Concatenated Transformer Models Based On Levels Of Agreements For Sexism Detection

Notebook for the EXIST Lab at CLEF 2024

Víctor Ruiz[1,*], Jorge Carrillo-de-Albornoz[1] and Laura Plaza[1]

[1]NLP & IR Group, Universidad Nacional de Educación a Distancia (UNED), 28040, Spain

Abstract

The EXIST 2024 lab aims to develop algorithms capable of detecting and classifying sexist messages. While previous EXIST tasks focused solely on sexism in tweets, the 2024 edition expands to include sexism in memes as well. The tasks involve identifying sexism, determining the intention behind the source, and categorizing the type of sexism, in both tweets and memes. This edition adopts the learning with disagreement paradigm (lately called perspectivism), meaning different annotations are provided for each instance in the datasets. Participants can submit two types of outputs for all tasks: hard labels, where the expected label is the majority label, and soft labels, where the expected results are the probability of each label for every instance. Our system employs a concatenation of various models. Initially, models are used to identify instances with a higher level of agreement. Once these instances are selected, another model generates predictions for those with a lower level of agreement. These models utilize different fine-tunings of an mDeBERTa-v3 model on the EXIST datasets. Our approach yielded strong and consistent results, achieving top rankings in Tasks 4 and 5.

Keywords

Sexism Detection, Learning with Disagreements, Large Language Models, Natural Language Processing

1. Introduction

The rise of social media as a dominant platform for communication has brought about unprecedented opportunities for social interaction and expression. However, this has also led to the proliferation of harmful behaviours, including sexism, which poses significant challenges to the promotion of safe and inclusive online environments. Sexism on social media manifests in various forms, ranging from overtly offensive language to subtle biases and microaggressions, affecting the well-being of individuals and perpetuating gender inequalities [1].

Addressing sexism on social networks is crucial not only for protecting users but also for upholding the integrity of online discourse. Traditional approaches to detecting harmful content have often relied on manual moderation or simplistic keyword-based algorithms, which are either unsustainable or insufficiently nuanced to capture the complexity of sexist language. Recent advances in natural language processing (NLP) and machine learning (ML) have paved the way for more sophisticated methods, yet challenges remain to effectively identify and categorise the multifaceted nature of sexism [2].

The EXIST challenges arise with the aim of defining a framework for researchers to develop systems capable of identifying sexism in text messages on social networks [3, 4]. This year, this challenge is held at CLEF 2024 and features several new developments compared to the EXIST competitions of previous years. The main novelty is the inclusion of a new dataset of memes, in English and Spanish, in which multimodal algorithms can be trained for classification [5]. Other tasks have also addressed the detection of sexism in images; however, EXIST 2024 is the first to be considered within the good practices of Perspectivism [6]. Since EXIST 2023, each instance in the EXIST datast has been annotated by six annotators from different genders and age groups [7]. There EXIST dataset is composed of two

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

[*] Corresponding author.

victor.ruiz@lsi.uned.es (V. Ruiz); jcalbornoz@lsi.uned.es (J. Carrillo-de-Albornoz); lplaza@lsi.uned.es (L. Plaza)
0000-0002-0877-7063 (V. Ruiz)

subsets: a dataset of tweets and a dataset of memes with six groups of annotators per cohort, to which demographic information about the annotators is added this year.

In this paper, we present the approaches proposed by the Victor-UNED team for the EXIST challenge at CLEF 2024. Our method aims to explore the impact of using different models to predict cases with lower inter-annotator agreement by applying a series-wise transformer-based approach. Instances where annotator agreement is not in the majority can be challenging for models to label accurately. In this paper, we compare the results of our three models. Firstly, a simple model that uses soft labels to train, secondly an approach which concatenates two models that filter cases depending on their level of agreement, and last a variation of the second approach in which the last model is an ensemble which contains cohorts differences to deal with cases with most disagreement. These systems establish new milestones in sexism detection strategies and have the potential to be extended with further research.

This paper is organised as follows. Section 2 is an overview of the state of the art on sexism detection and perspectivism. Section 3 presents EXIST 2024 tasks and datasets. In Section 4, we explain in detail our proposed methods. In Section 5, results are presented. Section 6 discussed such results. Finally, conclusions and future research directions are summarised in Section 7.

2. Related Work

For decades, the research community has focused on developing automatic methods to detect hate speech on social networks. Hate speech encompasses various forms of discrimination, including racism, xenophobia, and sexism. Traditionally, academic efforts targeted hate speech in general, employing methods ranging from rule-based systems to the more recent machine learning and deep learning algorithms [8]. However, significant differences exist among the types of hate speech. For instance, sexism can manifest not only as direct and explicit violent or hateful messages but also through trends, humor, or ideological expressions. Consequently, sexism requires distinct study and new approaches to address the diverse forms it can take. The EXIST challenges were established with this objective in mind. Over the past four years, the competition has introduced innovations to enhance the detection of sexism on social networks.

EXIST 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. First two editions, EXIST 2021 & 2022 were held in the IberLEF Spanish evaluation forum. These editions' dataset contained more than 11,000 tweets and gabs single labelled in two tasks: sexism identification, where texts can be sexist or non-sexist, and sexism categorization, where texts can be labelled as (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence, (v) misogyny and non-sexual violence. Results showed that transformer-based methods were the most implemented technique and ensembles of Large Language Models were the best performing models [3, 4].

Last year, EXIST 2023 published a new dataset with more than 10,000 tweets both in English and Spanish [7]. The difference with respect to the set from previous editions is that this new dataset follows perspectivism good practices. Perspectivism is a trend in dataset creation that encourages researchers to avoid aggregating annotations into gold standards and publish datasets with every annotation [6]. This trend is related with Learning With Disagreement paradigm, which argues that in subjective tasks, like sexism detection, disagreements help to represent actual opinions [9]. Models should learn from disagreements and provide a probability with their prediction, instead of reproducing biased opinions.

EXIST 2023 dataset's tweets were labelled by six groups of annotators according to gender (male, female) and age groups (18-22, 23-45, 46+). Different annotations per instance brought in disagreements in the annotations. Participants were encouraged to work with disagreements and provide soft labels, that is, probabilities for every label to be assigned. Results, again, showed the dominance of LLMs and how ensembles of transformer models can improve results in multiclass categories [10].

EXIST 2024 includes a new dataset of memes, together with previous edition tweets dataset, they form this year's provided data to train models [5]. Memes are images with text created with humoristic purposes [11]. They are shared on social networks, and sometimes they can contain hidden sexist behaviours hidden in a naive style. Memes have been studied from the hateful speech point of view [12] and even SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) [13] proposed examples of misogyny in memes. The aim is to create a memes' dataset for sexism detection that include Learning With Disagreement.

3. EXIST 2024

While previous editions of EXIST featured a dataset of tweets for textual classification, EXIST 2024 [5, 14] aims to expand the task and make it multimodal to cover a broader spectrum of sexism on social media. This year, the EXIST dataset includes a dataset of tweets and a new dataset of memes. Both datasets contain approximately half of the instances in English and the other half in Spanish.

The set of tweets corresponds to the EXIST 2023 dataset[7, 10], which consists of more than 3,200 tweets per language in the training set and approximately 500 tweets per language in the development set. Additionally, the test set contains about 1,000 tweets per language. On the other hand, the training set of memes contains more than 2,000 memes per language and approximately 500 memes for the test set.

Each instance in the dataset has been labelled by six annotators from different population groups by gender and age. The age groups are 18-22 y.o., 23-45 y.o., and over 46 y.o. The gender groups are male and female. Other genders were not considered due to the lack of individuals in the Prolific annotation application[1]. The intention of this annotation is to frame the competition within the Learning with Disagreements paradigm, in which the diversity of opinions is also included in the models through training with agreement levels. The level of agreement depends on the number of annotators who agree when labelling each instance. Additionally, this year the following demographic information of the annotators has been provided: country, race, and education level [5].

In EXIST 2024, six tasks are proposed: three targeting sexism in tweets and the same three tasks targeting sexism in memes. These three tasks are: 1) sexism detection, 2) source intention classification, and 3) sexism categorization. For each task, participants could participate with hard (gold) labels or with soft labels (probabilistic results).

- **Tasks 1 & 4 Sexism Detection**: This task is a binary classification in which cases are distinguished between those whose content is sexist, talks about sexist events, or is related to sexism, and those that are not. Task 1 involves tweets, while task 4 involves memes.
- **Tasks 2 & 5 Source Intention Classification**: This task is a multiclass classification between three classes in the case of memes: (i) direct, (ii) judgmental, and (iii) negative, or between four classes in the tweet dataset: (i) direct, (ii) judgmental, (iii) reported, and (iv) negative. Each class describes the intention of the tweet or meme's author when creating it, allowing us to study the roles of social networks in spreading sexist messages. Task 2 involves tweets, while task 5 involves memes.
- **Tasks 3 & 6 Sexism Categorization**: This task is a multiclass classification between six classes: (i) ideological and inequality, (ii) stereotyping and dominance, (iii) objectification, (iv) sexual violence, (v) misogyny and non-sexual violence, and (vi) negative. This task is also multilabel, meaning each annotator can choose more than one label for each instance. These labels are chosen to reflect the different aspect of a woman that can be attacked. Task 3 involves tweets, while task 6 involves memes.

It is worth noting that this competition takes into account the hierarchy between labels, so when predicting them, it will have a greater impact on the error to mistakenly predict positive instances as negative and vice versa than to mistake the types of intention or types of sexism [15].

[1] https://www.prolific.com/

4. System Description

Next, we will describe the systems we employed for each task in EXIST 2024. Although the competition comprises six tasks, they can be categorized into three main types: Sexism Identification, Source Intention, and Sexism Categorization, each applied to both the tweet dataset and the meme dataset. Our models are trained separately for each dataset, but we have developed three distinct types of models, one for each task category. Therefore, we will first explain the models used for tasks 1 and 4, followed by those for tasks 2 and 5, and finally those for tasks 3 and 6.

Our systems utilize a concatenation of models that filter instances based on their predicted soft labels. Soft labels represent the probabilities of each possible label for a given instance. These probabilities are calculated from the proportion of labels in the annotations. For example, if four out of six annotators indicate that a meme is positive while the other two annotate it as negative, the corresponding soft labels for that instance will be: YES: 0.67, NO: 0.33. We have trained our models using each instance along with their corresponding soft labels, so the output of these models will be the soft labels used as predictions.

The soft labels indicate the level of agreement among annotators for each case. A higher probability for one of the soft labels, combined with lower probabilities for the others, signifies greater agreement for that case. Agreement can favor any of the possible categories, whereas disagreement is characterized by a lack of clear majorities for any category. In this work, we will differentiate between types of instance classifications based on the level of agreement. Each instance has 6 annotations. We will define majority agreement as cases where 5 or 6 annotations are identical, and minority agreement as cases where 3 or 4 annotations are identical.

A point to highlight is that in the EXIST tasks, the hierarchy of predictions is key. In the Source Intention and Sexism Categorization tasks, the metric used for evaluation, ICM, penalizes misclassifications between negatives and a type of positives much more than misclassifications among different types of positives. For this reason, it is important to correctly determine the results for the first task and transfer these results to the other tasks.

Based on the state-of-the-art in offensive language detection, the results of previous editions of EXIST, and our own experience testing different models, we decided to use mDeBERTa-v3 to train our systems. This latest version of the mDeBERTa architecture improves upon BERT and RoBERTa by implementing disentangled attention and an enhanced mask decoder [16]. Additionally, this model demonstrates remarkable performance in tasks involving both English and Spanish text. Using a multilingual model instead of two monolingual ones allows us to access twice the number of training cases, as we can train with the entire dataset regardless of the language, rather than only half of it in one language.

4.1. Tasks 1 & 4

For Tasks 1 and 4, our hypothesis posits that a higher level of agreement among annotators increases the likelihood that an algorithm will predict the majority label. Therefore, additional effort is needed to develop models capable of distinguishing cases with low levels of agreement. We propose a two-part prediction process. First, a model generates predictions for every instance in the dataset. We then filter out cases where the predicted soft labels indicate high agreement, meaning the soft labels correspond to approximately five or more identical labels in the same category. These high-agreement instances are separated from the rest of the set. Next, a second model, ideally a more effective classifier for low-agreement cases (trained with more data, a different model, etc.), generates predictions for this subset. This approach was employed in models for Task 1 and Task 4.

System 1. Baseline. In our experiments, the mDeBERTa-v3 model consistently delivered the best results. For Task 1, we observed optimal performance when the model was trained exclusively on the EXIST 2024 tweets training dataset. For Task 4, superior results were achieved by training the model on both the tweets and memes datasets. Attempts to enhance the model by augmenting the dataset with EXIST 2021 data or reconfiguring the dataset to treat each case as if it had a unique label resulted in poorer performance on the development dataset. The final model was trained using the soft labels from the EXIST 2024 training set and applied directly to generate the solutions.

In the case of Task 1, fine-tuning was done using the tweet dataset, while for Task 4, both the meme and tweet datasets were used. This is because, in the tests, increasing the dataset in Task 4 helped improve the results. These datasets will also be used to train the models for the rest of the runs of Tasks 1 and 4.

System 2. Detection of cases with high agreement. This system has two steps: 1) prediction of cases with majority consensus, 2) prediction of the remaining cases. In the first step, mDeBERTa-v3 was trained to detect cases with a majority of consensus among annotators. The cases detected as majority consensus are removed from the test dataset, and in the second step, another mDeBERTa-v3 model fine-tuned on the training set is applied to predict the labels of the remaining cases.

In Task 1, Run 2 involves employing two models: the first model distinguishes positive cases with majority agreement, while the second model identifies the remaining cases. This approach addresses a slight imbalance towards negative cases in Task 1. Utilizing the first model to discriminate positives resulted in marginal enhancements in prediction accuracy.

In contrast, Task 4 entails a three-step detection process. The first step discriminates negatives with majority agreement, followed by the second step which identifies positives with majority agreement, and finally, the third model predicts the remaining cases. This sequence of models was found to be more effective due to the slight imbalance towards negatives in the training dataset, primarily stemming from the contribution of the tweet dataset (10,000). Notably, the meme dataset exhibits an imbalance towards positives in both the training and development sets, although the distribution of the test set remains unknown.

System 3. Detection of high agreements using demographic information . This system has two steps: 1) prediction of cases with majority consensus, 2) prediction of the remaining cases through the implementation of an ensemble model that contains information for each group of annotators. First, we use a model based on mDeBERTa-v3 to detect cases with majority consensus among annotators. This model identifies extreme agreements, both negative and positive, in the same step based on the value of their predicted soft labels. To predict the labels of cases with greater disagreement, we use an ensemble of six models, one for each cohort of the dataset for the corresponding task.

4.2. Tasks 2 & 5

For Tasks 2 and 5, our goal is to transfer the cases detected as negative in the first task and train an algorithm to capture the proportion of the remaining labels. A key advantage of this task is the constraint that the value of the soft labels for each case must sum to 1. These tasks are multiclass detection tasks. As we previously explained, Task 2 has 4 possible labels, while Task 5 has 3. Additionally, the level of imbalance in these tasks is much higher than in the previous ones: approximately half of the cases will be negative (not sexist), and failing to predict them correctly significantly affects the evaluation metric. Therefore, in the end, we adjust the model predictions to the results obtained in Tasks 1 and 4.

In these tasks, our objective was to compare two methods. The first approach trains and predicts by including the negatives, and then it adjusts predictions to match the results of Tasks 1 and 4. The second approach trains and predicts using the remaining labels without negatives, and at the end they add labels from negatives obtained from Tasks 1 or 4 and adjusts the rest of the labels to math these results.

System 1. Training and prediction with negatives. We trained a model based on mDeBERTa-v3 on the dataset for Tasks 2 and 5, keeping the negatives. At the end, the predictions for the negative label are replaced by the label from the first task for the same case, and the rest of the labels are adjusted according to their predictions so that their sum equals 1.

System 2. Training and prediction without negatives. We trained a model based on mDeBERTa-v3 on the dataset for Tasks 2 and 5, but without the negatives. At the end, the predictions for the negative label from the first task are added, and the rest of the labels are adjusted according to their predictions so that their sum equals 1.

4.3. Tasks 3 & 6

In Tasks 3 and 6, we apply a similar procedure to the previous tasks. We transfer the cases detected as negative in the first task and train an algorithm to capture the proportion of the remaining labels. However, in this case, the soft labels for each case can sum to more than 1, so we do not have the constraint from Task 2.

These tasks are multiclass and multilabel detection tasks. As we previously explained, Tasks 3 and 6 have 6 possible labels. The level of imbalance in these tasks is also high because approximately half of the cases are negative (not sexist). The rest of the classes have different distributions, although the differences in proportions between them are smaller. Since this task is multilabel, the sum of the soft labels for each category per case does not necessarily have to equal one.

Again, in this task, we aimed to evaluate whether it was more convenient to train and predict by including the negatives, although later adjusting to the results of Tasks 1 and 4, or if, on the other hand, it was preferable to train and predict using the remaining labels without negatives and add them at the end.

System 1. Training and prediction with negatives. We trained a model based on mDeBERTa-v3 on the dataset for Tasks 3 and 6, keeping the negatives. At the end, the predictions for the negative label are replaced by the label from the first task for the same case.

System 2. Training and prediction without negatives. We trained a model based on mDeBERTa-v3 on the dataset for Tasks 3 and 6, but without the negatives. At the end, the predictions for the negative label from the first task are added.

5. Results

Following, we will present the results of the systems on the test set. For all tasks, results with both hard labels and soft labels have been provided, except for tasks 3 and 6, where only soft labels have been provided.

The ranking metric utilized is ICM (Information Contrast Measure), developed by the organizing team. ICM is a similarity function grounded in information theory, extending Pointwise Mutual Information (PMI) to assess system outputs in classification problems. This metric has variants tailored for both hard labels and soft labels.

The results of the systems are shown in Tables 1-6. In addition to the model results, the baselines published by the competition for each task are added to the tables.

Gold standard: It is the ideal score obtained if our system's output exactly matched the annotations of the test set.

Majority Class: It is the score obtained by a model if it predicted using only the category that appears most frequently in the task.

Minority Class: It is the score obtained by a model if it predicted using only the least frequent category in the task.

5.1. Task 1 & Task 4

5.1.1. Task 1

In Table 1, we find the results for the sexism detection task in the tweet dataset. Among the results with soft labels, the best-performing model was System 1, which is consistent with our previous analyses. The positions of our models in the ranking are, respectively: System 1, 11th; System 2, 14th; and System 3, 17th. The results improve slightly when tested only on the Spanish dataset, while the scores slightly decrease in the English dataset. Regarding the results with hard labels, we find that in the ranking, our models score consecutively: System 3, 24th; System 1, 25th; and System 2, 26th. It is noteworthy that System 3 achieves better results in the English dataset than the other two models, while it performs worse in the Spanish dataset.

Our approaches obtained better positions in the ranking in soft labels evaluation than in hard evaluation. That was expected. Learning With Disagreements surveys [9] have shown that introducing disagreement in the labels may affect negatively to hard label evaluation, whereas in soft label evaluation it produces a big improvement with respect to models trained with hard labels. These changes depend on the level of agreement of the dataset, so the more annotators agree between them, the better the performance of soft labels models in hard labels evaluation. However, we are dealing with a task extremely subjective, whose annotations do not have majority consensus.

In this task, there were 38 runs for soft evaluation and 68 runs for hard evaluation. Our best model, VictorUNED_1, is 0.0640 away of the best model in Task 1 soft evaluation, NYCU_NLP_1. The same model, NYCU_NLP_1, is the best model in Task 1 hard evaluation, but this time our best model, VictorUNED_3, has a difference of 0.0522. Differences were lower in the case of hard evaluation.

Table 1
Results of the runs submitted for Task 1 (soft and hard labels).

	Rank Soft	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank Hard	ICM-Hard	ICM-Hard Norm	F1
EXIST2024-test_gold	0	3.1182	1.0000	0.5472	0	0.9948	1.0000	1.0000
EXIST2024-test_majority-class	36	-2.3585	-2.3585	0.1218	68	-0.4413	0.2782	0.0000
EXIST2024-test_minority-class	40	-3.0717	0.0075	5.3572	70	-0.5742	0.2114	0.5698
Best Model	1	1.0944	0.6755	0.9088	1	0.5973	0.8002	0.7944
VictorUNED_1	**11**	**0.6952**	**0.6115**	1.0691	25	0.4914	0.7470	0.7542
VictorUNED_2	14	0.6797	0.6090	**0.9818**	26	0.4863	0.7444	0.7535
VictorUNED_3	17	0.6479	0.6039	1.0930	**24**	**0.4934**	**0.7480**	**0.7602**

5.1.2. Task 4

In comparison to other participants, the results based on soft labels have been notably strong, with System 1 and System 2 models securing the first and second positions in the ranking, respectively, and the System 3 model achieving the seventh position. This success can be attributed to the consistent performance of the models across both Spanish and English datasets. While individually the models may not excel in either language, their high rankings across both languages contribute to their overall performance. Notably, results on the English dataset surpass those on the Spanish dataset by several points, a trend observed in other models as well.

In contrast, results based on hard labels have yielded lower rankings: System 2 occupies the 7th position; System 1 got the 13th position; and System 3, the 17th position. These results indicate that there is potential for improvement, both in English and Spanish.

In this task, there were 36 runs for soft evaluation and 51 runs for hard evaluation. Our model VictorUNED_1 is the best model in Task 4 soft evaluation. However, the model RoJiNG-CL_3 is the best model in Task 4 hard evaluation and has a difference of 0.1095 with VictorUNED_2. Differences in the hard evaluation were higher than in the case of first task, even though they have approximately the same difference in F1 (about 0.05).

Table 2
Results of the runs submitted for Task 4 (soft and hard labels).

	Rank Soft	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank Hard	ICM-Hard	ICM-Hard Norm	F1
EXIST2024-test_gold	0	3.1107	1.0000	0.5852	0	0.9832	1.0000	1.0000
EXIST2024-test_majority-class	36	-2.3568	0.1212	4.4015	39	-0.4413	0.2782	0.0000
EXIST2024-test_minority-class	38	-3.5089	0.0000	5.5672	46	-0.5742	0.2114	0.5698
Best Model	1	-0.2925	0.4530	1.1028	1	0.3182	0.6618	0.7642
VictorUNED_1	**1**	**-0.2925**	**0.4530**	**1.1028**	13	0.0641	0.5326	0.7051
VictorUNED_2	2	-0.3135	0.4496	1.2834	**7**	**0.1028**	**0.5523**	**0.7154**
VictorUNED_3	7	-0.3761	0.4395	1.1562	17	0.0364	0.5185	0.6991

5.2. Task 2 & Task 5

5.2.1. Task 2

In Table 3 we find the results for the source intention classification task in the tweet dataset. The results of the soft labels have reached the fifth and sixth positions in the ranking for System 2 and System 1, respectively. The results of the hard labels have been lower in the ranking: System 1 occupies the 19th position, and System 2 occupies the 20th position. The results are quite consistent between languages, with hardly any variations for both soft labels and hard labels.

In this task, there were 33 runs for soft evaluation and 48 runs for hard evaluation. Our best model, VictorUNED_2, is 0.1120 away of the best model in Task 2 soft evaluation, NYCU_NLP_2. In Task 2 hard evaluation, ABCD Team_1 is the best model and has a difference with our best model, VictorUNED_1, of 0.1043. Differences were slightly lower in the case of hard evaluation, even though we obtain a better position in the ranking. Also, differences with respect to the best results were higher than in the first task.

Table 3
Results of the runs submitted for Task 2 (soft and hard labels).

	Rank Soft	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank Hard	ICM-Hard	ICM-Hard Norm	F1
EXIST2024-test_gold	0	6.2057	1.0000	0.9128	0	1.5378	1.0000	1.0000
EXIST2024-test_majority-class	27	-5.4460	0.0612	4.6233	39	-0.9504	0.1910	0.1603
EXIST2024-test_minority-class	35	-32.9552	0.0000	8.8517	46	-3.1545	0.0000	0.0280
Best Model	1	-0.2543	0.4795	1.8344	1	0.4059	0.6320	0.5677
VictorUNED_1	6	-1.6549	0.3667	1.8132	**19**	**0.0851**	**0.5277**	**0.3257**
VictorUNED_2	**5**	**-1.6440**	**0.3675**	**1.7971**	20	0.0815	0.5265	0.3256

5.2.2. Task 5

In Table 4 we find the results for the second task in the meme dataset. In the case of soft labels, the second model has obtained very good results, achieving the top position in the ranking. On the other hand, the first model has obtained the 7th position. In the case of hard labels, once again, our first model has been the best model in the subtask while our second model occupies the third position. As in task 4, the results, for both soft and hard labels, are slightly higher results in ICM Norm in the English dataset.

In this task, there were 16 runs for soft evaluation and 20 runs for hard evaluation. Our two models reached top positions, each one in a different evaluation. VictorUNED_2 was the best model in soft evaluation, with a difference of 0.0022 with respect to the second model. VictorUNED_1 was the best model in soft evaluation, with a difference of 0.0048 with respect to the second model.

Table 4
Results of the runs submitted for Task 5 (soft and hard labels).

	Rank Soft	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank Hard	ICM-Hard	ICM-Hard Norm	F1
EXIST2024-test_gold	0	4.7018	1.0000	0.9325	0	1.4383	1.0000	1.0000
EXIST2024-test_majority-class	14	-5.0745	0.0000	5.5565	17	-1.0445	0.1369	0.1839
EXIST2024-test_minority-class	18	-18.9382	0.0000	8.0245	21	-2.0637	0.0000	0.0697
VictorUNED_1	7	-2.0053	0.2867	2.0028	**1**	**-0.2397**	**0.4167**	**0.3873**
VictorUNED_2	**1**	**-1.2453**	**0.3676**	**1.6235**	3	-0.2668	0.4073	0.3850

5.3. Task 3 & Task 6

5.3.1. Task 3

Table 5 shows the results for the sexism categorization task in the tweet dataset. In this task, we have only competed with soft labels, obtaining the 15th and 16th positions in the ranking. The results by language are consistent with each other, slightly higher in the English dataset.

In this task, there were 31 runs for soft evaluation and 32 runs for hard evaluation. Our best model, VictorUNED_2, is 0.2333 away of the best model in Task 3 soft evaluation, NYCU_NLP_1. Differences with respect to the best results were higher than in the previous tasks.

Table 5
Results of the runs submitted for Task 3 with soft labels on the whole test set.

	Rank Soft	ICM-Soft	ICM-Soft Norm
EXIST2024-test_gold	0	9.4686	1.0000
EXIST2024-test_majority-class	28	-8.7089	0.0401
EXIST2024-test_minority-class	33	-46.1080	0.0000
Best Model	1	-1.1762	0.4379
VictorUNED_1	**15**	**-5.5936**	**0.2046**
VictorUNED_2	16	-5.6190	0.2033

5.3.2. Task 6

In Table 6 we find the results for the third task in the meme dataset. Once again, in this task, we have only competed with soft labels, obtaining the 6th and 7th positions in the ranking for our System 1 and System 2 models, respectively. The results by language are consistent with each other, slightly higher in the English dataset.

In this task, there were 20 runs for soft evaluation and 23 runs for hard evaluation. Our best model, VictorUNED_1, is 0.0860 away of the best model in Task 6 soft evaluation, ROCurve_1.

Table 6
Results of the runs submitted for Task 6 (soft labels).

	Rank Soft	ICM-Soft	ICM-Soft Norm
EXIST2024-test_gold	0	9.4343	1.0000
EXIST2024-test_majority-class	13	-9.8173	0.0000
EXIST2024-test_minority-class	22	-50.0353	0.0000
Best Model	1	-4.7893	0.2462
VictorUNED_1	**6**	**-6.4124**	**0.1602**
VictorUNED_2	7	-6.4777	0.1567

6. Discussion

As we have noted in the previous section, even though the same three tasks were repeated in two different datasets, results from tweets dataset and memes dataset highly differ between them. From models' results in the ranking, we can state that the memes' dataset added extra difficulties to the tasks. One of the reasons may be that annotators judge both text and image when annotating memes. In our case, we have only used the textual part because of its simplicity and because we realized that it captures more meaning that only using images.

Remarkably, our most successful outcomes in the competition were achieved with the memes' dataset, with three of our models ranking among the top performers across six subtasks. This outcome supports our belief in prioritizing text as the primary element for content analysis in memes. However, it is crucial to recognize that text alone does not encompass the entirety of meme content. Despite our systems achieving superior positions in the memes' dataset, our results in the tweets dataset surpassed those in the memes' dataset. Consequently, limiting predictions to text alone, when the annotation process considers additional elements, may constrain the model's performance potential.

We put an especial effort in defining appropriate models for Tasks 1 and 4. Experiments with mDeBERTa-v3 showed that this model reached better results than other multilingual transformer-based models like XLM-RoBERTA or multilingual DistilBERT. Experiments were aimed to find a model that could obtain an increase in the prediction of soft labels when agreement was not majority, however, when predicting on soft labels mDeBERTa trained on soft labels has been the best model, and results in the competition confirmed it again. Even though we expected that our approaches would obtain better results in soft evaluation, some of the most interesting results were given by hard evaluation. Our approaches that concatenate models have obtained better results in hard labels evaluation.

As the next step of our concatenated systems, our objective was to tailor the output of a system predicting between second task categories to the results obtained in the preceding task. It is evident that a binary classification approach would likely yield superior results compared to a multiclass classification. Thus, leveraging our new predictions based on the previous ones was anticipated to yield positive outcomes. And indeed it did. Our models achieved higher rankings in Task 2 and Task 5 compared to the others, with top rankings in Task 5.

It's noteworthy that models achieving better results in soft labels were trained without negative labels, while those performing better in hard labels were trained with negative labels. Replicating similar experiments could shed light on how such differences in training impact output variations.

We tried to apply a similar method to Sexism Categorization task, however, results were not as promising. This task demands a more detailed analysis of the corpus and its labelling. Probably adding a method that could capture more information, maybe more information from images in the case of the memes' dataset or vocabulary in tweets, could end in a fine-grained classification of instances.

In general, our results showed very little variance between results in Spanish and English. Since the dataset is balanced between both languages, and the model we chose, mDeBERTa, shows very good results in both language, even in some of them we have obtained better results in Spanish.

7. Conclusions

In this paper, we present the approaches of Victor-UNED team for all tasks of EXIST at CLEF 2024. This challenge encourages researchers to develop algorithms to detect sexism in social networks with two datasets, a textual one from tweets and another one with memes, composed by text and images. Instances in the dataset were labelled by six different annotators, introducing disagreements in the annotations. Thus, there are two ways of providing predictions: soft labels, that is, probabilities of every possible label; and hard labels, which are unique labels per instance.

Our aim in this work was to study the effect of models where low agreement is met. We have built models that predict series wise labels, from higher agreement to lower agreement. From the state of the art reached in previous EXIST editions, we have studied which transformer-based model was the most suitable for the tasks, leading to the use of mDeBERTa-v3. We have trained our models on soft labels, and we have explored different data augmentation techniques and multiple embedding forms.

Finally, our systems have stood out especially in the memes' dataset, where they have achieved first position in the ranking for three subtasks. For Tasks 1 and 4 our proposed custom models, System 2 and 3, have obtained better results than our mDeBERTa trained in soft labels. For Tasks 2 and 5 models trained without negative labels obtained better results in soft labels results. Further research is needed for Tasks 3 and 6, to implement more information, especially in memes.

References

[1] S. Dimitry, A. Murphy, Beyond# metoo: The gender equity iceberg, 2017.

[2] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "call me sexist, but..." : Revisiting sexism detection using psychological scales and adversarial samples, Proceedings of the International AAAI Conference on Web and Social Media 15 (2021) 573–584. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18085. doi:10.1609/icwsm.v15i1.18085.

[3] 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.

[4] 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.

[5] L. Plaza, J. Carrillo-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.

[6] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 6860–6868. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25840. doi:10.1609/aaai.v37i6.25840.

[7] 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. Experimental IR Meets Multilinguality, Multimodality, and Interaction, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[8] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Surv. 51 (2018). URL: https://doi.org/10.1145/3232676. doi:10.1145/3232676.

[9] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, J. Artif. Int. Res. 72 (2022) 1385–1470. URL: https://doi.org/10.1613/jair.1.12752. doi:10.1613/jair.1.12752.

[10] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023 – Learning with Disagreement for Sexism Identification and Characterization (Extended Overview), in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023.

[11] S. Sharma, F. Alam, M. S. Akhtar, D. Dimitrov, G. D. S. Martino, H. Firooz, A. Halevy, F. Silvestri, P. Nakov, T. Chakraborty, Detecting and understanding harmful memes: A survey, 2022. arXiv:2205.04274.

[12] R. L. Tamayo, B. Chulvi, P. Rosso, Everybody hurts, sometimes overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter, Procesamiento del Lenguaje Natural 71 (2023) 383–395.

[13] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[14] L. Plaza, J. Carrillo-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.

[15] E. Amigo, A. 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399. doi:10.18653/v1/2022.acl-long.399.

[16] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. arXiv:2111.09543.