Combining Transformer Based Language Models with Socio-demographic Information for Improving Sexism Detection in Social Media

Notebook for the EXIST Lab at CLEF 2023

Jacobo Pedrosa-Marín[1,*], Jorge Carrillo-de-Albornoz[1,2] and Laura Plaza[1,2]

[1]NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), 28040, Spain
[2]RMIT University, 3000, Australia

Abstract

Detecting and addressing sexism in social networks is crucial for fostering inclusive and respectful digital spaces. The third edition of the EXIST competition emphasizes the importance of incorporating annotator disagreement into the classification process, recognizing the inherent challenges and diversity of perspectives in identifying sexist content. In this paper, we present our participation in the EXIST 2023 campaign where we propose systems for Task 1 (Sexism Identification) and task 2 (Source Intention Identification), both for hard and soft evaluation contexts. We adopted a primary strategy that involved data augmentation to enhance the training dataset. By leveraging techniques such as translation and the use of transformers, we aimed to expand the available data and capture a broader range of linguistic patterns and expressions related to sexism. Additionally, the EXIST 2023 dataset allows to identify and exploit annotators characteristics such as gender and age. We have used this socio-demographic information to train different models that capture each age-gender cohort singularities, and used different strategies to combine them in a final decision, in the hard approaches, and a probability representation, in the soft approaches. The results achieved suggest that having different models for the different cohorts improves the efficiency of the classification.

Keywords

Sexism Detection, Sexism Identification Learning with disagreement, Transformer Models, Natural Language Processing

1. Introduction

In recent years, the rise of social media platforms such as Twitter and Facebook has brought about a significant transformation in communication and society. These platforms have provided users with new means to express their ideas, thoughts, and knowledge. These platforms hold tremendous potential for information dissemination, and researchers have extensively examined their impact in various fields, including politics and medicine. However, the proliferation of hate speech on these platforms has emerged as a growing concern. The rise of hate speech

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
jacobo.pedrosa@lsi.uned.es (J. Pedrosa-Marín); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); lplaza@lsi.uned.es (L. Plaza)
0009-0004-1224-3760 (J. Pedrosa-Marín)

poses a significant challenge, demanding careful attention and effective solutions to ensure a safe and inclusive online environment [1].

Detecting and preventing hate speech in social medial can be challenging, especially considering the overwhelming volume of data generated on these platforms every second, which makes it necessary to employ automated methods and advanced technologies to process and classify the content efficiently. Over the years, numerous studies and competitions have emerged, focusing on the analysis and automation of online community management. These initiatives aim to address various challenges associated with content detection and moderation, including Anomaly Detection [2], Phishing Detection [3], Toxicity Detection [4] and Sexism Detection [5, 6].

In this paper, we focus on a particular form of harmful content: sexist expressions. Sexism encompasses actions or attitudes that exhibit prejudice or discrimination towards individuals based on their gender. It is closely linked to societal beliefs and expectations regarding the roles that individuals should adhere to, with its repercussions primarily impacting women. Detecting sexism in online platforms is crucial for creating inclusive and respectful digital spaces. It enables the identification and moderation of harmful content, promotes gender equality, and helps to prevent the perpetuation of discriminatory behaviors.

The EXIST challenges serves as an avenue for researchers and practitioners to develop and present their approaches and models in tackling the complex task of sexism detection in social media [7]. The 2021 and 2022 editions of the competition were held in the IberLEF forum [1] and were the first shared tasks on sexism detection in social networks whose aim was to identify and classify various forms of sexism, ranging from explicit and hostile expressions to more nuanced or even benevolent behaviors involving implicit sexism. With participation from over 50 teams from research institutions and companies worldwide, the substantial interest shown by the research community underscores the significance of the problem at hand.

The third edition of the EXIST challenge at CLEF builds upon the tasks addressed in previous years, while facing a new challenge: the identification of the author's intention behind sexist messages. However, the main innovation is the adoption of the "learning with disagreements" paradigm [8] and the participation of annotators of different genders and ages. This approach aims to mitigate the potential "label bias" by incorporating diverse perspectives and sensitivities from different population groups, ensuring a more comprehensive reflection of viewpoints and recognizing that annotators' socio-demographic backgrounds can shape their labeling decisions [9, 10].

This paper presents the participation of the JPM-UNED team in EXIST 2023 at CLEF. Our approach includes the use of transformers along with different strategies (such as voting and aggregation) to leverage the collective knowledge and the disagreement among the annotators to derive the most reliable predictions. By considering the socio-demographic variances among annotators and employing tailored models and strategies, our objective is to enhance the accuracy and robustness of our classification approach. This approach contributes to a more deeper understanding of the complexities involved in sexism detection and enhances the overall effectiveness of our model.

This paper is organized as follows: Section 2 reviews the state of the art in sexism detection

[1] https://sites.google.com/view/iberlef2022/home

and learning with disagreements; Section 3 presents this edition of EXIST, including an overview of the three tasks proposed and the dataset provided; Section 4 presents the systems developed for participating in the competition; Section 5 discusses the results; and Section 6 summarizes the main conclusions and discusses potential improvements for future work.

2. Related Work

In this section, we first examine the importance of sexism detection in social networks, review previous research in this area, and provide an overview of previous editions of the EXIST competition. We then briefly discuss the state-of-the-art in Learning With Disagreements (**LeWiDi**) to consider different approaches for managing tasks of this nature.

The detection of sexism has traditionally been regarded as a distinct form of hate speech [11]. It can be approached through various methods, such as data-driven models that incorporate n-grams and additional features [12], classical machine learning models [13], and deep learning models that utilize LSTM and CNN architectures [14]. Additionally, certain studies have explored the utilization of offensive lexicons like Hurtlex [15].

However, it is important to recognize that sexism is not always expressed as hate speech. As highlighted in [7], sexism can take on a "friendly" or even "humorous" tone, such as in the case of benevolent sexism and sexist jokes [16]. Consequently, novel approaches are necessary to detect the various forms of sexism, ranging from hostile and explicit to subtle and seemingly benign expressions.

In the following section, we present the EXIST 2021 and 2022 editions, which introduced the challenging task of detecting sexism in all its nuanced manifestations.

2.1. The EXIST Challenges

In the previous two editions of the EXIST competition, two tasks were proposed. Task 1 focused on sexism identification, aiming to detect whether a given post contains sexist content or not. Task 2, on the other hand, focused on sexism categorization, aiming to classify the type of sexism present in a post into one of the following five classes: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence y (v) misogyny and non-sexual violence.

In the 2021 edition of EXIST, the majority of submissions for both tasks relied on transformer-based models for classification. Out of the 23 participating teams, 14 teams utilized BERT, a widely used transformer model, as the foundation of their solutions. Additionally, 10 teams employed BETO, a version of BERT that is trained on Spanish text. Furthermore, 5 teams leveraged XLM-R, a multilingual variant of RoBERTa that supports multiple languages, including Spanish [5]. In the 2022 edition of EXIST, all participating teams utilized transformer-based solutions. Among these solutions, 8 teams used BERT, 5 teams used BETO, and 4 teams used RoBERTa. The widespread adoption of transformer-based models in both editions of EXIST highlights their effectiveness in tackling the challenges associated with sexism detection in social networks [6].

2.2. Working with Disagreements

Usually, in the fields of Artificial Intelligence (AI) and Natural Language Processing, datasets are built with instances having a single class or interpretation referred to as the "gold standard". However, this approach fails to capture the nuances of human behavior, which often involves disagreement and varying perspectives. Tasks that involve subjectivity or ambiguity inherently introduce the possibility of biased annotations influenced by the perspectives of individual annotators. Moreover, socio-demographic factors, such as education, age or gender, have the potential to influence the annotation process and introduce biases.

An emerging solution that is gaining popularity is to engage multiple annotators, ideally representing diverse demographic strata and socioeconomic contexts, and retain the labels provided by each annotator instead of relying solely on a gold standard. This approach allows the systems to incorporate various perspectives for each instance, enabling them to learn from different points of view.

Working with a dataset that lacks a unanimous label for each instance offers valuable insights, but it also requires ways to manage the divergent opinions among annotators. The current state of the art in learning with disagreement can be categorized into four main categories:

- **Judgements aggregation**: Methods that operate under the assumption that only a single "truth" exists for each instance typically aggregate all crowd annotations into a single label, commonly referred to as the "silver" label. There are multiple approaches to tackle this challenge, with the most straightforward being the adoption of a "majority vote". However, one of the widely employed techniques is the utilization of **Probabilistic aggregation methods**, which leverage the probabilities assigned to each label by individual annotators [17, 18, 19].
- **Filtering hard items:** this approach utilizes the disagreement information to filter the dataset by removing instances with significant disagreement. Within this category, [20] proposed two approaches. The first approach involves directly discarding instances that exhibit disagreement among annotators. The second approach is to train separate models for each annotator and discard predictions that demonstrate substantial disagreement.
- **Learning directly from crowds:** This classification approach acknowledges the absence of a single truth or gold standard and instead focuses on training a classifier directly from the crowd, utilizing probabilistic distributions or soft labels. The objective is to capture the collective knowledge of the annotators and incorporate their diverse perspectives into a single model. There are various strategies to approach this. For instance, [21] propose a "repeated labeling method" where replicas of each instance are created for each label, enabling multiple annotations per instance. Another interesting method in this category is presented by [22], which involves adding a crowd layer after the output layer during training.
- **Using both hard labels and information about disagreements:** These methods utilize both gold labels and disagreement to train the models. One approach is to train with hard labels, while incorporating crowd information as part of the loss function during the training process. This allows the model to learn from both the ground truth labels and the disagreements among the crowd annotations, improving its performance and capturing the collective knowledge of the annotators [23].

As we can see, there are various approaches available to address the challenge of working with disagreements. For further information on this topic, please refer to the survey conducted by [23]. This survey provides more in-depth insights and details on different methods and techniques that can be employed to handle disagreement in various tasks or domains.

3. The EXIST 2023 Lab at CLEF 2023

3.1. EXIST 2023 Tasks

The EXIST 2023 edition proposes the following three tasks: (i) sexism detection, (ii) source intention classification, and (iii) sexism categorization (see [10] for a detailed description). For each task, participants may provide both hard (a single "gold" label) and soft (a probabilistic label) outputs.

- **Task 1 - Sexism Detection:** The first task is a binary classification task where systems must decide whether or not a given tweet is sexist.
- **Task 2 - Source Intention Classification:** This task aims to categorize the message according to the intention of the author, which provides insights in the role played by social networks in the emission and dissemination of sexist messages. In this task, we propose a ternary classification task: (i) direct sexist message, (ii) reported sexist message and (iii) judgmental message.
- **Task 3 - Sexism Categorization:** Each sexist tweet must be categorized in one or more of the following categories, that reflect the facets of a woman's life that are the focus of the sexist message: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence and (v) misogyny and non-sexual violence.

3.2. EXIST 2023 Dataset

The EXIST 2023 dataset comprises tweets in both English and Spanish. The training set consists of over 3,200 tweets per language, while the development set includes 500 tweets per language. Additionally, the test set contains 1,000 tweets per language. To ensure diverse perspectives and mitigate label bias, each tweet in the dataset has been annotated by six individuals recruited through the Prolific service [2]. The annotators' gender (male/female[3]) and age (18-22 years old, 23-45 years old, +46 years old) are taken into account during the labeling process. Consequently, each tweet is labeled by annotators from a different gender and a different age groups.

For a more comprehensive understanding of the EXIST 2023 dataset, please refer to [10].

4. System description

In this section, we present our systems developed for participating in Task 1 and Task 2 of the EXIST 2023 competition. Our approach builds upon the methods discussed in the previous

[2] https://app.prolific.co

[3] Only male and female genders were consider for availability reasons

section and utilizes Transformer pre-trained models. To establish a baseline, we first examine the models that have been employed in previous editions of the competition. We then train and test our models using the EXIST 2021 Dataset for each language, employing the same configuration. The results for Spanish tweets are presented in Table 1, while Table 2 showcases the results for English tweets.

Table 1
Models review Spanish - baselines

Model	F-Measure
xlm-roberta-base	0.703
xlm-roberta-large	0.447
bert-base-multilingual-cased	0.701
distilbert-base-multilingual-cased	0.706
PlanTL-GOB-ES/roberta-base-bne	**0.751**
PlanTL-GOB-ES/roberta-large-bne	0.741
bertin-project/bertin-roberta-base-spanish	0.747
dccuchile/bert-base-spanish-wwm-cased	0.709
CenIA/distillbert-base-spanish-uncased	0.713

Table 2
Models review English - baselines

Model	F-Measure
xlm-roberta-base	0.652
xlm-roberta-large	0.676
bert-base-multilingual-cased	0.721
distilbert-base-multilingual-cased	0.701
roberta-large	0.288
distilbert-base-uncased	**0.741**
bert-base-uncased	0.733

After analyzing the results of our initial experiments (see Tables 1 and 2, we have selected the model with the best performance in Spanish, which is **PlanTL-GOB-ES/roberta-base-bne**, and in English, which is **distilbert-base-uncased**. Our remaining experiments will be conducted using these two models as the foundation. Based on these findings, we have decided to utilize these two models as the base for our future experiments.

As a previous step, we have also created an augmented version of the EXIST 2023 dataset by translating tweets from English to Spanish and vice versa. With this expanded dataset in hand, we have two parallel approaches to enhance the training process and our analysis.

Our initial approach involved fine-tuning the pre-trained models we had selected using the "repeated labeling" technique. This technique involved converting each instance in the datasets into six duplicated instances, with each instance assigned a unique label representing a single annotator. We will refer to this approach as "Learning from raw disagreement".

In our second approach, we trained six different models for each language, corresponding

to **each cohort based on age-gender combinations**. We utilized the individual votes of the annotators to train these models. To determine the final label for each instance, we combined the outputs of each cohort using various methods. Some methods were based on the labels themselves, while others utilized the probability distributions returned by the models:

- **Majority vote:** This approach determines the label by considering the majority vote among the six models. In case of a tie, we decided to set the label as "NO" in the first task. For the second task, the order of preference for tie-breaking is JUDGMENTAL > REPORTED > DIRECT.
- **Gender vote:** This approach adds, for each instance, the number of "votes" that each label receives from each gender's models. The label that receives more votes is selected. In case of a tie (two different labels obtain the same number of votes from the two genders' models), then the label selected by the "females" is returned as it has shown better results, as indicated in table 3.
- **Age vote:** This approach involves incorporating the "votes" received by each label from the models of each age group (18-22, 23-45, and 46 or more) for each instance. In each age group, we adopt the same decision rule as the previous method. In the case of a tie, the preferred label for the first task is "NO". For the second task, the order of preference is JUDGMENTAL > REPORTED > DIRECT.
- **Probability distribution mean:** This method computes the mean probability for each label and the six models, and the label with the highest mean probability is returned.
- **Probability distribution gender:** This method computes the most probable label for each gender by adding the scores of each gender's model. The label with the highest probability is selected. Since it is based on probability, it is assumed to be very difficult to have a tie-breaking situation.
- **Probability distribution Age:** This method computes the most probable label for each age range by adding the scores of each range's model. The label with the highest probability is selected.
- **Probability distribution cohort:** This method calculates the probability for each label/-cohort by considering the probability outputted by the model. The label with the highest probability is then selected and returned as the final prediction.

Finally, we employed all of these methods to train models using the augmented version of the EXIST 2023 train set and assessed their performance on the EXIST 2023 development set. Table 3 presents the results for Task 1.

Based on the obtained results, we have selected the outputs from the "Learning from raw disagreements" method as the first run for Task 1. This method has exhibited promising performance, and we will also employ its results to filter out instances labeled as NO-SEXIST, as it has demonstrated high accuracy in classifying this label.

As the second run for Task 1, we have submitted the "Gender vote" method. This approach consolidates the votes of the annotators from each gender to determine the final label for each instance.

For the third run in Task 1, we employ the outputs from the "Probability distribution (age)" method. This approach considers the probability distributions assigned by each age range cohort to determine the final label for each instance.

Table 3
TASK 1 evaluation results on the EXIST development dataset

Method	ICM hard-hard	F-measure hard-hard (YES)	F-Measure hard-hard (NO)	F-measure hard-hard (macro-F)	ICM hard-soft	ICM soft-soft
Learning from raw disagreements	0.5873	0.8068	0.8272	0.817	0.4393	0.7447
Majority vote	0.5841	0.8059	0.826	0.816	0.4563	-
Age vote	0.573	0.8	0.8246	0.8123	0.4214	-
Gender vote Male preference	0.5883	0.8091	0.8255	0.8173	0.4458	0.3913
Gender vote Female preference	**0.5942**	**0.8105**	0.8279	**0.8192**	**0.4758**	0.4125
Probability distribution	0.575	0.8029	0.8233	0.8131	0.4306	-
Probability distribution (mean)	0.575	0.8029	0.8233	0.8131	0.4306	0.7363
Probability distribution (age)	0.5785	0.8037	0.8247	0.8142	0.4274	-1.5735
Probability distribution (gender)	0.575	0.8029	0.8233	0.8131	0.4306	0.785
Probability distribution (cohort)	0.5841	0.8038	**0.828**	0.8159	0.4485	**0.8499**

A similar approach is followed for addressing Task 2. This is a multi-class classification Task instead of a binary classification one. Initially, tweets labeled as non-sexist (from Task 1) were eliminated using the approach that achieved the best F-Measure score for the "NO" (non-sexist) class, which involved using the Probability distribution cohort method. Subsequently, we repeated the previous steps to predict the source intention. The results of this approach evaluated on the development dataset are displayed in Table 4.

Table 4
TASK 2 evaluation results on the EXIST development dataset

Method	ICM hard-hard	FMeasure hard-hard (JUDGEMENTAL)	F-Measure hard-hard (NO)	FMeasure hard-hard (REPORTED)	F-measure hard-hard (DIRECT)	F-measure hard-hard (macro-F)	ICM hard-soft	ICM soft-soft
Learning from raw disagreement	0.2936	**0.3178**	0.828	0.3394	0.5945	0.5199	-6.8743	-
Majority vote	0.2936	**0.3178**	0.828	0.3394	0.5945	0.5199	-6.8743	-
Gender vote	0.2811	0.2812	0.828	0.3293	0.5951	0.5084	-6.6572	-
Age vote	0.3106	0.3015	0.828	0.3509	**0.6083**	0.5222	-6.7823	-
Probability distribution	0.29	0.2667	0.828	**0.3681**	0.5935	0.5141	-6.6209	-
Probability distribution (mean)	**0.3142**	0.2974	0.828	0.3659	0.5988	**0.5225**	-6.5681	-1.2093
Probability distribution (age)	0.279	0.2588	0.828	0.3294	0.5922	0.5021	**-6.4868**	-2.2366
Probability distribution (gender)	0.3021	0.3099	0.828	0.358	0.5937	0.5224	-6.8162	**-1.1792**

We have selected the outputs from the "Majority vote" method as the first run for Task 2, the "Probability distribution (mean)" method for the second run, and the "Probability distribution (gender)" method for the third run.

5. Evaluation and results

Table 5 provides a summary of the strategies employed in each of the runs that were ultimately submitted for both tasks.

For each of the tasks, the organization performed three types of evaluations:

- Hard-hard: the hard system output is compared against the hard ground truth.
- Hard-soft: the hard system output is compared against the soft ground truth.
- Soft-soft: the soft system output is compared against the soft ground truth.

For all tasks and evaluation types (hard-hard, hard-soft, and soft-soft), the official metric used is ICM (Information Contrast Measure) [24]. ICM is a similarity function that extends Pointwise

Table 5
Strategies employed in the runs submitted for Task 1 and Task 2

Task	Run	Method
Task 1	1	Repeated labeling
Task 1	2	Gender vote
Task 1	3	Probability distribution (cohorts)
Task 2	1	Majority vote
Task 2	2	Probability distribution (mean)
Task 2	3	Probability distribution (gender)

Mutual Information (PMI) and is employed to evaluate system outputs in classification problems by measuring their similarity to the ground truth categories. An extended version of ICM, known as ICM-soft, has been specifically developed for the task to accommodate both soft system outputs and soft ground truth assignments. The results of our three runs for Task 1 and Task 2, evaluated using the three types of evaluation, are presented in Tables 6, 7, 8, 9, 10, and 11. Each table provides details on various evaluation metrics for both Spanish and English languages, as well as the combined results for both languages. In each table, the first column within each column group indicates the ranking position of each run. The first and second rows of values in each table represent the gold and best results achieved for the respective task in each evaluation.

5.1. Task 1 - Sexism Identification

Regarding Task 1, as shown in Tables 6, 7, and 8, the third run utilizing the 'probability distribution (cohorts)' method outperformed the others in terms of hard metrics for all languages. However, the first run, based on 'repeated labeling,' achieved better results in the soft-soft evaluation. Our best approach secured the 19th position out of 57 participants in the hard-hard evaluation, the 17th position in the hard-soft evaluation, and the 12th position in the soft-soft evaluation. The higher ranking in the soft evaluations suggests that our approach effectively captures the different perceptions of sexism among distinct population cohorts.

Table 6
Results for the hard-hard evaluation for Task 1: This table presents the results obtained in the hard-hard evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM, Normalized ICM, and F-measure. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold	0	0.9948	1	1		0.9999	1	1		0.9798	1	1
Best score	1	0.6575	0.785	0.8109	1	0.6995	0.8011	0.8384	1	0.6004	0.7693	0.776
JPM_UNED_1	28	0.5057	0.6883	0.756	20	0.514	0.6783	0.7748	30	0.4819	0.6972	0.7308
JPM_UNED_2	33	0.4863	0.6759	0.7533	24	0.5016	0.6701	0.7784	39	0.4556	0.6812	0.7204
JPM_UNED_3	19	**0.5223**	**0.6989**	**0.7623**	14	**0.545**	**0.6988**	**0.7885**	29	**0.4844**	**0.6987**	**0.7284**

Table 7
Results for the hard-soft evaluation for Task 1: This table presents the results obtained in the hard-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft and ICM-Soft normalized. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL			ES			EN		
	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm
Gold	0	3.1182	1	0	3.1177	1	0	3.1141	1
Best score	1	1.1977	0.6897	1	1.3487	0.6892	1	0.9695	0.6905
JPM_UNED_1	19	0.1685	0.5235	17	0.3485	0.5135	31	-0.1243	0.5327
JPM_UNED_2	32	0.1075	0.5136	23	0.2927	0.5037	38	-0.1879	0.5235
JPM_UNED_3	17	**0.2041**	**0.5292**	13	**0.3927**	**0.5212**	28	**-0.0924**	**0.5373**

Table 8
Results for the soft-soft evaluation for Task 1: This table presents the results obtained in the soft-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft, ICM-Soft normalized and Cross entropy. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold	0	3.1182	1	0.5472	0	3.1177	1	0.5208	0	3.1141	1	0.577
Best score	1	0.903	0.6421	0.796	1	0.9527	0.6196	0.7672	1	0.8157	0.6683	1.0198
JPM_UNED_1	12	**0.6779**	**0.6058**	**0.8023**	18	0.6536	0.5671	**0.7588**	11	**0.6632**	**0.6463**	0.8512
JPM_UNED_2	18	0.5972	0.5927	**0.8852**	16	**0.6641**	**0.5689**	0.8116	15	0.4853	0.6207	0.9677
JPM_UNED_3	29	0.2467	0.5361	2.2342	31	0.1576	0.4799	2.1581	24	0.3506	0.6012	2.3196

5.2. Task 2 - Source Intention Identification

For the second task (categorizing the tweets according to the intention of the source), as shown in Tables 9, 10 and 11, the second run utilizing the 'probability distribution (mean)' method obtains the best results in this task and achieves the second position in hard-soft evaluation for Spanish. Over both languages, our best approach secured the 9th position in the hard-hard evaluation, the 22nd position in the hard-soft evaluation, and the 3rd position in the soft-soft evaluation.

Table 9
Results for the hard-hard evaluation for Task 2: This table presents the results obtained in the hard-hard evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM, Normalized ICM, and F-measure. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1	Rank	ICM-Hard	ICM-Hard Norm	F1
Gold	0	1.5378	1	1	0	1.6007	1	1	0	1.4449	1	1
Best score	1	0.4887	0.7764	0.5715	1	0.5711	0.7732	0.6059	1	0.3677	0.781	0.5224
JPM_UNED_1	11	0.1673	0.7079	0.5032	10	0.1986	0.6911	0.5281	12	0.1024	0.727	0.4661
JPM_UNED_2	9	**0.1862**	**0.712**	0.5054	8	**0.2351**	**0.6992**	0.5341	13	0.0995	0.7264	0.4649
JPM_UNED_3	10	0.1806	0.7108	**0.5092**	9	0.2231	0.6965	**0.5383**	11	**0.1034**	**0.7272**	**0.4673**

Table 10
Results for the hard-soft evaluation for Task 2: This table presents the results obtained in the hard-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft and ICM-Soft normalized. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL			ES			EN		
	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm	Rank	ICM-Soft	ICM-Soft Norm
Gold	0	6.2057	1	0	6.2431	1	0	6.1178	1
Best score	1	-2.3974	0.7803	1	-1.8502	0.7684	1	-3.265	0.7943
JPM_UNED_1	23	-7.5078	0.6498	21	-6.8073	0.6266	25	-8.9034	0.6707
JPM_UNED_2	22	**-7.3346**	**0.6542**	19	**-6.5622**	**0.6336**	23	**-8.8583**	**0.6717**
JPM_UNED_3	24	-7.5205	0.6495	22	-6.8533	0.6253	24	-8.8978	0.6708

Table 11
Results for the soft-soft evaluation for Task 2: This table presents the results obtained in the soft-soft evaluation. The three main columns represent the results for different languages: the first column shows results for all instances, the second column focuses on Spanish instances, and the third column focuses on English instances. Each column includes the following metrics for each run: ICM-Soft, ICM-Soft normalized and Cross entropy. The first row indicates the gold standard result, while the second row represents the best result. The remaining rows display the results obtained by our three evaluation runs.

RUN	ALL				ES				EN			
	Rank	ICM-Soft	ICM-Hard Soft	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Gold	0	6.2057	1	0.9128	0	6.2431	1	0.8926	0	6.1178	1	0.9354
Best score	1	-1.3443	0.8072	1.7833	1	-1.2317	0.7861	1.6415	1	-1.1471	0.8407	1.8001
JPM_UNED_2	3	**-1.675**	**0.7988**	**2.5549**	2	**-1.4414**	**0.7801**	**2.4472**	6	-2.1062	0.8197	2.6757
JPM_UNED_3	5	-1.6888	0.7984	2.5561	3	-1.5006	0.7785	2.4511	4	**-2.0436**	**0.8211**	**2.674**

6. Conclusions

This paper presents the participation of the JPM-UNED team in the Task 1 and the Task 2 of the EXIST 2023 Lab at CLEF, which focuses on the classification of sexism in social networks with disagreement. We have investigated different approaches to learning with disagreement, leveraging the current state of the art. Furthermore, we essayed different data augmentation techniques, such as incorporating translations of tweets between English and Spanish in the training set.

Among the approaches we explored, some of them proposed different variations of the judgement aggregation method, which combines the judgments or opinions of the multiple annotators and models to arrive at the final label. Notably, our best results were obtained in Task 2, where we secured the second position in the soft-soft evaluation for the Spanish language. This achievement was made possible by employing the Judgement Aggregation approach that leverages the viewpoints of the six different cohorts.

One limitation of our work stems from the size of the dataset. As some of our approaches involved splitting the dataset into six cohorts, the resulting training datasets were relatively small, which presented challenges in effectively training the models.

7. Acknowledgments

This research is funded by FAIRTRANSNLP-DIAGNOSIS: Measuring and quantifying bias and fairness in NLP systems, grant PID2021-124361OB-C32, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. This work has been also funded by the Ministry of Universities and the European Union through the EuropeaNextGenerationUE funds and the "Plan de Recuperación, Transformación y Resiliencia".

References

[1] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[2] A. R. G. Prasad, V. Seshagiri, L. Ravindranath, et al., A catalytic spectrophotometric method for the analytical determination of trace amounts of mercury (ii), Chem. Sci. Jour (2010) 1–8.

[3] S. Y. Jeong, Y. S. Koh, G. Dobbie, Phishing detection on twitter streams, in: Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2016 Workshops, BDM, MLSDA, PACC, WDMBF, Revised Selected Papers 20, Springer, 2016, pp. 141–153.

[4] J. Nogués Graell, Detection of toxicity in social media. a study on semantic orientation and linguistic structure (2022).

[5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[6] 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.

[7] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer, 2023, pp. 593–599.

[8] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.

[9] 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: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[10] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[11] O. Istaiteh, R. Al-Omoush, S. Tedmori, Racist and sexist hate speech detection: Literature

review, in: 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA), IEEE, 2020, pp. 95–99.
[12] H. Abburi, S. Sehgal, H. Maheshwari, V. Varma, Knowledge-based neural framework for sexism detection and classification., in: Proceedings of IberLEF@ SEPLN, 2021, pp. 402–414.
[13] L. Altin, H. Saggion, Automatic detection of sexism in social media with a multilingual approach, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). CEUR Workshop Proceedings Series, 2021, pp. 415–419.
[14] H. Abburi, P. Parikh, N. Chhaya, V. Varma, Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach, Data Sci. Eng. 6 (2021) 359–379.
[15] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: 5th Italian Conference on Computational Linguistics, CLiC-it 2018, volume 2253, CEUR-WS, 2018, pp. 1–6.
[16] L. I. Merlo, B. Chulvi, R. Ortega, P. Rosso, When humour hurts: linguistic features to foster explainability, 2023-03.
[17] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the em algorithm, J. Royal Stat. Soc. Series 28 (1979) 20–28.
[18] F. Rodrigues, F. Pereira, B. Ribeiro, Gaussian process classification and active learning with multiple annotators, in: Proceedings of ICML, 2014, pp. 433–441.
[19] P. Welinder, P. Perona, Online crowdsourcing: Rating annotators and obtaining cost-effective labels, in: IEEE CVPR Workshops, 2010, pp. 25–32.
[20] D. Reidsma, R. Akker, Exploiting 'subjective' annotations, in: Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics, 2008, pp. 8–16.
[21] V. S. Sheng, F. Provost, P. G. Ipeirotis, Get another label? improving data quality and data mining using multiple, noisy labelers, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 614–622.
[22] F. Rodrigues, M. Lourenço, B. Ribeiro, F. C. Pereira, Learning supervised topic models for classification and regression from crowds, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 2409–2422.
[23] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.
[24] 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), Dublin, Ireland, 2022, pp. 5809–5819.