both multilingual and monolingual BERT models, translating data, and implementing ensemble strategies for the identification and classification of sexism in English and Spanish. Similarly, [15] employs a multi-task learning approach that addresses distinct tasks from a unified representation, aiming to enhance model performance by leveraging information derived from different tasks. Another notable approach by [16] combines the final four hidden states of XLM-RoBERTa with a TextCNN equipped with three kernels. This integration is designed to improve sexism detection, further incorporating abusive word lexicons to demonstrate enhanced effectiveness compared to the use of the transformer's final layer.

In the EXIST2022 challenge, the second place team [17] based their system on an ensemble of five different models for Spanish (XLM-R, RoBERTa, and three BERT models) and another five models for English (XLM-R, RoBERTa, BERT, hateBERT, and ALBERT). They also translated all English tweets to Spanish and vice versa, additionally masking randomly selected tokens to augment the data. The third-place team's system [18] combined linguistic features with state-of-the-art transformers using ensemble techniques, their most effective model being a weighted ensemble of transformers. The team that achieved first place in the EXIST2023 competition [19] employed mBERT and XLM-RoBERTa along with ensemble techniques, further solidifying that transformers remain the optimal approach for this task. Our work, contextualizes within the state of the art by utilizing both Binary Relevance and Classifier Chain architectures alongside the XML-RoBERTa model family with the aim of balancing computational efficiency and robust performance.

3. Objectives

As previously stated, certain aspects of a woman's life may be the focus of sexist attitudes, and the ability to automatically identify which of these aspects of women are being more frequently attacked in social networks will facilitate the development of policies to combat sexism. This study aims to classify tweets identified as sexist according to the type of sexism involved. This is a multi-label classification task. In this manner, each tweet identified by the system as sexist is to be assigned one or more of the following categories:

- **Ideological and Inequality**: the text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression;
- **Stereotyping and Dominance**: the text expresses false ideas about women that suggest they are more suitable to fulfil certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hard work, etc.), or claims that men are somehow superior to women;
- **Objectification**: the text presents women as objects apart from their dignity and personal aspects or assumes or describes certain physical qualities that women must have to fulfil traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women's bodies at the disposal of men, etc.);
- **Sexual Violence**: the text includes or describes sexual suggestions, requests for sexual favours or harassment of a sexual nature (rape or sexual assault);
- **Misogyny and Non-Sexual Violence**: the text expresses hatred and violence towards women, different to that with sexual connotations.

The objective of this task is to classify tweets, both in English and Spanish, according to whether they contain sexist expressions or behaviours. Initially, the classification should identify whether a given tweet contains sexist content, and subsequently, the category of sexism present in the tweets.

4. Dataset

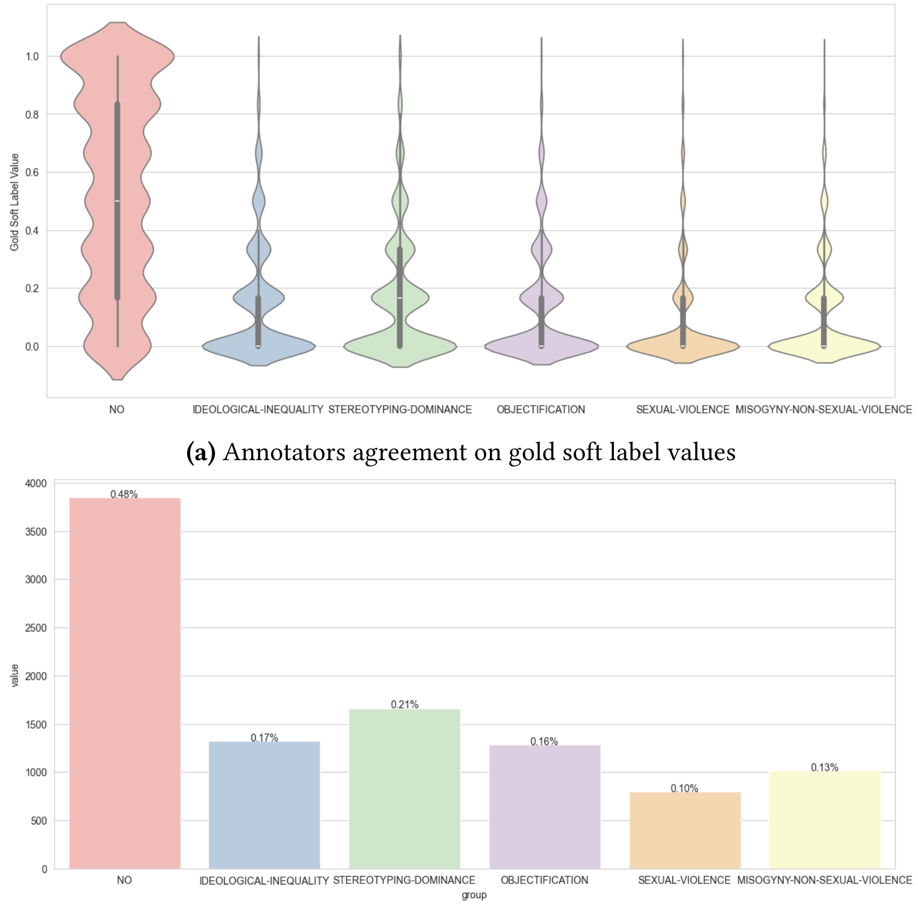
The EXIST 2024 Tweets Dataset comprises over 10,000 labelled tweets. In particular, the challenge presents a standard splitting of the dataset into three subsets: a training set comprising 6,920 tweets, a development set comprising 1,038 tweets, and a test set comprising 2,076 tweets. The entirety of the dataset is bilingual, with a ratio of 0.9 to 1 between English and Spanish tweets (3,749 and 4,209 respectively). This ratio was estimated from the training and development datasets. The aforementioned splitting has been disregarded during the development phase, in favour of an alternative internal splitting to maintain an internal test set. The dataset entries have been shuffled, keeping 80% of the size for our internal development set (training and validation sets), while reserving the remaining 20% for our internal test set.

Each tweet in the dataset is represented as a JSON object containing various attributes. These attributes include a unique identifier for the tweet, the language of the text, and the text of the tweet itself. Additionally, metadata about the annotators is provided, including the number of annotators, and their unique identifiers, gender, age group, ethnicity, level of education, and country of residence. The dataset also includes sets of labels for the three tasks about sexism in tweets.

The EXIST 2024 dataset was annotated by collecting the opinions of various annotators regarding the presence and, if any, nature of sexism in the provided tweets. Six annotators voted on each tweet, selecting one or more of six categories, with the restriction that selecting “NO” precludes selecting any other category. An “UNKNOWN” label is used when an annotator does not provide a label, but this is not a class to be predicted.

The dataset provided for the challenge features both soft and hard gold labels. The soft labels indicate the proportion of annotators who selected each category as shown in 1, reflecting the multi-label nature of the problem. Therefore, the sum of the “NO” label and the highest value among other labels cannot exceed one.

Gold hard labels are derived from soft labels using a probability threshold. If a category is chosen by more than one annotator, it becomes a hard label. Tweets without a category that exceeds the threshold are excluded from the evaluation.



(a) Annotators agreement on gold soft label values

(b) Gold hard labels distribution

Figure 1: (a) Violin plot showing the distribution of soft gold labels in the training and development sets. The y-axis represents the value of the label and thus the level of agreement between the annotators for a given class. For each class, the width of the violin represents, visually and not in precise scale, the frequency of that particular value for that class. In addition, the distribution is also shown by the box plot contained in each violin. (b) Bar chart showing the distribution of gold hard labels in the training and development sets.

The first step of preprocessing involved the removal of “irrelevant” features for our approach to the task (“labels_task2”, “labels_task1”, “labels_task3”, “annotators”, “number_annotators”, “gender_annotators”, “age_annotators”, “ethnicities_annotators”, “study_levels_annotators”, “countries_annotators”). The features in question represent supplementary metadata that is not strictly necessary nor present when addressing standard cases of classification approaches. In our view, building a model that relied on these features would have resulted in the creation of a highly customised system, making it challenging to extend to general cases of online sexism identification in the absence of datasets annotated in a manner compatible with the one provided in this instance.

Training the model on features related to the personal characteristics of the annotators would also have introduced new biases into the system, making it more likely to produce predictions driven by elements such as the ethnicity and gender of the annotators, with potentially discriminatory implications.

Building on [20], we implemented a preprocessing pipeline to improve the classification performance. Specifically, we applied language-agnostic functions to remove all URLs, user tags, numbers and dates, useless spaces and inverted exclamation/question marks at the beginning of Spanish phrases, as well as any syntagmas that did not contain relevant information for categorising the tweet. All instances of multiple exclamation marks, multiple question marks and mixed question and exclamation marks were identified and reformatted to reduce the variability introduced by the alternation and repetition of these two characters, which are often unevenly distributed. We identified and reduced all repetitions of punctuation and letters, including extended words, to just 2 repetitions so that they retained a different meaning from the single occurrence of the same character, while also ensuring that all repetitions were consistent. The final issue we addressed was the omission of a space between the period at the end of a sentence and the following word. This syntactic error was a common occurrence in the dataset.

Figure 2 illustrates two exemplary tweets in English and Spanish, respectively. These tweets demonstrate the processing of the source tweet and its appearance following the decoding of the tokenization.

Tweet Processing Example

English tweet:

Source tweet:

```
@user5 Wow!!! https://example.com insaneee I can’t evennn believe it???!!!
```

Processed:

```
Wow !! insanee I can’t evenn believe it ?!
```

Processed and decoded from the tokenizer:

```
<s> Wow!! insanee I can’t evenn believe it?!</s>
```

Spanish tweet:

Source tweet:

```
@usuario3 ¡Mira esto!!! https://ejemplo.com ¿¿¿Qué??? ¡¡Es increíble!!
```

Processed:

```
Mira esto !! Qué ?? Es increíble !!
```

Processed and decoded from the tokenizer:

```
<s> Mira esto!! Qué?? Es increíble!!</s>
```

Figure 2: Example tweets in English and Spanish demonstrating the effect of the processing function.

After preprocessing, we created a new split of the labelled dataset by randomly shuffling its entries, keeping 80% of the size for our internal development set (training and validation sets), while reserving the remaining 20% for our internal test set.

5. Baseline models

In addition to the dataset, we were provided with a baseline for each task. This served as an initial reference point for comparing the performance of various models. This approach enabled us to evaluate how well or poorly a model performed in comparison to an unsophisticated or simple system. In our case, considering only the third task, we had two baselines: one for the majority class and one for the minority class.

The majority class baseline is a non-informative system where all instances are labelled with the majority class, while the minority class is a non-informative system where all instances are classified as the minority class. The term “non-informative” is used to describe a system or model that does not utilise any significant information or features of the data to make predictions. Instead, it simply assigns all instances to a particular class, regardless of the actual data.

The majority class is the “NO” class (Figure 3a), and the minority class is the “SEXUAL-VIOLENCE” class (Figure 3b).

```
{
  "test_case": "EXIST2024",
  "id": "100001",
  "value": {
    "IDEOLOGICAL-INEQUALITY": 0.0,
    "STEREOTYPING-DOMINANCE": 0.0,
    "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.0,
    "SEXUAL-VIOLENCE": 0.0,
    "OBJECTIFICATION": 0.0,
    "NO": 1.0
  }
}
```

(a) Majority Class Instance

```
{
  "test_case": "EXIST2024",
  "id": "100001",
  "value": {
    "IDEOLOGICAL-INEQUALITY": 0.0,
    "STEREOTYPING-DOMINANCE": 0.0,
    "MISOGYNY-NON-SEXUAL-VIOLENCE": 0.0,
    "SEXUAL-VIOLENCE": 1.0,
    "OBJECTIFICATION": 0.0,
    "NO": 0.0
  }
}
```

(b) Minority Class Instance

Figure 3: Labels for all the instances of the majority and minority baselines

6. Resources employed

The development of RoBEXedda models was constrained by a limited number of resources, as the shared machine assigned to us by the University of Pisa was also exploited by other students at the same time. The machine was equipped with a NVIDIA V100 with 32 GB of memory. Alternatively, Google Colab with the free plan was employed. An important mention goes to the Weight & Biases [21] library, adopted for model selection. This permitted the training of distinct configurations in parallel across multiple machines, with all results and plots being recorded directly on the library’s website. This was achieved through the Sweep paradigm.

7. Proposed methodology

A fundamental stage in developing RoBEXedda involved searching for cutting-edge approaches that fit well with our objectives. In addition to the techniques mentioned above that are used in the preprocessing phase of the data, the use of AdamW [22] as an optimiser, and the implementation of two distinct architectures based on the principles of Classifier Chain [23] and Binary Relevance [24], respectively, deserves a more in-depth mention. These techniques have been combined with original insights and approaches identified by our team. Both state-of-the-art approaches and integrations of original techniques are discussed in this chapter.

7.1. Architectures

The architectures that we evaluated differed in the classification head that was placed on top of the pretrained transformer. To address the multilingual nature of the task while respecting our computation constraints, we focused our model selection on pretrained models from the XML-RoBERTa family [10]. Our pipeline included, after the preprocessing phase, a tweet tokenization phase. After studying the dataset, we decided to set the length of the transformer input at 128 tokens, as this was found to be the optimal length for the average and maximum length of the tokenized tweets shown in Figure 4.

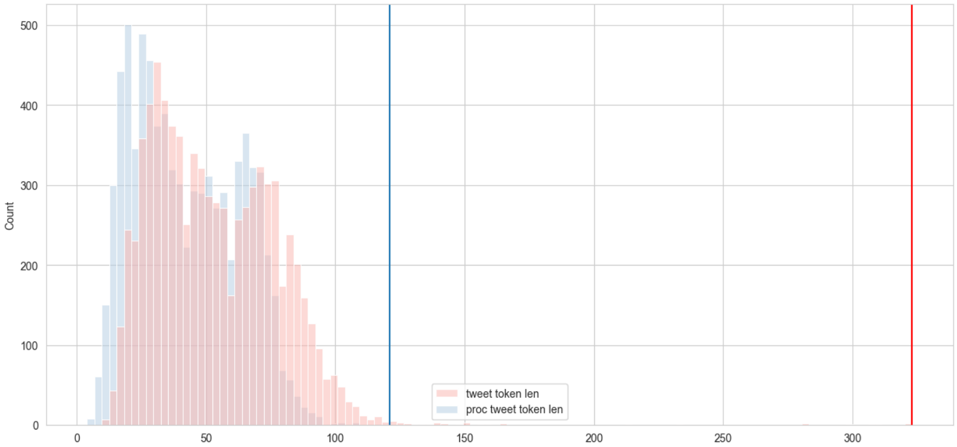


Figure 4: Distribution of tokenized input tweet lengths across the entire dataset. Raw tokenized data is shown in red with a vertical red line indicating the maximum length, while processed tokenized tweets are shown in blue with a vertical blue line indicating the maximum length.

Two main architectural archetypes, Classifier Chain and Binary Relevance, were considered during the model selection phase. We aimed to study the performance of the two architectures in tackling the task analysed. Among the three model proposals submitted by our team, two of them were indeed selected by us as the best Classifier Chain model and the best Binary Relevance model according to the validation metrics considered during model selection.

7.1.1. Binary Relevance Architecture

The first architecture is based on the concept of Binary Relevance (BR) [24], a very simple technique, often used as a baseline in multi-label classification problems. BR is a problem decomposition technique that assumes that each label is independent of the others and can therefore be treated separately. Furthermore, BR is a computationally efficient technique, making it a practical choice for our context.

The BR-based architecture consists of two fully connected hidden feedforward layers with GELU activation function, placed on top of the pretrained transformer, receiving as input the contextual embedding of the classification token produced by it. The output of the transformer does not go through the internal pooling or classification layer of the transformer but is taken from the last block of multi-head bidirectional attention. The head ends with a linear classification layer, followed by the application of a sigmoidal function to the 6 computed outputs to obtain the 6 different probabilities, one for each class. This approach is illustrated in Figure 5.

7.1.2. Classifier Chain Architecture

In Classifier Chain architectures, classifiers are chained together in a directed structure so that predictions from individual labels become features for other classifiers. Such methods are known in the literature for their flexibility and effectiveness, achieving state-of-the-art performance on many datasets and multi-label evaluation metrics [23].

As discussed in the previous chapter, the soft labels that our model should predict do not represent a probability distribution on mutually exclusive classes (since several soft labels can be predicted simultaneously), with the exception of the label “NO”, the only one whose value has a relationship with the values of the others.

In particular, the target values represent the proportion of annotators who have chosen a set of labels to associate with each specific tweet. In the case of the “NO” class, it is not possible to select this label in conjunction with any of the remaining five labels. However, multiple categories of sexism can

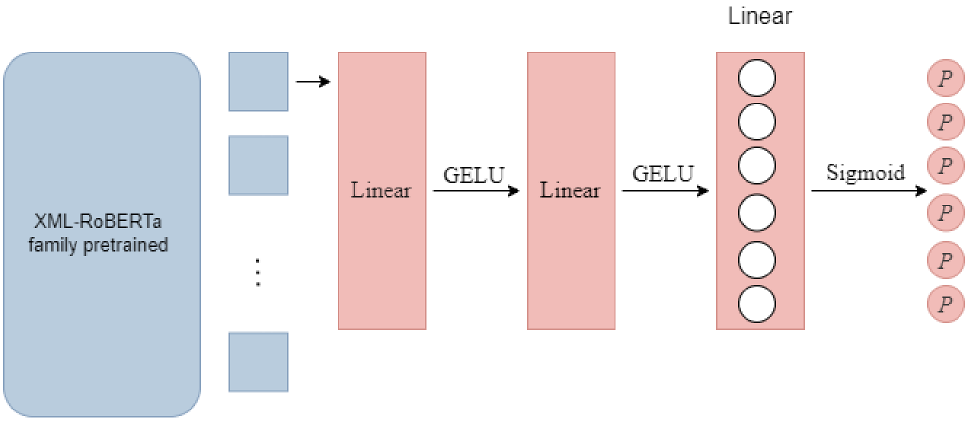


Figure 5: Classification head architecture using Binary Relevance concept.

be selected without any restrictions. The sum of the “NO” label and the maximum value among the remaining labels cannot exceed one. This intuition led to the development of an original architectural idea, which consists of using a Classifier Chain model that can use its prediction of the “NO” label as a feature for predicting the remaining labels.

The proposed Classifier Chain architecture comprises two modules, both of which constitute the multi-label classifier head, situated at the top of the pretrained transformer. Both modules receive contextual embeddings produced by the transformer following the processing of an observed tweet.

The first module comprises three fully connected feedforward layers, with GELU activation functions in the hidden neurons and a sigmoid activation function in the output layer. The objective of this module is to output the prediction of the value of the “NO” label associated with the input tweet.

The second module is analogous to the first in structure and its objective is to predict the values of the remaining five soft labels. In light of the success of Classifier Chain architectures, we hypothesised that the prediction produced by the first module could be used as input to the second module, thereby serving as a feature in the prediction of the remaining five soft labels. A noteworthy design choice is that the prediction of the “NO” label is given as input to the second module at a higher level of the architecture (the second hidden layer rather than the first). The rationale behind this decision is that the prediction of the first classifier (the first module) can more effectively represent a high-level feature of the subsequent classifier, at a higher level of abstraction, than the contextual embedding returned by the transformer.

During the training phase, the second module was trained using the teacher forcing technique, where the input from the previous classifier in the chain was replaced by the corresponding gold label. Figure 6 shows the design of the approach.

7.2. Training

The pretrained model is employed in conjunction with the classification heads, which were based on Classifier Chain and Binary Relevance architectures respectively described in sections 7.1.2 and 7.1.1. The training parameters include learning rate, dropout, optimiser, hidden layer size, batch size and epochs, which are optimised during model selection. One of the state-of-the-art techniques explored in the training process is the AdamW optimiser.

AdamW (Adaptive Moment Estimation with Weight Decay) is an optimisation algorithm that combines the properties of Adam with a weight decay mechanism. Adam is known to adapt individual learning rates for each parameter using estimates of the first and second moments of the gradients. AdamW differs from Adam for the weight decay that is applied separately from the updating of the gradients. This approach allows more precise control of the weight decay and avoids unwanted interference between the learning rate and the weight decay itself, which complicates the optimal choice of these hyperparameters and improves convergence efficiency. This facilitates the choice of learning

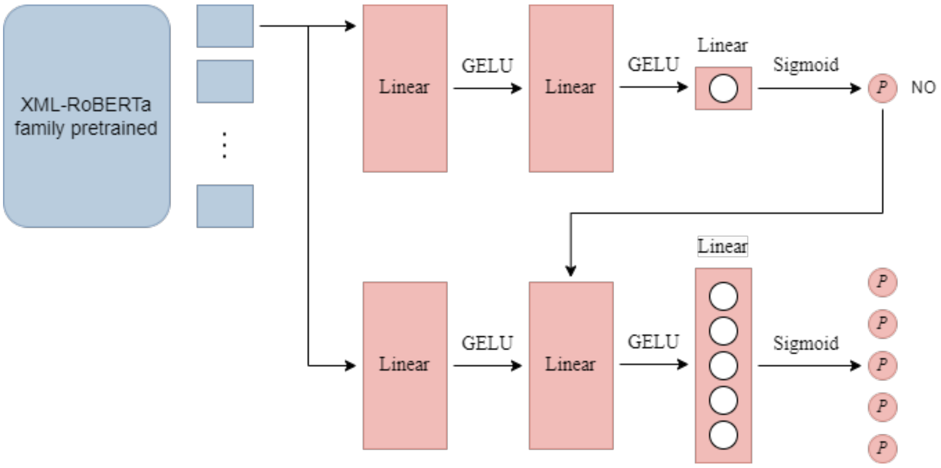


Figure 6: Classification head architecture using Classifier Chain concept.

parameters and leads to more efficient convergence. Studies have shown that AdamW tends to produce models with a greater capacity for generalisation than Adam [25].

This finding was confirmed during our preliminary exploration phase. Indeed, we observed that AdamW performed better than Adam, so we decided to directly employ it. In addition, we noticed that also compared to Stochastic Gradient Descent (SGD), it significantly reduces the time needed to find an effective combination of hyperparameters, allowing for more efficient and faster tuning [22].

To train our models, we employed the Binary Cross-Entropy (BCE) loss, which also served as the primary validation metric. In addition to the BCE loss, we evaluated our models using other validation metrics described below to ensure a comprehensive assessment of performance.

In particular, we exploited the PyEvALL (The Python library to Evaluate ALL) framework [26] that offers several assessment metrics including F1 score, ICM (Information Contrast Measure) [27], and a soft version of ICM (ICM-Soft). All of these additional metrics were observed during the model selection process. The sole criterion for the selection was the validation metric (BCE), except for one of the three candidate models, which was selected based on the ICM-Soft.

The ICM-Soft criterion represents an extension of ICM, a measure that has been demonstrated to be analytically superior to cases where categories have a hierarchical structure and items may belong to more than one category. However, in contrast to its standard counterpart, the ICM-Soft accepts both soft system outputs and soft ground truth assignments.

7.3. Model Selection

The initial phase of model selection involved an analysis designed to gain a first understanding of the influence of hyperparameters on model performance. To facilitate this process, we employed the W&B (Weights & Biases) library (wandb) to train distinct configurations of parameters in parallel across multiple machines, with all results and plots being recorded directly on the library's website. The utilisation of the sweeps and agents features enables the automation of hyperparameter search by defining a search space and strategy and running the experiments according to this configuration.

The objective was to minimise the validation loss. To achieve this objective, a preliminary random search [28] was conducted, during which several pretrained models from the XML-RoBERTa family were tested. The optimal choice was identified as "sdadas/xlm-roberta-large-twitter" [29].

The initial search was followed by a Bayesian search. The fixed parameters for the model are the Batch Size fixed at 64, the maximum number of Epochs set to 15 (with early-stopping, patience 2), AdamW as the optimizer, and the pretrained model "sdadas/xlm-roberta-large-twitter".

Table 1 shows the ranges of the other hyperparameters that were explored during the Bayesian search.

Table 1
Hyperparameter Ranges for Model Selection.

Parameter	Distribution	Values/Range
classifier_type	Categorical	chain, ff
dropout	Quantized Uniform	min: 0.2, max: 0.7, q: 0.05
hidden_layer_size	Categorical	128, 256, 512, 1024
learning_rate	Log Uniform	min: -13, max: -10

To prevent overfitting, during Bayesian search we employed early stopping. This ensured that the model did not continue to train beyond the point where its performance on the validation set started to degrade. After this fine-grade search, we can see the top 10 runs in Figure 7.

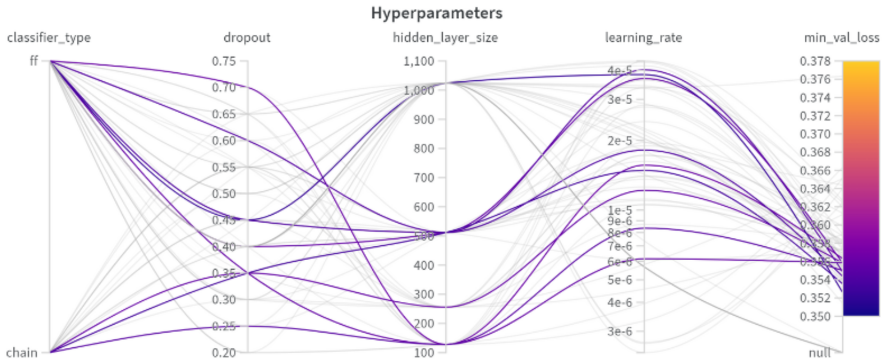


Figure 7: Top 10 runs from the final Bayesian search. Each line in the graph represents a specific model run, characterised by specific hyperparameter values and the corresponding minimum validation loss result.

From this graph, which highlights the top 10 runs found during model selection, we can understand that both types of classifiers (Binary Relevance and Classifier Chain) can achieve competitive results. Dropout tend to be slightly more effective between 0.2 and 0.4, indicating that minimal regularisation is preferable. The results indicate that smaller hidden layer sizes (equal to or less than 512) are more common among the best runs. Furthermore, smaller learning rates are associated with a lower minimum validation loss, which highlights the importance of precise fine-tuning of the learning rate to improve model convergence. In any case, the most crucial hyperparameters were identified as the values of the hidden layer size and the learning rate. By operating within the range of interest, adjusting these parameters to more specific values led to changes in the results. The hidden layer size was identified as the most influential factor in the search for optimal models, exhibiting a strong negative correlation (lower values perform better) with respect to the minimum validation loss. The learning rate also showed significant importance, with a moderate negative correlation. In contrast, the value related to dropout did not have a significant impact, as it showed minimal importance and a low positive correlation.

The model selection process led to the identification of the most promising hyperparameter configurations. From these, three RoBEXedda models were selected for submission to the challenge (a maximum of three candidates per team were allowed). These models have been selected by us for specific features and are identified as “Best BR”, “Best Chain”, and “Best ICM-Soft”. Best BR and Best Chain represent respectively the Binary Relevance model and the Classifier Chain model that obtained the best BCE loss on the validation set. Best ICM-Soft is the model chosen for obtaining the best ICM-Soft on the validation set (also featuring the Binary Relevance architecture).

All three RoBEXedda models share the parameters of maximum number of ‘epochs’ (15), ‘batch size’ (64), and ‘early stopping patience’ (2). The “Best ICM-Soft” model was trained with ‘learning rate’ of 3.6936026e-5, ‘training epochs’ of 4, ‘dropout’ percentage of 0.4, and ‘hidden layer size’ of 512. The “Best Chain” model was trained with ‘learning rate’ of 1e-5, ‘training epochs’ of 7, ‘dropout’ percentage of

0.2, and 'hidden layer size' of 128. The "Best BR" model was trained with 'learning rate' of 1.8176664e-5, 'training epochs' of 4, 'dropout' percentage of 0.25, and 'hidden layer size' of 512. Table 2 shows a brief summary of the selected models.

Table 2
The core features on which the models have been selected are presented in bold.

Model	Description
Best ICM-Soft	Model (Binary Relevance) with best **ICM Soft** on Validation Set
Best Chain	**Classifier Chain** model with best **BCE Loss** on Validation Set
Best BR	**Binary Relevance** model with best **BCE Loss** on Validation Set

8. Results

Following the selection of the models, an internal assessment was conducted to evaluate the system performance. This section will discuss both the internal assessment phase and the scores achieved by our models in the EXIST 2024 challenge.

8.1. Internal Assessment

Following the retraining on both the training and validation sets, the three RoBEXedda models identified in model selection phase were evaluated on an internal test set. Tables 4 - 5 show the results of the internal test set, averaged over five different weight initialisations.

Table 3
Results on internal test set (Soft Metrics).

	ICM-Soft	ICM-Soft Norm	BCE
Gold Label	9.82805	1	0.24241
Best ICM-Soft	**-2.5544**	**0.370045**	0.35034
Best Chain	-2.70407	0.362431	0.35194
Best BR	-2.56684	0.369412	**0.34715**
Majority Class	-8.99827	0.042214	17.73328
Minority Class	-40.6109	0	33.08923

Table 4
Results on internal test set (Hard Metrics).

	ICM	ICM Norm	F1 score
Gold Label	2.41576	1	1
Best ICM-Soft	**0.16134**	**0.533393**	**0.612121**
Best Chain	0.102222	0.521157	0.610467
Best BR	0.0856231	0.517722	0.606697
Majority Class	-1.84935	0.117231	0.110364
Minority Class	-3.29148	0	0.0328426

8.2. Challenge Results

After the final assessment, the RoBEXedda models were retrained on the entire dataset, and subsequently employed to generate the predictions on the official blind test set, submitted for our participation in the challenge (Task 3 Soft-Soft). Table 6 shows the results of our approaches compared to the official baselines described in Section 5 and the gold labels.

Table 5
F1 score for each class in the internal test set. The classes I-I, S-D, Obj, S-V, M-NS refer respectively to, Ideological Inequality, Stereotyping Dominance, Objectification, Sexual Violence, Misogyny Non-Sexual Violence.

	NO	**I-I**	**S-D**	**OBJ**	**S-V**	**M-NS**
Gold Label	1	1	1	1	1	1
Best ICM-Soft	0.82795	0.62189	0.52088	0.56732	0.58133	0.55333
Best Chain	0.82510	0.60382	0.52410	0.56695	0.57788	0.56491
Best BR	0.82269	0.61063	0.52307	0.56401	0.57009	0.54966
Majority Class	0.66218	0	0	0	0	0
Minority Class	0	0	0	0	0.19705	0

In the Task 3 Soft-Soft competition, our models achieved the 4th, 5th, and 6th position in the global ranking, ranking our team (Medusa), just behind the "NYCU-NLP" team (Task Winner), whose three models took the 1st, 2nd and 3rd positions in the ranking. In particular, the Best ICM-Soft model achieved the 4th position, the Best Chain model achieved the 5th position, and the Best Binary Relevance model achieved the 6th position.

Table 6
Task 3 Soft-Soft final results.

		English and Spanish		**Only English**		**Only Spanish**	
Rank	Model	**ICM-Soft**	**ICM-Soft Norm**	**ICM-Soft**	**ICM-Soft Norm**	**ICM-Soft**	**ICM-Soft Norm**
0	Gold Label	9.4686	1	9.1255	1	9.6071	1
1^{st}	Task Winner	-1.1762	0.4379	-1.2583	0.4311	-1.1280	0.4413
4^{th}	**Best ICM-Soft**	**-2.2055**	**0.3835**	**-2.0694**	**0.3866**	**-2.2859**	**0.3810**
5^{th}	**Best Chain**	-2.4010	0.3732	-2.2945	0.3743	-2.4730	0.3713
6^{th}	**Best BR**	-2.4142	0.3725	-2.3419	0.3717	-2.4397	0.3730
28^{th}	Majority Class	-8.7089	0.0401	-8.2105	0.0501	-9.0314	0.0300
33^{th}	Minority Class	-46.1080	0	-46.9473	0	-45.4260	0

A first observation is that, although the outcomes are essentially comparable, all RoBEXedda models demonstrate a slight advantage in English with respect to Spanish. This can be attributed to the composition of the training data for the pretrained model, which comprised 50.9% English tweets and 14.4% Spanish tweets [30]. A second consideration is that the use of a Classifier Chain did not result in enhanced efficacy compared to the Binary Relevance approach. One potential explanation for this finding is that the Binary Relevance architecture is more effective in representing dependencies between the labels in the analysed task.

In the final ranking, the Best ICMSoft model emerged as the most effective between our choices. It is noteworthy that it was the sole model selected based on the ICM Soft measure, which is not always synchronised with the BCE loss.

This emphasises the significance of considering alternative evaluation metrics when selecting models. Exploring model selection based on this measure could prove to be an intriguing avenue for future research.

9. Conclusion and future directions

Participation in the EXIST 2024 challenge aimed at categorising sexist content in tweets has provided valuable insights into the detection and classification of online misogyny. Utilising a robust dataset of over 10,000 tweets in both English and Spanish, we developed and evaluated three distinct neural network models based on Binary Relevance and Classifier Chain architectures. The results demonstrate the potential of advanced machine learning techniques in addressing the pervasive issue of online sexism and underscore the importance of continued research and development in this critical area.

Although the results obtained are far from perfect, we believe that our analyses have nevertheless led to interesting insights. One of them is the fact that good results in the task of classifying sexist behaviour in social networks can be achieved with limited resources. Indeed, as mentioned above, the development of RoBEXeddA was carried out in particularly narrow time slots within the EXIST 2024 time window, distributed on a few shared machines.

The lack of computational resources is not the sole point of improvement in our process. Indeed, the project is open to numerous possible future developments. It would be of interest to undertake a model selection process that could screen larger pretrained transformer models. Among potential future additions to our project, it might be worthwhile to test other state-of-the-art techniques, such as data augmentation and Ensemble Learning, which were not included in the challenge preparation in favour of producing an efficient system in the shortest possible time.

An additional intriguing attempt would be to conduct a separate pre-training of the classifier head, preceding the entire fine-tuning of the model. This is because, following our tests, training the random initialised head required a much higher learning rate than what was allowed in the model's finetuning. Therefore, we could have obtained more stable training curves and encouraged the learning of an initial representation of the dataset's features.

References

[1] 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 the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 813–854. URL: https://ceur-ws.org/Vol-3497/paper-070.pdf.

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

[3] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[4] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes (extended overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024- Conference and Labs of the Evaluation Forum, 2024.

[5] L. Plaza, J. Carrillo-de-Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, EXIST 2024: sexism identification in social networks and memes, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, volume 14612 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 498–504. URL: https://doi.org/10.1007/978-3-031-56069-9_68. doi:10.1007/978-3-031-56069-9_68.

[6] Clef 2024 conference and labs of the evaluation forum, 2024. https://clef2024.clef-initiative.eu/index.php.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017,

Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[8] M. Zhang, Y. Li, X. Liu, X. Geng, Binary relevance for multi-label learning: an overview, Frontiers Comput. Sci. 12 (2018) 191–202. URL: https://doi.org/10.1007/s11704-017-7031-7. doi:10.1007/S11704-017-7031-7.

[9] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains for multi-label classification, Mach. Learn. 85 (2011) 333–359. URL: https://doi.org/10.1007/s10994-011-5256-5. doi:10.1007/S10994-011-5256-5.

[10] 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: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/V1/2020.ACL-MAIN.747.

[11] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, CoRR abs/2402.06196 (2024). URL: https://doi.org/10.48550/arXiv.2402.06196. doi:10.48550/ARXIV.2402.06196. arXiv:2402.06196.

[12] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, Taking the human out of the loop: A review of bayesian optimization, Proc. IEEE 104 (2016) 148–175. URL: https://doi.org/10.1109/JPROC.2015.2494218. doi:10.1109/JPROC.2015.2494218.

[13] J. Gonzalo, M. Montes-y-Gómez, P. Rosso, Iberlef 2021 overview: Natural language processing for iberian languages, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Álvarez-Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 1–15. URL: https://ceur-ws.org/Vol-2943/Overview_iberLEF_2021.pdf.

[14] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual BERT and ensemble models, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Álvarez-Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 356–373. URL: https://ceur-ws.org/Vol-2943/exist_paper2.pdf.

[15] F. M. P. del Arco, M. D. Molina-González, L. A. U. López, M. T. Martín-Valdivia, Sexism identification in social networks using a multi-task learning system, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Álvarez-Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 491–499. URL: https://ceur-ws.org/Vol-2943/exist_paper16.pdf.

[16] A. Jiang, A. Zubiaga, QMUL-SDS at EXIST: leveraging pre-trained semantics and lexical features for multilingual sexism detection in social networks, in: M. Montes, P. Rosso, J. Gonzalo, M. E. Aragón, R. Agerri, M. Á. Álvarez-Carmona, E. Á. Mellado, J. Carrillo-de-Albornoz, L. Chiruzzo, L. A. de Freitas, H. Gómez-Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. del Arco, M. Taulé (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with

the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 469–483. URL: https://ceur-ws.org/Vol-2943/exist_paper14.pdf.

[17] V. Ahuir, J. González, L. Hurtado, Enhancing sexism identification and categorization in low-data situations, in: M. Montes-y-Gómez, J. Gonzalo, F. Rangel, M. Casavantes, M. Á. Á. Carmona, G. Bel-Enguix, H. J. Escalante, L. A. de Freitas, A. Miranda-Escalada, F. J. Rodríguez-Sanchez, A. Rosá, M. A. S. Cabezudo, M. Taulé, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022, volume 3202 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper5.pdf.

[18] J. A. García-Díaz, S. M. J. Zafra, R. C. Palacios, R. Valencia-García, Umuteam at EXIST 2022: Knowledge integration and ensemble learning for multilingual sexism identification and categorization using linguistic features and transformers, in: M. Montes-y-Gómez, J. Gonzalo, F. Rangel, M. Casavantes, M. Á. Á. Carmona, G. Bel-Enguix, H. J. Escalante, L. A. de Freitas, A. Miranda-Escalada, F. J. Rodríguez-Sanchez, A. Rosá, M. A. S. Cabezudo, M. Taulé, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022, volume 3202 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3202/exist-paper14.pdf.

[19] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, AI-UPV at EXIST 2023 - sexism characterization using large language models under the learning with disagreements regime, CoRR abs/2307.03385 (2023). URL: https://doi.org/10.48550/arXiv.2307.03385. doi:10.48550/ARXIV.2307.03385. arXiv:2307.03385.

[20] D. Effrosynidis, S. Symeonidis, A. Arampatzis, A comparison of pre-processing techniques for twitter sentiment analysis, in: J. Kamps, G. Tsakonas, Y. Manolopoulos, L. S. Iliadis, I. Karydis (Eds.), Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings, volume 10450 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 394–406. URL: https://doi.org/10.1007/978-3-319-67008-9_31. doi:10.1007/978-3-319-67008-9_31.

[21] L. Biewald, Experiment tracking with weights and biases, 2020. URL: https://www.wandb.com/, software available from wandb.com.

[22] Y. Pan, Y. Li, Toward understanding why adam converges faster than SGD for transformers, CoRR abs/2306.00204 (2023). URL: https://doi.org/10.48550/arXiv.2306.00204. doi:10.48550/ARXIV.2306.00204. arXiv:2306.00204.

[23] J. Read, B. Pfahringer, G. Holmes, E. Frank, Classifier chains: A review and perspectives, J. Artif. Intell. Res. 70 (2021) 683–718. URL: https://doi.org/10.1613/jair.1.12376. doi:10.1613/JAIR.1.12376.

[24] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, A. Bahamonde, Binary relevance efficacy for multilabel classification, Prog. Artif. Intell. 1 (2012) 303–313. URL: https://doi.org/10.1007/s13748-012-0030-x. doi:10.1007/S13748-012-0030-X.

[25] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[26] Pyevall, 2024. https://github.com/UNEDLENAR/PyEvALL.

[27] E. Amigó, A. D. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 5809–5819. URL: https://doi.org/10.18653/v1/2022.acl-long.399. doi:10.18653/V1/2022.ACL-LONG.399.

[28] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2012) 281–305. URL: https://dl.acm.org/doi/10.5555/2503308.2188395. doi:10.5555/2503308.

2188395.

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

[30] S. Dadas, OPI at semeval-2023 task 9: A simple but effective approach to multilingual tweet intimacy analysis, in: A. K. Ojha, A. S. Dogruöz, G. D. S. Martino, H. T. Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023, Association for Computational Linguistics, 2023, pp. 150–154. URL: https://doi.org/10.18653/v1/2023.semeval-1.21. doi:10.18653/V1/2023.SEMEVAL-1.21.