# Integrating Annotator Information in Transformer Fine-tuning for Sexism Detection

SINAI participation at EXIST Lab in CLEF 2023

María Estrella Vallecillo-Rodríguez[1], Flor Miriam Plaza-del-Arco[2], Luis Alfonso Ureña-López[1], María Teresa Martín-Valdivia[1] and Arturo Montejo-Ráez[1]

[1]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain
[2]Bocconi University, Via Sarfatti, 25, Milan, 20100, Italy

**Abstract**

This paper describes the participation of SINAI research team in the sEXism Identification in Social neTworks (EXIST) Shared Task at CLEF 2023. Specifically, we participated in Task 1 (sexism identification), Task 2 (sexism intention), and Task 3 (sexism categorization). For the three tasks, we propose three different systems, one based on a fine-tuning of a transformer pretrained model with hard labels, another exploring a data augmentation strategy, and the third system that integrates socio-demographic annotators features in order to check if this information helps in detecting sexism content. In EXIST shared task, the organizers propose three evaluation methods, Hard-Hard evaluation where they compare the hard label from the output of the system with the hard label of the ground truth, Hard-Soft evaluation which consists in evaluate the hard label from the output of the system with the soft labels of the ground truth and Soft-Soft where they compare the soft labels of the system with soft labels of the ground truth. Our team ranked 1st in Task 1 for Soft-Soft evaluation method, 8th in Task 2 for Soft-Soft evaluation, and 7th in Task 3 with Hard-Hard evaluation among the participants, achieving 0.903, -2.29, and 0.1472 of the ICM metric, respectively.

**Keywords**

Sexism detection, Text classification, Transformers, Natural language processing, Social media

## 1. Introduction

According to The Oxford Dictionary, sexism is the "prejudice, stereotyping, or discrimination, typically against women, on the basis of sex" [1]. In our daily lives, sexism manifests itself when people undervalue the views expressed by women, whether in spoken or written conversations containing fixed sayings and expressions. Today, with the prevalence of social media, sexist comments have become alarmingly prevalent, spreading rapidly and driving more instances

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

*Corresponding author.

†These authors contributed equally.

mevallec@ujaen.es (M. E. Vallecillo-Rodríguez); flor.plaza@unibocconi.it (F. M. Plaza-del-Arco); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia); amontejo@ujaen.es (A. Montejo-Ráez)
0000-0001-7140-6268 (M. E. Vallecillo-Rodríguez); 0000-0002-3020-5512 (F. M. Plaza-del-Arco); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. T. Martín-Valdivia); 0000-0002-8643-2714 (A. Montejo-Ráez)

of sexism. Moreover, identifying these comments can be challenging due to the various ways in which they can be expressed. To address these challenges, the scientific community has established numerous academic events and shared tasks. These initiatives aim to address specific issues related to the detection and classification of sexism. For example, EVALITA [2] and AMI [3] focus on the identification of misogyny, while HateEval [4] focuses on the detection of hate speech directed against women and immigrants. In addition, shared tasks such as EDOS [5] aim to develop more accurate and explainable systems for sexism detection, and EXIST [6, 7] attempts to classify sexism according to the different facets of women that are affected. The efforts of increasing the scope of sexism detection, contribute to a more complete understanding of the different types of sexism and how they are expressed.

This paper describes the participation of SINAI in sEXism Identification in Social neTworks (EXIST) shared task [8, 9] at CLEF 2023. This task aimed to capture sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviors. The main purpose of this task is to contribute to developing applications that can detect sexism. For this purpose, the organizers propose three different tasks. Task 1: Sexism Identification is related to identifying if a comment is sexist or not. Once a comment is classified as sexist, Task 2: Source Intention consists of classifying the intention of the author with this sexist comment, and Task 3: Sexism Categorization, when a comment is classified as sexist, to try to detect which facets of women are attacked with this comment. An important proposal of the task is *"The learning with disagreement paradigm"* where the organizers propose to build systems that are able to consider the different perspectives that people have when identifying sexism. For this reason, task organizers propose three evaluation methods (Hard-Hard, Hard-Soft, and Soft-Soft) that are explained in Section 4.2 Our team SINAI has participated in the three tasks.

Our proposal for addressing EXIST task is the integration of the different features of the annotators in systems that detect sexism to gather the diversity of subjective views in sexism detection. We expect to obtain a more accurate prediction for this specific task. There are some previous works that try to integrate external knowledge into systems to obtain better results. For example, in misogyny detection task, Frenda et al. [10] introduces an approach based on aesthetic features captured by character n-grams, sentiment information, and a set of lexicons built by analyzing misogynistic tweets. Nevertheless, misogyny detection is not the only task where researchers try to integrate extra information. In hate speech detection there are some proposals that use a multi-task learning paradigm to combine different phenomena that are inextricably related to the expression of offensive language such as sentiments, emotions, target, irony, sarcasm, and constructiveness, among others [11, 12, 13]. They show that the integration of these features helps the detection of hate speech. Finally, Pérez et al. [14] evaluated the impact of incorporating contextual information in hate speech related to the news posted on social media. This study shows evidence that adding contextual information improves hate speech detection performance for systems that perform binary and multi-label prediction tasks.

The rest of the paper is structured as follows: In Section 2 we describe the different strategies used to develop the systems for the shared task. The used data and the methodology followed to achieve the goal of the task are described in Section 3. The results obtained in our experiments during the development phase and the evaluation phase are shown in Section 4. Finally, we conclude with a discussion in Section 5.

## 2. System Overview

Sexism identification has been approached as a text classification task. We propose two different architectures, namely *base architecture* (BA) and *integrating annotator's information* (IAI). The *integrating annotator's information* solution takes into account information related to the annotators of the texts with the aim of allowing the system to consider the different perspectives of the annotators.

### 2.1. Base Architecture

The basic architecture is a Transformers architecture [15] for text classification. In this architecture, we have texts that are tokenized and then passed to the model. The model preprocesses the input and generates an output. To perform text classification, we get the classification token from the last hidden state of the model. This token is intended to encode the whole input sequence. This token is then passed as input to a feed-forward network, that classifies the text and produces the result of this classification in the number of classes that each specific task demands. This architecture is followed for two proposed experiments called *Baseline* and *Baseline with Data Augmentation.*

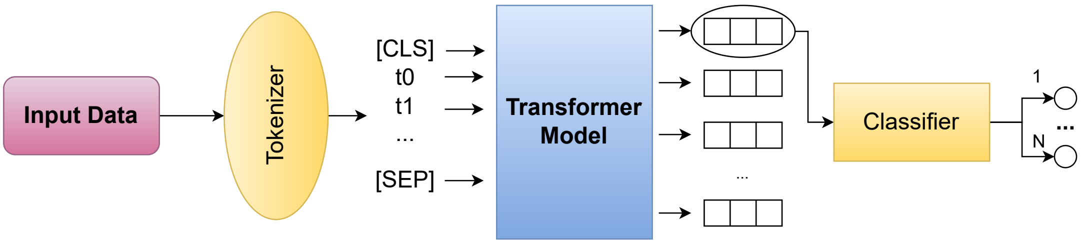
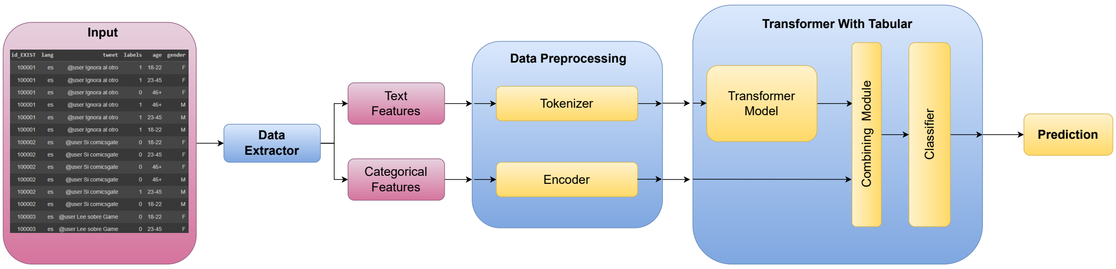


**Figure 1:** Base architecture proposed for EXIST shared task. N represents the number of output nodes and depends on the task because it corresponds to the number of labels to classify.

### 2.2. Integrating Annotator's Information

This design is a multimodal architecture called *Transformer With Tabular* [16] that received a DataFrame as input. The DataFrame is preprocessed, extracting the columns that represent the text to classify, categorical information and numerical information. Later, text features are tokenized, the categorical features are encoded and the numerical features are transformed to an adequate format. Then, text features are passed to a Transformer model. The Transformers model output is passed to a combining module that incorporates the numerical and categorical features to generate a unique tensor that is passed to a classifier, so final predictions are generated. In our case, we do not have numerical features, so they are not represented in Figure 2, which shows this architecture. The experiment called *Integrating Annotators' Information* implements this architecture.

The combining module of the architecture represents the strategy to combine the different features. In this case, we have explored the following methods:



| id_EXIST | lang | tweet | labels | age | gender |
|---|---|---|---|---|---|
| 100001 | es | @user Ignora al otro | 1 | 18-22 | F |
| 100001 | es | @user Ignora al otro | 1 | 23-45 | F |
| 100001 | es | @user Ignora al otro | 0 | 46+ | F |
| 100001 | es | @user Ignora al otro | 1 | 46+ | M |
| 100001 | es | @user Ignora al otro | 1 | 23-45 | M |
| 100001 | es | @user Ignora al otro | 1 | 18-22 | M |
| 100002 | es | @user Si comicsgate | 0 | 18-22 | F |
| 100002 | es | @user Si comicsgate | 0 | 23-45 | F |
| 100002 | es | @user Si comicsgate | 0 | 46+ | F |
| 100002 | es | @user Si comicsgate | 0 | 46+ | M |
| 100002 | es | @user Si comicsgate | 1 | 23-45 | M |
| 100002 | es | @user Si comicsgate | 0 | 18-22 | M |
| 100003 | es | @user Lee sobre Game | 0 | 18-22 | F |
| 100003 | es | @user Lee sobre Game | 0 | 23-45 | F |

**Figure 2:** Proposed architecture to incorporate annotator information for EXIST shared task.

- **Concat**. Concatenate transformer model output and categorical features before the classifier.
- **Multi Layer Perceptron (MLP) on categorical features then concat**. We apply MLP on categorical features then concat transformers model output and processed categorical features before the classifier.
- **Attention on categorical features**: Attention-based summation of transformer outputs, and categorical features queried by transformer outputs before classifier.
- **Gating on categorical features then sum**. Gated summation of transformer outputs, and categorical features before classifier.
- **Weighted feature sum on transformer and categorical features**: Learnable weighted feature-wise sum of transformer outputs, and categorical features for each feature dimension before classifier.

# 3. Experimental Setup

## 3.1. Data

To run our experiments, we use the dataset provided by the organizers. The dataset is composed of comments extracted from Twitter that can contain sexist popular expressions and terms both in English and Spanish. This dataset is labeled by a total of 6 individuals with different socio-demographic characteristics, such as gender (male or female) and age (18-22, 23-45, or +46). Moreover, in this dataset, the organizers, instead of providing a gold label for each text, give the participants the label assigned by each annotator and their personal information such as the gender and age for all tasks. The objective of the organizers with this information is to gather the diversity of views in a subjective task like sexism detection. A set of 10,034 tweets are annotated as sexist or not sexist and sexist posts are designated with more specific labels related to the intention of the author of the tweet and the category of the sexism. Table 1 shows the dataset size in each split. For task 1 we have to classify if a text is sexist or not (*YES, NO*). For task 2 we have to classify the intention of the author for sexist comments (*NO, DIRECT, REPORTED, JUDGEMENTAL*). For task 3, we have to detect which facets of women are more attached. As we can see in this task we can select one or more labels (*NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, MISOGYNY-NON-SEXUAL-VIOLENCE*).

**Table 1**
Instances of the dataset in each split and language.

| Dataset Split | #Spanish Instances | #English Instances | #Total |
|---|---|---|---|
| Train | 3660 | 3260 | 6920 |
| Development | 549 | 489 | 1038 |
| Test | 1098 | 978 | 2076 |

### 3.2. Data Preprocessing and Data Augmentation

Due to the fact that the texts of the dataset are from Twitter, the language of these comments is an informal language, which contains hashtags, mentions, emojis, and URLs that enter noise into the models. In order to reduce noise and the variability of the data, we perform data preprocessing. This preprocess in the text will help to identify patterns in texts in an easier way. To preprocess the texts, we apply the following steps:

- Remove # in hashtags.
- Replace user's mentions by the string "user".
- Replace the URLs by the string "url".

In addition, the systems we developed that use an architecture incorporating annotator's information (Section 2.2) require data that includes the annotator's features. To fulfill this requirement, we replicated each text six times, assigning different annotator features (age and gender) to each replicated instance, along with the label assigned by the annotator.

### 3.3. Experiments and Selected Models

To achieve the goal of EXIST shared task, we propose three experiments for each task. Each experiment has a different configuration and different models are selected to run the experiments. The proposed experiments are the following:

- **Baseline**. This experiment employs the base architecture explained in Section 2.1. To conduct our experiments we selected four multilingual models due to the multilingual nature of the EXIST dataset (tweets are in English and Spanish). The selected models are mDeBERTa [17] in both base [18] and large [19] versions, and XLM-RoBERTa [20] in both base [21] and large [22] versions. This experiment consists of a fine-tuning of the selected models.
- **Baseline with Data Augmentation.** This experiment uses the base architecture explained in Section 2.1. To run this experiment, we perform Data Augmentation. This experiment consisted of a fine-tuning of the mDeBERTa base model with an augmented version of the data, where each text was repeated six times and associated with the label of each annotator.
- **Integrating Annotators' Information.** In this experiment, we utilize the integrating annotator's information (Section 2.2). This experiment consists of testing which combination method to integrate information related to the annotators performs the best. The selected model for this experiment was a mDeBERTa base model.

As can be seen, for baseline with data augmentation and integrating the annotator's information experiment, we use mDeBERTa base model. The reason to select this model is that we think that the mDeBERTa base model has fewer parameters and is more efficient in comparison with the other selected models. The choice of such pre-trained models is due to the multilingual nature of the tasks, as the task are proposed for two different languages, English and Spanish.

### 3.4. Training Approach and Hyper-Parameters Search

During the competition, we established two phases. In the first phase, the development one, we train our model with the train set and evaluate it with the development set. In the second phase, the evaluation one, we train our models with the train and development sets for the final models that will be used for predicting over the test set. More details are given in the next sections.

**Hyper-parameter optimization:** Hyper-parameter optimization is an important step during the training of a model, due to the fact that the models have a big amount of hyper-parameters that should be optimized to adjust the model to a specific task. All of our experiments implement hyper-parameter optimization. For the hyper-parameter optimization, we use Optuna [23], an automatic hyper-parameter optimization software framework. Specifically, we opted for a random search method.

**Table 2**

Search space used to optimize hyper-parameters for the selected models in tasks 1, 2, and 3 of the EXIST 2023 shared task.

| Hyper-parameter | Search Space |
|---|---|
| Learning-rate | (2e-5, 2e-7) |
| Weight-decay | (0.01, 4e-5) |
| Batch-size | [8, 16, 32] |

## 4. Results

In this section, we present the results obtained by the system developed as part of our participation in EXIST 2023. The experiments are conducted in two phases, the development phase, where we select the best models, and the evaluation phase where we evaluate the selected models.

### 4.1. Development Phase

In this phase, we train the models with the train and the development splits of the dataset. Then, we select the best model for each task. To evaluate our systems we use macro-F1, which is the harmonic average of precision and recall averaged over different classes. For each task of EXIST, we propose three strategies explained in Section 3.3. The results obtained in the development phase are shown in Tables 3, 4, 5, and 6.

Table 3 shows the results obtained by the developed baseline systems for all tasks. As can be seen, the large version of all models improves the base version due to the fact that they are more complex models and can capture more information. If we observe the results for each task, we can see that for Task 1, mDeBERTa Large achieves the best result in macro-F1 score with 0.8618 followed by XLMRoBERTa Large with a macro-F1 of 0.8606. The mDeBERTa and XMLRoBERTa base models showed good results with 0.8519 and 0.8558 macro-F1 scores respectively. In task 2, mDeBERTa Large outperforms the rest of the models (0.6113). It was followed by the mDeBERTa Large model with a macro-F1 of 0.6092. The mDeBERTa Base and XLMRoBERTa Base models obtained slightly lower results with macro-F1 of 0.5736 and 0.5577 respectively. Finally, for task 3, XLMRoBERTa Large obtained the best result in macro-F1 score in task 3 with a score of 0.37, the rest of the models had lower results (0.37 to 0.3483).

**Table 3**
Results of the different models in the Baseline experiment for tasks 1 (Sexism Identification), 2 (Source Intention), and 3 (Sexism Categorization) on EXIST 2023 development set. The selected model for one of the runs in the evaluation phase is shown in bold.

| Model | Task 1 | Task 2 | Task 3 |
|---|---|---|---|
| mDeBERTa Base | 0.8519 | 0.5736 | 0.3254 |
| **mDeBERTa Large** | **0.8618** | **0.6113** | 0.3483 |
| XLMRoBERTa Base | 0.8585 | 0.5577 | 0.3515 |
| **XLMRoBERTa Large** | 0.8606 | 0.6092 | **0.3700** |

Once base models were set (in terms of pre-trained model selected and best performing hyperparametrization), we proceeded with the rest of our experiments. In Table 4, we can see the results obtained in the systems that integrate the information from the annotators, either by adding only the labels assigned by the annotators (Experiment baseline with data augmentation) or with the socio-demographic information provided and the labels (Experiment integrating annotators' information). For this task, we can observe that the evaluated strategies achieve relatively close macro-F1 scores with a range from 0.7534 to 0.7567. The weighted strategy achieved the best result in macro-F1 (0.7567).

The results of the different proposed experiments for task 2 are presented in Table 5. In this Table, we can see the result of the experiments that include annotator information (Baseline with Data Augmentation and Integrating Annotators' Information). Due to the unbalanced classes in the dataset, we propose two experiments: one without class weight in the computation of the loss during neural network training, so we train the model with the unbalanced data, and the other with class weighting to deal with the imbalance. We establish the weight of each class as the result of dividing the number of samples for that class by the total number of samples. In this task, we can see that the strategy of MLP on categorical features then concat outperforms the other strategies with a 0.4851 when class weighting is used. Moreover, we can observe that the use of class weighting outperforms the models that do not use class weighting in all of the experiments except in the Weighted strategy where the use of class weighting is the worst strategy (0.4803 to 0.4818). Regarding the base strategy (experiment baseline with data

**Table 4**
Results of the different proposed strategies that incorporate annotators' information (experiment base with data augmentation and experiment integrating annotator information) for Task 1: Sexism Identification on EXIST 2023 development set. The selected model for one of the runs in the evaluation phase is shown in bold.

| Strategy | macro-F1 |
|---|---|
| **Base DA** | **0.7537** |
| Attention | 0.7560 |
| Concat | 0.7534 |
| Gating | 0.7548 |
| CategoricalMLP | 0.7562 |
| **Weighted** | **0,7567** |

augmentation), we can observe that it obtains a good result, close to CategoricalMLP strategy when class weights are used with a 0.4826 macro-F1 score.

**Table 5**
Results of the different proposed strategies that incorporate annotators' information (experiment base with data augmentation and integrating annotator information) for Task 2 (Source Intention) on EXIST 2023 development set. The selected model for the evaluation phase is shown in bold.

| Strategy | macro-F1 | |
|---|---|---|
| | Without Class Weight | Weight Sklearn |
| **Base DA** | 0.4721 | **0.4826** |
| Attention | 0.4657 | 0.4744 |
| Concat | 0.4661 | 0.4805 |
| Gating | 0.4666 | 0.4774 |
| **CategoricalMLP** | 0.4642 | **0.4851** |
| Weighted | 0.4818 | 0.4803 |

Due to the fact that task 3 is a multilabel classification task, we propose two experiments, one without weight class and the other with class weight to try the unbalanced data. The strategy used to assign class weight to each class is different than task 2. In this task, we assigned a weight of positive examples, where we divided the total of examples of the dataset by the frequency of that class. As in the previous tasks, in Table 6 we can see the results for the experiment with Data Augmentation and the one that integrates information from the annotators. The results show that the best strategy is Gating with class weight where the strategy achieves 0.4109 in macro-F1 scores. Moreover, we can observe that the use of class weight outperforms the models that do not use class weight in all of the experiments. Finally, if we observe the result of the base strategy with class weight, we can see that outperforms other strategies such as Attention, CategoricalMLP, and Weighted. Nevertheless, the differences between strategies are close in the range of 0.4087 to 0.4138.

As a resume, at the end of this phase, we selected three models to submit the result to the evaluation phase for all of the proposed tasks. We decided to send one model of each proposed

**Table 6**
Results of the different proposed strategies that incorporate annotators' information (experiment base with data augmentation and integrating annotator information) for Task 3 (Sexism Categorization) on EXIST 2023 development set. The selected model for the evaluation phase is shown in bold.

| Strategy | macro-F1 | |
|---|---|---|
| | Without Class Weight | With Class Weight |
| **Base DA** | 0.3109 | **0.411** |
| Attention | 0.3149 | 0.4087 |
| Concat | 0.3164 | 0.4130 |
| **Gating** | 0.3107 | **0.4138** |
| CategoricalMLP | 0.3124 | 0.4109 |
| Weighted | 0.3179 | 0.4089 |

experiment. For Task 1, we selected mDeBERTa Large with the baseline strategy, and two mDeBERTa Base models, once with the baseline with data augmentation and other integrating annotators' information with the strategy weighted feature sum on transformer output. In Task 2, we decided to submit for baseline model a mDeBERTa Large model, a mDeBERTa Base model for baseline with data augmentation experiment, and a mDeBERTa Base with the strategy MLP on categorical features then concat such as strategy that includes socio-demographic annotator's information. Finally, for Task 3, we selected a XLMRoBERTa Large model such as baseline model, and two versions of mDeBERTa Base model, one with the Baseline Data Augmentation strategy and the other incorporating annotator information with the strategy gating on categorical features and then sum.

## 4.2. Evaluation Phase

In this section, we present the official results obtained by our submissions. For each task, we present submitted three runs based on the systems that reported the best performance during the pre-evaluation phase. The three runs are related to the best model in the baseline models, the best model in the baseline with data augmentation, and the best model integrating the features of the annotators.

To evaluate the system in the evaluation phase, organizers use the official Information Contrast Measure (ICM) [24] which is a similarity function that tries to calculate the similarity of the output label to the ground truth. The goal of the ICM is to penalize fewer an error in similar classes and more an error between totally different classes. For example, if we have an error between sexist and non-sexist it will be more penalized than if we had it between two specific types of sexism. This happens because sexist and no sexist are two different classes and two specific types of sexism share that they are sexist.

Due to the fact that the organizers propose a learning with disagreement paradigm to consider the different points of view of the annotators, each task is evaluated in three modes: hard-hard, soft-soft, and hard-soft. Each mode and its associated metrics are described below and defined by organizers [8, 9]

- **Hard-Hard evaluation.** In this type of evaluation considers the output of the system such as the traditional setting (one or more categories for each instance of the dataset) and the ground of truth such as a gold label (majority vote of the label assigned by annotators).

  - **ICM-Hard.** The official metric for the ranking.
  - **ICM-Hard Norm.** The ICM-Hard metric normalized.
  - **F1.** F1-score. In Task 1 the F1-score corresponds to the sexist class *"YES"*. In the other tasks, this metric represents the average F1 score for all classes.

- **Soft-Soft evaluation.** This evaluation method considers the output of the system (probability for each class, for each instance) and the ground of truth such as a full set of human annotations with their variability (the proportion of human annotators that have selected each category).
  - **ICM-Soft.** The official metric for the ranking.
  - **ICM-Soft Norm.** The ICM-Soft metric normalized.
  - **Cross Entropy.** The result of the cross entropy measure used only for Task 1 and Task 2.
- **Hard-Soft evaluation.** This mode of evaluation considers the output of the systems such as traditional settings (one or more categories for each instance of the dataset) and the ground of truth such as a full set of human annotations with their variability (the proportion of human annotators that have selected each category).
  - **ICM-Soft.** The official metric for the ranking.
  - **ICM-Soft Norm.** The ICM-Soft metric normalized.

#### 4.2.1. Task 1

In task 1 (Sexist Identification), we send the following run, from best to worst results:

- **Run 1: Baseline.** The hyper-parameter setting for this model is 1.732959582004883e-5 for learning rate, 0.0004027852347237054 for weight decay, and 16 for batch size.
- **Run 2: Model integrating annotators' information with the strategy weighted feature sum on transformer output.** In this run, we establish a learning rate of 1.5415726856440183e-05, a weight decay of 0.005162783476382758, and a batch size of 8.
- **Run 3: Data Augmentation Baseline.** The hyper-parameter used in this run are 7.28772895885929e-06 of learning rate, 0.007902681589455288 of weight decay, and 16 of batch size.

Table 7 shows the results obtained by our team in task 1 for all of the evaluation methods. As can be observed, in Hard-Hard evaluation method SINAI_1 run outperforms the other two runs submitted to the task with a 0.5584 of ICM-Hard. The SINAI_2 and SINAI_3 achieve close values with 0.5543 and 0.544 respectively. We believe that the best run for our team in this evaluation method is because it follows the Baseline experiment. In this experiment, we train the system with the same strategy that the organizers are going to evaluate the system, with only one label assigned to each instance.

Regarding the Soft-Soft type of evaluation, we can see that SINAI_3 outperforms with a significant difference from the other runs with a 0.903 ICM-Soft. Moreover, the worst result of these runs is for SINAI_2 run with a -5.6559 of ICM-Soft. We think that SINAI_3 obtained the best results due to the fact that, in Baseline Data Augmentation experiment, each comment of the dataset has assigned all the labels selected by annotators, and with this strategy, the system can adjust in a better way the probability of each class for the texts.

Finally, in the Hard-Soft method, we can observe that the best run is SINAI_2 with a 0.2678. Nevertheless, the result of the other run is too close to SINAI_2 with 0.264 of SINAI_1 and 0.2513 of SINAI_3. In our opinion, in this type of evaluation, our systems obtain similar results even though they are trained in different ways due to the fact that the hard label of our systems is similar to each instance of the dataset.

**Table 7**
Results of SINAI Team submission on Task 1: Sexism Identification. The best result for each evaluation method is in bold.

| Run | Hard-Hard | | | | Soft-Soft | | | | Hard-Soft | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ranking | ICM-Hard | ICM-Hard Norm | F1 | Ranking | ICM-Soft | ICM-Soft Norm | Cross Entropy | Ranking | ICM-Soft | ICM-Soft Norm |
| **SINAI_3** | 11 | 0.5440 | 0.7127 | 0.7715 | **1** | **0.9030** | **0.6421** | **0.7960** | 11 | 0.2513 | 0.5368 |
| **SINAI_1** | **8** | **0.5584** | **0.7219** | **0.7804** | 24 | 0.4863 | 0.5748 | 1.5759 | 9 | 0.2640 | 0.5389 |
| **SINAI_2** | 9 | 0.5543 | 0.7192 | 0.7719 | 56 | -5.6559 | -0.4175 | 7.3080 | **8** | **0.2678** | **0.5395** |

### 4.2.2. Task 2

In task 2 (source intention), we send the following runs, from best to worst results:

- **Run 1: Baseline.** The hyper-parameter configuration for this model is 1.598807925394166e-5 for learning rate, 0.0005069050078567268 for weight decay, and 16 for batch size.
- **Run 2: Model integrating annotators' information with the strategy MLP on categorical features then concat.** The hyper-parameter result of hyper-parameter optimization for this model are 1.362373572892371e-05 for learning rate, 0.00634231591793882 for weight decay, and 16 for batch size.
- **Run 3: Data Augmentation Baseline.** With a learning rate of 1.861739374638094e-05, weight decay of 0.004698351934634685, and batch size of 32.

The results obtained by our team in Task 2 for all of the evaluation methods are presented in Table 8. As can be seen, in Hard-Hard evaluation method SINAI_2 run outperforms the other two runs submitted to the task with a -0.0496 of ICM-Hard. The SINAI_1 and SINAI_3 achieve ICM-Hard values of -0.5959 and -0.9649 respectively. We believe that the best run for our team in this evaluation method is because it integrates annotators' information that helps the model to recognize the different sexist intentions.

Regarding the Soft-Soft type of evaluation, we can see that SINAI_3 outperforms the other runs with a -2.29 ICM-Soft. In addition, the worst result of these runs is formed SINAI_2 with a -10.9851 of ICM-Soft. We think that SINAI_3 obtained the best results due to the fact that it follows the Baseline Data Augmentation experiment, and like Task 1, for Soft to Soft method, this experiment seems to be the best strategy.

Finally, in the Hard-Soft method, we can observe that the best run is SINAI_2 with a -10.9830. Nevertheless, the results of the other runs are relatively close to this run with -11.0357 of ICM for SINAI_1 and -15.0466 for SINAI_3. In our opinion, in this type of evaluation, and for systems that include more possible labels to classify, our system is better able to match the class most voted by annotators for each text instance.

**Table 8**
Results of SINAI Team submission on Task 2: Sexism Intention. The best result for each evaluation method is in bold.

| Run | Hard-Hard | | | | Soft-Soft | | | | Hard-Soft | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Ranking | ICM-Hard | ICM-Hard Norm | F1 | Ranking | ICM-Soft | ICM-Soft Norm | Cross Entropy | Ranking | ICM-Soft | ICM-Soft Norm |
| **SINAI_3** | 30 | -0.9646 | 0.4667 | 0.2544 | **8** | **-2.2900** | **0.7831** | **1.6753** | 34 | -15.0466 | 0.4573 |
| SINAI_1 | 25 | -0.5959 | 0.5453 | 0.2562 | 19 | -4.2437 | 0.7332 | 2.3710 | 32 | -11.0357 | 0.5597 |
| **SINAI_2** | **18** | **-0.0496** | **0.6617** | 0.4924 | 25 | -10.9851 | 0.5610 | 4.6237 | **31** | **-10.9983** | **0.5607** |

### 4.2.3. Task 3

In task 3 (sexist categorization), we sent the following runs, from best to worst results in the development phase:

- **Run 1: Model integrating annotator information with the strategy gating on categorical feats then sum.** For this model we use the following hyper-parameter, a learning rate of 1.59165455576808e-06, a weight decay of 0.006133244155950455, and a batch size of 16.
- **Run 2: Data Augmentation Baseline.** With a learning rate of 1.209486958019233e-05, a weight decay of 0.00824196801678852, and a batch size of 8 for hyper-parameter configuration.
- **Run 3: Baseline.** We use 1.725910088321729e-5 for the learning rate, 0.006144588290326373 for weight decay, and 8 for batch size.

In Table 9 we can see the results obtained by our system for Task 3 (Sexism Categorization) in all of the evaluation methods. For Hard-Hard method, the run that obtains the highest result is SINAI_2 with a 0.1472 of ICM-Hard, followed by SINAI_3 with a 0.0249 score in ICM-Hard and SINAI_1 with -0.3020. It can be noted that this system implements a system proposed in Data Augmentation Baseline, where the system has the same text repeated with the different labels assigned by each annotator. This type of experiment seems to be a good way to make a multilabel classification for systems evaluated with Hard-Hard method.

For the runs evaluated like Soft-Soft can be noticed that SINAI_3 outperforms the other runs with a significant difference in ICM-Soft. SINAI_3 achieve -7.1306 ICM-Soft value. In this case, SINAI_3 has a base architecture and it is trained with hard labels.

In Hard-Soft evaluation, the best run is SINAI_3 which achieves -12.1399 in ICM-Soft metric. This run outperforms significantly the other two runs SINAI_2 and SINAI_1 with a -27.2984 and -34.9858 score in ICM-Soft respectively.

As we can observe in the last two types of evaluation, the best run is for systems implemented in the baseline strategy. This may be because with so many classes to choose from, adding annotator information may add more noise to the models that try to predict the sexism category.

**Table 9**
Results of SINAI Team submission on Task 3: Sexism Categorization. The best result for each evaluation method is in bold.

| Run | Hard-Hard | | | | Soft-Soft | | | Hard-Soft | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Ranking | ICM-Hard | ICM-Hard Norm | F1 | Ranking | ICM-Soft | ICM-Soft Norm | Ranking | ICM-Soft | ICM-Soft Norm |
| **SINAI_3** | 10 | 0.0249 | 0.5971 | 0.5033 | **10** | **-7.1306** | **0.7013** | **12** | **-12.1399** | **0.6112** |
| **SINAI_2** | **7** | **0.1472** | **0.6203** | **0.5822** | 19 | -13.5493 | 0.5858 | 26 | -27.2984 | 0.3384 |
| SINAI_1 | 14 | -0.3020 | 0.5352 | 0.5306 | 24 | -34.9362 | 0.2010 | 28 | -34.9858 | 0.2001 |

## 5. Conclusion

This paper presents the participation of SINAI research group in the sEXism Identification in Social neTwork shared task at CLEF 2023. In all of the tasks, we explore some different ways to train models, a baseline strategy, a baseline strategy with data augmentation, and various types of methods used to integrate socio-demographic annotators features to implement learning with disagreement paradigm. For all the tasks, we can see how the results depend on the evaluation method. Focusing on Task 1, we can see how the strategy with data augmentation improves sexism identification. However, the other two strategies have good results. In Task 2, it can be noticed that integrating socio-demographic annotator information achieve the best results in the majority evaluation method, this can be possible because there are more label and this information can help to determine each category of sexism. Finally, in Task 3, the only task based on multilabel classification the baseline system obtains the best results. Maybe due to the fact that integrating more information can insert more noise and the model has more difficulty in the identification of the different categories of sexism.

We conclude that, in general, the incorporation of socio-demographic information from annotators does not help the models to conduct the sexist tasks as we thought at the beginning of this shared task. This could be due to a need for larger training datasets for the network to better mimic human behavior according to each annotator profile. Nevertheless, it seems that when the number of classes increases, as we have observed in Task 2, this information can be more relevant. In future work, we plan to make a more rigorous analysis of the error of our systems and the impact of each socio-demographic features in our systems. In addition, we want to search for more relevant characteristics that are related to sexism and can help to detect it.

## Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

## References

[1] Definition of sexism by the oxford english dictionary, 2023. URL: https://www.oed.com/view/Entry/177027rskey=577LIN&result=2&isAdvanced=false#eid.

[2] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: EVALITA@CLiC-it, 2018.

[3] E. Fersini, P. Rosso, M. E. Anzovino, Overview of the Task on Automatic Misogyny Identification at IberEval 2018, in: IberEval@SEPLN, 2018.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[5] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, 2023. arXiv:2303.04222.

[6] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6389.

[7] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[8] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality,

Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Thessaloniki, Greece, 2023.

[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: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[10] S. Frenda, G. Bilal, et al., Exploration of Misogyny in Spanish and English tweets, in: Third workshop on evaluation of human language technologies for iberian languages (ibereval 2018), volume 2150, Ceur Workshop Proceedings, 2018, pp. 260–267.

[11] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, M. T. Martín-Valdivia, A multi-task learning approach to hate speech detection leveraging sentiment analysis, IEEE Access 9 (2021) 112478–112489. doi:10.1109/ACCESS.2021.3103697.

[12] F. M. Plaza-del-Arco, M. D. Molina-González, L. A. Ureña-López, M.-T. Martín-Valdivia, Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection, Knowledge-Based Systems 258 (2022) 109965.

[13] S. Halat, F. M. Plaza-Del-Arco, S. Padó, R. Klinger, Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language (2022).

[14] J. M. Pérez, F. M. Luque, D. Zayat, M. Kondratzky, A. Moro, P. S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, V. Cotik, Assessing the impact of contextual information in hate speech detection, IEEE Access 11 (2023) 30575–30590. doi:10.1109/ACCESS.2023.3258973.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[16] K. Gu, A. Budhkar, A package for learning on tabular and text data with transformers, in: Proceedings of the Third Workshop on Multimodal Artificial Intelligence, Association for Computational Linguistics, Mexico City, Mexico, 2021, pp. 69–73. URL: https://aclanthology.org/2021.maiworkshop-1.10. doi:10.18653/v1/2021.maiworkshop-1.10.

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

[18] "hugging face mdeberta base model", 2021. URL: https://huggingface.co/microsoft/mdeberta-v3-base, last Accesed June 1, 2023.

[19] Hugging face mdeberta large model, 2021. URL: https://huggingface.co/microsoft/deberta-v3-large, last Accesed June 1, 2023.

[20] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[21] Hugging face xlm-roberta base model, 2019. URL: https://huggingface.co/xlm-roberta-base.

[22] Hugging face xlm-roberta large model, 2019. URL: https://huggingface.co/xlm-roberta-large.

[23] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperpa-

rameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.

[24] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819.