Sexism Identification in Tweets Using Ensembles & Augmentation: A Multilingual Approach

Notebook for the EXIST Lab at CLEF 2025

Syeda Rija Hasan Abidi[*,†], Muhammad Shoaib Khursheed[*,†], Sarah Faisal Sikandar[†], Sabahat Zahra[†], Faisal Alvi and Abdul Samad

Computer Science Program, Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan

Abstract

In this rapidly advancing digital world, misogynistic dialogue has evolved beyond traditional boundaries, now seeping into online spaces such as Twitter or X. This study aims to detect and classify instances of sexism in tweets. We explore three tasks posed by CLEF lab's EXIST 2025 challenge, focusing on English and Spanish tweets. These tasks comprise: Binary classification to detect sexist tweets (1.1), multi-class classification to classify the author's intention (1.2), and multi-label classification to identify various dimensions of sexism (1.3). Our methodology leverages several large language models (LLMs), prominently multilingual BERT and XLM_Roberta, combined with an ensemble learning approach. We employ data augmentation techniques such as cross-translation, EASE, and AEDA, and develop separate models for English and Spanish to optimize language-specific predictions. Model evaluation is conducted using hard labels, derived through majority annotator voting, and soft labels, derived from class probability distributions. We achieved the 4th rank for Spanish predictions and 13th in English and Spanish combined soft evaluation for task 1.1. In the soft evaluation for tasks 1.2 and 1.3, our team ranked 6th and 5th, respectively.

Keywords

Deep Learning, LLM, Augmentation, AEDA, Ensemble, BERT, RoBERTa, AI, Sexism

1. Introduction

The rise of digital communication has amplified gender-based discrimination and hate speech, most of which goes undetected. Models that can detect sexism in tweets are imperative to address this ever-growing crisis, enabling platforms such as X (formerly Twitter) to take appropriate actions against harmful content and mitigate sexism in online interactions.

In this research, we participate in the CLEF EXIST Lab's Task 1: Sexism Detection in Tweets, which aims to systematically identify and analyze sexist language [1]. The task is divided into three subtasks with primary objectives: Identification, Source Intention, and Categorization.

1. **Subtask 1: Sexism Identification** Subtask 1 involves binary classification (**YES** or **NO**) to decide whether or not a given tweet contains or describes sexist expressions or behaviors (i.e., exhibits sexism, describes a sexist situation, or criticizes a sexist situation).
2. **Subtask 2: Source Intention** If subtask 1 classifies a tweet as sexist, subtask 2 identifies the author's intention. This is a ternary classification of the classes:
    - **Direct:** The intent is to be sexist or endorse sexism.
    - **Reported:** The intent is to report a sexist situation, experienced by a woman, in the first or third person.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*]Corresponding author.

[†]These authors contributed equally.

syedarija02@gmail.com (S. R. H. Abidi); mk07149@st.habib.edu.pk (M. S. Khursheed); sf08449@st.habib.edu.pk (S. F. Sikandar); sz08447@st.habib.edu.pk (S. Zahra); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@st.habib.edu.pk (A. Samad)

0009-0007-6208-8537 (S. R. H. Abidi); 0009-0005-4441-0030 (M. S. Khursheed); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

- **Judgmental:** The intent is to judge a sexist situation to condemn sexism.

3. **Subtask 3: Sexism Categorization** If subtask 1 classifies a tweet as sexist, subtask 3 categorizes the tweet based on the type(s) of sexism it contains. The categories include:
   - **Ideological and inequality**
   - **Stereotyping and dominance**
   - **Objectification**
   - **Sexual violence**
   - **Misogyny and non-sexual violence**

Subtasks 1.1, 1.2, and 1.3 are referred to as Tasks 1, 2, and 3, respectively, throughout this paper.

2. Literature Review

Given the hierarchical nature of this challenge and sexism in textual information, approaches to all three tasks overlap. In previous years, many approaches to identifying sexism in tweets involved encoding-based transformer models, augmentation, LLMs/Prompt Engineering, and ensemble learning. Before fine-tuning, the preprocessing and selection of relevant data are also crucial [2].

Top-performing models in EXIST 2024 were transformer-based architectures fine-tuned to EXIST datasets. These architectures included BERT, RoBERTa, DeBERTa, and their multilingual variants. Teams like BAZI [3] and Victor-UNED [4] focused on training a single encoding-based model using soft labels, achieving top-10 rankings in soft evaluation in 2024 [2]. The models that performed exceptionally well were multilingual-BERT (mBERT) and XLM_RoBERTa, trained on text from multiple languages, enabling them to capture cross-lingual semantic similarities and transfer knowledge across languages. By learning shared representations across languages, the models excel in handling diverse languages, including those with limited labeled data, making it effective for various NLP applications, such as machine translation and information retrieval [5]. Naebzadeh et al. from the NICA team utilized sdadas/xlm-roberta-large-twitter, google-bert/bert-base-multilingual-uncased, and FacebookAI/xlm-roberta-base, where mBERT outperformed all models in all tasks. They attribute the performance to the simplicity and reduced number of parameters, as they are less prone to overfitting and generalize better [6]. For Team frms [7], XLM_Roberta performed best for Task 1, and mBERT performed best for Task 2 [2]. However, for other teams like BAZI [3], XLM_Roberta was the best for Task 2. This shows that multiple models must be accommodated for the best overall results.

Ensembling is a technique that combines the capabilities of multiple models. It has shown promise for all three EXIST tasks from the work done by UMUTeam [8], where they created ensembles of two Spanish (BETO and MarIA) and two multilingual models (deBERTa v3 and XLMTwitter). Two methods of ensembling were implemented: Knowledge Integration (KI), which merges all information into one model, and Ensemble Learning (EL), which trains separate models and combines their predictions. Through KI, they reached the 8th position in Task 2 [2]. For English tweets specifically, EL performed better. Their best results were in Spanish, which suggests that BETO and MarIA may be good options to integrate for Spanish. Also employing EL, Team Awakened [9] merged mono-lingual, multi-lingual, and domain-specific models like twitter-xlm-roberta-base-sentiment and roberta-hate-speech-dynabench-r4 to create an ensemble of models and used weighted voting to assign higher weights to high-performing models. Additionally, equal contribution voting with EL was used by teams like CIMAT-CS-NLP [10] and Medusa [2] [11].

Other than EL, teams tweaked the weights of the attention layer to create better sentence representation. Team Penta-nlp's approach accounted for the attention weights of the sentence representation, which helped their models yield the best results for each task. Their results again underscore that BERT-based models are best trained to capture the pattern of sexism when considering the attention layer [12]. Prompt engineering through LLMs also spanned the experimental space in EXIST 2024. Zero-shot and Few-shot led to the best predictions in hard-hard evaluation compared to encoding-based transformers. Teams like CIMAT-CS-NLP [10] used summed zero-shot responses from Gemini, while

ABCD [13] created one Llama-2 response per annotator, simulating the annotation process and improving classification robustness. For hard evaluation in Task 3, the top-performing team was the ABCD Team. They used LLMs like Llama 2 and T5 and models like XLM_RoBERTa. They divided the datasets into six subsets corresponding to each annotator group, applied prompt engineering, and fine-tuned transformer models on each subset. With LLMs and prompt engineering, Team ABCD achieved the highest ICM-Hard Norm scores of 0.6320 for task 2 and 0.5862 for task 3 [13]. Plaza et al. note that employing LLMs with encoding-based transformers seems to be most efficient in hard evaluations of Tasks 2 and 3 [2].

Data augmentation is another technique prevalent in best-performing approaches. The NYCU-NLP team achieved top ranks in all soft-soft and hard-hard evaluations for task 1 [14]. They employed augmentation through back-translation using the Google Translate API and AEDA (An Easier Data Augmentation). While these augmentations are highly efficient in enriching the dataset, Rahman et al. show that the EASE (Extract Units, Acquire Labels, Sift, and Employ) method for augmentation performs better in low-resource experiments [15]. In any case, data augmentation has proven to increase model performance. Like Team ABCD, Team NYCU's approach also incorporates annotators' metadata, such as age, gender, and ethnicity, in the tweet embedding, resulting in a unified vector representation for each tweet. This approach tunes the models to nuanced and specific biases and stereotypes attached to certain ages, cultures, and genders [14]. One improvement in this approach would be identifying which metadata is significant to the prediction and which is irrelevant. This would also help one highlight the underlying biases attached to specific demographics in society at large.

This review suggests that incorporating mBERT and xlm_RoBERTa with other transformer-based models to create ensembles, language-specific fine-tuning, and data augmentation has shown much promise for sexism identification and classification in previous years.

3. Dataset Overview

The dataset contains English and Spanish tweets, each annotated by six annotators. Three datasets are expected from the lab: "Training" for training the models, "Development" for validation and pre-testing, and "Testing" for final testing and creating prediction files to be submitted to the lab. The stats of these datasets are shown in Table 1.

Table 1
EXIST Datasets: Statistics

Dataset	Total Count	English (EN)	Spanish (ES)
Training	6920	3260	3660
Development	1038	489	549
Testing	2076	978	1098

Tasks & Annotations - Learning with Disagreement Whether a tweet is sexist or not is highly subjective. Even with clear descriptions of the term and its meaning, many points of view arise, all of which may be valid considering the socio-cultural norms and other factors that vary from individual to individual. To assign such diversely identified labels to tweets, EXIST adopts the *Learning with Disagreements* paradigm [1]. This paradigm involves 6 annotators for each tweet, with 6 distinct standpoints. The differing standpoints are encapsulated in the metadata dictionary given with each tweet in all datasets, where annotators' demographics such as gender, age, ethnicity, nationality, and study level are stated. The overall distribution of demographics is largely normalized, with each tweet containing 1 annotator from each age group and an equal number of males and females. The rest of the demographic groups vary from tweet to tweet, but overall, the dataset contains equal representation from each group of individuals.

Tweet annotation for each annotator is illustrated hierarchically in Figure 1. A tweet is either labeled sexist or non-sexist (Task 1 - Binary - Green). If it is not labeled sexist, Tasks 2 and 3 are automatically

assigned the label “NO" for that annotator. If the tweet is labeled sexist, annotators assign one of three labels to identify source intention (Task 2 - Multi-class - Yellow) and any number of labels out of five to categorize the tweet (Task 3 - Multi-label - Blue). Hard labels are obtained by finding the majority vote of 6 annotators for each task. The hard label statistics are given in Table 2.

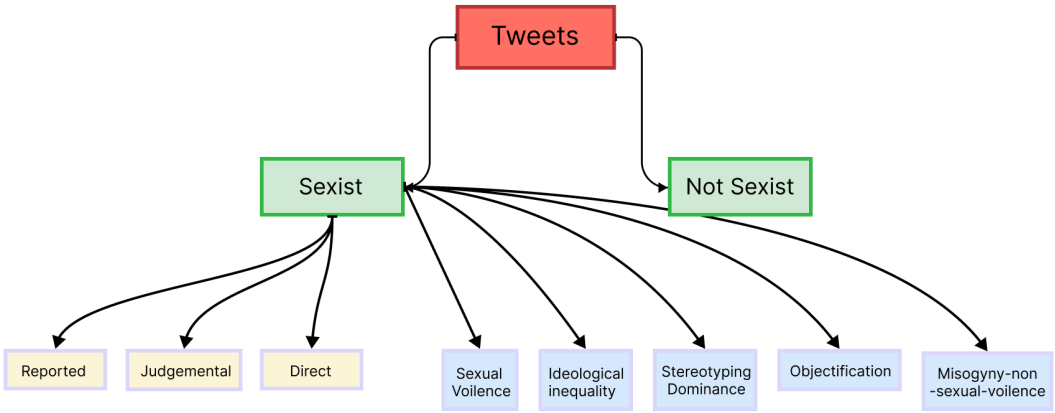

Figure 1: Annotation Hierarchy

Table 2
Hard Labels Statistics

Task	Training	Development
Task 1		
NO	3367	479
YES	2697	455
Task 2		
NO	3367	479
DIRECT	1294	204
REPORTED	459	75
JUDGEMENTAL	376	83
Task 3		
NO	3367	479
OBJECTIFICATION	1103	183
SEXUAL-VIOLENCE	675	123
STEREOTYPING-DOMINANCE	1423	241
IDEOLOGICAL-INEQUALITY	1113	212
MISOGYNY-NON-SEXUAL-VIOLENCE	856	158

4. Approaches

For the development phase, the “Development" dataset is used for testing instead of validation to compare and contrast the performance across various approaches. In this phase, validation is performed by splitting the “Training" dataset into an 80/20 ratio, where 80% is used for training and 20% is used for validation.

4.1. Models Overview

The following models were used in all three tasks:

- **BERT multilingual base model (uncased):** Trained in 102 languages, including English and Spanish, using a masked language modeling (MLM) objective [16].

- **BERT multilingual base model (cased):** Similar to uncased, however, this model treats capitalization as a different word, whilst uncased treats it the same.
- **XLM_RoBERTa:** Multilingual variant of RoBERTa model trained on 100 different languages [17].

In addition to these multilingual baselines, we employed a variety of monolingual and region-specific models tailored to English and Spanish for ensembling in Task 1.

- **distilroberta-base:** A distilled and computationally efficient variant of RoBERTa trained exclusively on English data.
- **bert-base-uncased:** BERT trained on English text with uncased tokens.
- **roberta-base:** An enhanced version of BERT trained on larger corpora with improved training strategies, including dynamic masking and longer sequences.
- **PlanTL-GOB-ES/roberta-base-bne:** A RoBERTa variant specifically pre-trained on Spanish text.
- **dccuchile/bert-base-spanish-wwm-cased:** A Spanish BERT model utilizing whole-word masking (WWM) during pretraining.
- **xlm-roberta-base:** Although multilingual, it was also included as a Spanish model due to its balanced multilingual training and demonstrated effectiveness on Spanish tasks.

4.2. Data Augmentation

To enhance our dataset, we applied data augmentation through translation, EASE, and AEDA. The training dataset was split into Spanish and English tweets for language-specific fine-tuning. To enrich and increase the size of the datasets, each was translated into the other language while keeping all other parameters constant. This involved translating all English tweets into Spanish and adding them to the Spanish dataset, and vice versa. The translation was done via the Helsinki-NLP translation model. We refer to this technique as cross-translation and the resulting datasets as cross-translated datasets.

In addition to cross-translation, we incorporated an enhanced version of the EASE (Extract Units, Acquire Labels, Sift, and Employ) approach [15]. First, meaningful units (sentences or facts) are extracted using the NLTK library. Then, pre-trained models generate labels for these units. Next, shorter-length samples are filtered, and the augmented data is integrated with the original training set. To further refine the data, we introduced an additional layer of synonym replacement before merging the augmented set back into the original dataset.

We also implemented AEDA (An Easy Data Augmentation), which involved randomly introducing punctuation marks at different positions in tweets. This augmentation strategy was employed by the NYCU team, who dominated the charts of all three tasks in 2024 [14][2]. Training on the AEDA-augmented cross-translated dataset significantly improved the scores of all three tasks, as demonstrated in Section 5. As shown in Table 3, through these augmentations, the size of the training dataset was increased by more than double.

Table 3
EXIST Datasets Post Augmentation: Statistics

Dataset	Total Count	English (EN)	Spanish (ES)
Cross-Translation	13840	6920	6920
EASE-S + Cross-Translation	25840	12920	12920
AEDA + Cross-Translation	24408	12204	12204

4.3. Base Architecture

Our architecture primarily consists of twelve independently fine-tuned models. For each of the three tasks, we use four models: two for English (EN) and two for Spanish (ES) (for hard and soft probabilities

each). Each model within a task is language-specific to aid tailored fine-tuning, accommodating linguistic nuances. This multilingual approach helps the models be better fine-tuned in their respective languages.

Our proposed framework comprises the following main components:

1. **Preparing Dataset:** The dataset is split into EN and ES and preprocessed for each task.
2. **Data Augmentation:** Cross-translation, AEDA, and EASE-S augmentation are applied. Spanish tweets are translated into English and vice versa, resulting in mirrored datasets. Thus, EN and ES datasets contain equal tweets (Table 1). AEDA and EASE are performed on cross-translated datasets with optimizations for each task.
3. **Fine-tuning the Models:** Each Model undergoes separate fine-tuning using either the EN or ES augmented datasets. The EN and ES models for a given task share the same architecture and training hyperparameters. The datasets contain labels from all six annotators for each task, which are replaced by the hard and soft labels extracted from gold files. Across tasks, different approaches may be used. For instance, Ensemble Learning is only used in Task 1, metadata is incorporated in Task 2, and EASE is used in Task 3. Fine-tuning may also be done separately for hard and soft labels to optimize evaluation-specific output, with separate models trained to predict soft and hard labels.
4. **Post-processing:** The prediction outputs from both EN and ES models are collected and aggregated. For soft labels, we calculate the probability distribution over possible classes. To reflect the real-world annotation process (six annotators), these probabilities are snapped to the nearest $\frac{1}{6}$ interval. This adjustment helps simulate annotator agreement more accurately. While the "snapping" works well for Task 3, additional steps are taken for Tasks 1 and 2 to ensure that the sum of all probabilities after finding the nearest $\frac{1}{6}$ equals 1. The output is then formatted to match the required submission specifications.
5. **Final Output:** We produce two types of outputs:
   - **Soft Labels:** Represented as probability distributions across classes. For Tasks 1 and 2 (multi-class), these probabilities must sum to 1. For Task 3 (multi-label), the sum may not equal 1.
   - **Hard Labels:** Derived from soft label probabilities by applying thresholds inspired by the gold label creation criteria:
     - **Task 1 (Multi-class):** Class selected if picked by more than 3 annotators.
     - **Task 2 (Multi-class):** Class selected if picked by more than 2 annotators.
     - **Task 3 (Multi-label):** Class selected if picked by more than 1 annotator.

     In cases where no class meets the threshold, 'NO' is selected as output.

4.4. Approach for Task 1: Sexism Identification in Tweets

This task asks for binary classification of tweets, determining whether a given tweet is sexist ("YES") or non-sexist ("NO"). We first filtered the dataset to include only the parts essential for Task 1, removing the rest of the metadata provided with each tweet. In hard evaluation, predictions are discrete. If at least 3 out of 6 annotators annotated a tweet "YES", it is processed as sexist. In contrast, soft evaluation considers the probability distribution over labels, capturing and returning the probabilities of whether the tweet is sexist or non-sexist, and evaluating the model's ability to predict probabilities close to human judgment.

4.4.1. Soft Models

In initial experimentation, Bert-base-multilingual-cased and DistilRoBERTa demonstrated strong baseline results. Cross-translation improved generalization across linguistic variations, while AEDA offered minimal gains on top of the cross-translated dataset. The most effective approach was language-aware ensemble models built upon the other approaches: AEDA and cross-translation.

For Ensemble Learning (EL), two ensembles were created for augmented EN and ES datasets:

- **EN**: `distilroberta-base`, `bert-base-multilingual-cased`
- **ES**: `bert-base-multilingual-cased`, `dccuchile/bert-base-spanish-wwm-cased`

Each model in the ensemble contributed weighted probabilities based on its performance on validation data. The optimal weights were determined using a trial-and-error approach with different weights.

At inference time, the system dynamically assigned weights based on the tweet's language: the English ensemble's output dominated predictions for English tweets, and a similar pattern was observed for the Spanish ensemble. This adaptive, language-specific ensembling strategy delivered our best results, achieving strong ICM, ICMNorm, and Cross Entropy scores shown in section 5.1.

4.4.2. Hard Models

Hard models for Task 1 also incorporated cross-translation, AEDA, and EL. Cross-translation preserved label semantics while improving model performance across languages, and AEDA improved generalization by making models more resilient to the informal and variable nature of tweets.

Similar to the soft models through EL, two ensembles were created for augmented EN and ES datasets:

- **EN**: `distilroberta-base`, `bert-base-uncased`, `roberta-base`
- **ES**: `PlanTL-GOB-ES/roberta-base-bne`, `dccuchile/bert-base-spanish-wwm-cased`, `xlm-roberta-base`

Ensemble weights were optimized via grid search and other common optimization techniques to maximize the overall F1 score. This unified approach using AEDA, cross-translation, and EL consistently outperformed any individual model or augmentation strategy.

4.5. Approach for Task 2: Source Intention in tweets

Task 2 involves a multi-class classification task where the source intentions of the tweets are identified. The approaches for this task included pre-processing, cross-translation, AEDA, and incorporating annotators' metadata.

The preprocessing steps for this task included:

- Removal of URLs and mentions.
- Removal of residual special characters except for basic punctuation.
- Reduced repetition of letters and punctuation to reduce noise, following Team Medusa [11].
- Normalizing whitespace.

XLM_RoBERTa and mBERT were fine-tuned on two versions of augmented datasets—cross-translated with AEDA and cross-translated without AEDA—for EN and ES separately, resulting in eight fine-tuned models. After training with each model and approach, EN and ES predictions were collected from separate models and merged, giving four ALL (EN+ES) prediction files with scores discussed in section 5.3.

AEDA For task 2, tweets with class "NO" are the overwhelming majority and result in class imbalance (Table 2). Hence, AEDA was only applied to tweets with hard labels "REPORTED," "JUDGMENTAL," or "DIRECT." The resulting datasets overcame the class imbalance and yielded improved results, especially with XLM_RoBERTa's Twitter variant as the base model.

Separate Training For Hard Models While the model fine-tuned with AEDA augmentation worked well for soft labels, it did not yield the same improvement for hard labels. Hence, various approaches were implemented to improve the scores for hard labels. These approaches included decreasing the class imbalance further by augmenting more tweets that belonged to the underrepresented class. Furthermore, for the previous models, the ground truth labels were taken as a vector of probabilities from the gold file for soft labels. This was replaced with the gold file for hard labels, and instead of a probabilities vector, the ground truth label was used for training.

Annotator's Metadata Another approach for this task was to utilize the annotators' metadata dictionary to capture socio-cultural biases that may arise while labeling tweets. This was done following the approach by the ABCD team, first ranked in Tasks 2 and 3 in 2024, where Quan et al. split the dataset into 6 subsets on the annotator's data and subsequently trained 6 component models on the split data [13]. We adopted this approach and trained 12 models, 6 for English and 6 for Spanish. The hard and soft labels were obtained by aggregating the predictions of each model, as described in section 4.3. Due to the long training time, larger models such as Roberta Large, used by Quan et al., were not fine-tuned, and a smaller augmented dataset without AEDA was used.

4.6. Approach for Task 3: Sexism Categorization in tweets

Task 3 was a multi-label classification task aimed at identifying multiple categories of sexism present in each tweet. We developed separate models for English and Spanish tweets and applied EASE-S and AEDA augmentation. These were done on top of cross-translation, enabling better alignment and generalization across both languages.

We fine-tuned BERT-multilingual-cased as the base model for both languages. While we also experimented with XLM_Roberta, the multilingual BERT variant consistently yielded better performance. Similarly, a combined augmentation pipeline (EASE-S + AEDA) was evaluated but did not outperform EASE-S or AEDA individually.

Each language model produced soft probabilities for each potential category, representing the likelihood that a tweet belonged to that category. These probabilities were then snapped to the nearest $\frac{1}{6}$ interval, reflecting the presence of six annotators. This snapping allowed the soft outputs to better emulate real-world annotation distributions and led to improved alignment with gold labels.

To convert the soft scores into hard labels, we applied a thresholding mechanism: any category with a probability greater than or equal to 0.167 (i.e., $\frac{1}{6}$) was included in the final label set for that tweet. If no category met this threshold, the tweet was labeled as "NO", indicating the absence of sexist content. This threshold was based on the assumption that a label assigned by at least one annotator should be considered valid, maintaining consistency with the annotation schema.

Finally, the predictions from the English and Spanish models were unified by merging their soft probability JSON files before thresholding, ensuring a comprehensive, multi-lingual label assignment.

5. Results and Discussions

The following sections present the results from our models across different tasks, organized into two phases: Dev Pre-testing & Experimentation and Test Results & Ranking. The pre-testing results were obtained using the Development (Dev) set, for which we had access to gold labels, enabling detailed experimentation, analysis, and comparison. This division allows us to highlight the range of techniques and experiments that informed model development—insights that would be lost if we only reported final test scores. Notably, the ICM-Soft Norm scores on the Dev set closely mirrored those on the Test set, reinforcing the reliability and generalizability of our approach. The Test Results & Ranking section shows the final results on the Test set and the rankings obtained through our runs.

5.1. Task 1: Soft Models

5.1.1. Dev Pre-testing & Experimentation

The results in Table 4 indicate a clear improvement when data augmentation techniques are applied compared to the baseline. The AEDA method outperforms both the Baseline and Cross-Translation across all three evaluation metrics: ICM-Soft, ICM-SoftNorm, and Cross-Entropy.

For ICM-Soft, the baseline score is negative (-2.388), suggesting poor alignment or inconsistency, while Cross-Translation significantly improves this to 0.526, and AEDA further enhances it to 0.712, indicating stronger soft-label consistency. Similarly, ICM-Soft Norm, which normalizes soft-label

Table 4
Task 1: Soft-soft Evaluation using bert-base-multilingual-cased

Metric	Baseline	Cross-Translation	AEDA
ICM-Soft	-2.388	0.526	**0.712**
ICM-Soft Norm	0.114	0.585	**0.615**
Cross Entropy	4.799	0.810	**0.943**

confidence, increases from 0.114 (Baseline) to 0.585 (Cross-Translation) and peaks at 0.615 with AEDA. Cross-entropy shows a notable decrease from 4.799 to 0.810 (Cross-Translation), and 0.943 with AEDA, indicating the predictions match the gold labels more with the cross-translated and AEDA-augmented model.

Table 5 shows the result of soft-soft evaluation scores for the three ensemble models Ensemble_EN, Ensemble_ES, and the Combined Ensemble. As can be seen among all the approaches, the final ensemble model, Combined Ensemble, achieves the highest ICM-Soft, ICM-SoftNorm, and Cross Entropy scores. This indicates that combining both English and Spanish ensembles attains the highest performance compared to the individual ensembles.

Table 5
Task 1: Soft-Soft Evaluation for Ensemble Models

Metric	Ensemble_EN	Ensemble_ES	Combined Ensemble
ICM-Soft	0.670	0.809	**0.870**
ICM-Soft Norm	0.608	0.630	**0.640**
Cross Entropy	0.966	1.023	**0.881**

5.1.2. Test Results & Ranking

We submitted one run for the test set, evaluated across both languages combined (ALL), Spanish (ES), and English (EN) subsets. As can be seen from Table 6, the results show variations across languages. On the test set, the Spanish subset showed the best performance in terms of ICM-Soft and ICM-Soft Norm, with values of 0.8426 and 0.6351, respectively, and achieved a rank of 4. The ALL subset followed with an ICM of 0.6767, ICM-Soft Norm of 0.6085, and rank 13. The English subset dropped behind in ICM-Soft and ICM-Soft Norm, scoring 0.5034 and 0.5808, and ranked 24. The Spanish subset's overall stronger performance highlights the effectiveness of the base models used in the Spanish ensemble and our cross-lingual augmentation strategies. As for the English ensemble, a bigger and better model might have given better results. While the ranking performance is consistent with trends seen in development, further refinements may be needed to improve English-specific generalization and better balance performance across all languages.

Table 6
Task 1: Soft-soft Evaluation on Test Set & Ranking

Run	Language	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
Run 3	ALL	13	0.6767	0.6085	0.8894
Run 2	ES	4	0.8426	0.6351	0.8609
Run 1	EN	24	0.5034	0.5808	1.0516

5.2. Task 1: Hard Models

5.2.1. Dev Pre-testing & Experimentation

Table 7 compares the performance of different techniques applied to Task 1 under hard evaluation metrics: ICM-Hard, ICM-Hard Norm, and FMeasure. We started with a baseline model (DistilRoBERTa

trained without augmentation, ensembling, or cross-translation), which gave relatively low scores of 0.223, 0.612, and 0.700 for ICM-Hard, ICM-Hard Norm, and F1 score, respectively. These are better than the baseline, but not very good, indicating that the base model alone was not sufficient to accurately classify sexist content in tweets. This could be due to less training data or its multilingual capacity. With cross-translation and ensemble modeling together, the results across all metrics improved significantly, yielding scores of 0.531, 0.766, and 0.843 for ICM-Hard, ICM-Hard Norm, and F1, respectively. This was a tremendous shift from the baseline, indicating the importance of enriching the training data and using more language-specific models. Finally, we incorporated AEDA into the already-designed ensembles and cross-translation setup. This gave the best overall results, with the highest scores in every metric. ICM-Hard rose to 0.575, ICM-Hard Norm became 0.788, and F1 increased to 0.858.

Table 7
Task 1: Hard-Hard Evaluation on Dev-set

Metric	Baseline	Base Model	Ensemble + Cross Translation	Ensemble + AEDA + Cross Translation
ICM-Hard	-0.481	0.223	0.531	**0.575**
ICM-Hard Norm	0.256	0.612	0.766	**0.788**
F1 Score	0.316	0.700	0.843	**0.858**

5.2.2. Test Results & Ranking

As done for soft-soft evaluation, we submitted one run for the test set, broken down into ALL (EN+ES), EN, and ES. Overall, the performance on the test set is slightly lower than that on the development set. On the dev set, our best configuration (ensemble + cross-translation + AEDA) achieved strong results with ICM = 0.575, ICMNorm = 0.788, and F1 = 0.858. However, on the test set, the best F1 score achieved was 0.7750 for the ES subset, followed by 0.7558 for ALL and 0.7302 for EN (Table 8). This drop indicates the inability of the models to generalize to unseen variations in the test set. The same fall is seen in scores for ICM and ICM Norm that came out to be 0.4953 and 0.7490, respectively, for ALL, 0.4962 and 0.74810 for ES, and 0.4800 and 0.7449 for EN. Notably, the Spanish subset performed best across all metrics, indicating that our multilingual and cross-translation strategies were particularly effective for non-English data. Meanwhile, the English subset lagged, which may point to dataset-specific nuances and base model choice. Even though the overall performance was quite consistent with dev scores, the overall rankings were not comparable to the competition. With ALL achieving a rank of 56, ES achieving a rank of 50, and EN achieving a rank of 90, it is clear that other approaches, such as zero-shot learning, should be explored, and specific heed must be paid to English models.

Table 8
Task 1: Hard-Hard Evaluation on Test Set & Ranking

Run	Language	Rank	ICM-Hard	ICM-Hard Norm	F1 Score
Run 3	ALL	56	0.4953	0.7490	0.7558
Run 2	ES	50	0.4962	0.7481	0.7750
Run 1	EN	90	0.4800	0.7449	0.7302

5.3. Task 2: Source Intention in tweets

5.3.1. Dev Pre-testing & Experimentation

For this subtask, we experimented with five different approaches and tested them in the development phase. The models for both soft-soft and hard-hard evaluations were trained together with these five types of approaches. First, we evaluated scores for the cross-translation approach on mBERT and xlm_RoBERTa models. xlm_RoBERTa consistently performed better than mBERT, yielding overall better scores as shown for both soft-soft and hard-hard evaluations in Tables 9 and 10. The addition of AEDA

augmentation improves the performance of both models, mBERT ICM-SoftNorm goes from 0.36 to 0.38, and xlm_Roberta from 0.40 to 0.41. A similar improvement is seen in the hard-hard evaluations, where xlm_Roberta improves from 0.44 to 0.46 ICM Norm. This highlights that managing class imbalance and introducing more samples of underrepresented classes through AEDA augmentation helps models identify the overall patterns and nuances better.

The last approach of Metadata incorporation performed better than the mBERT models; however, it did not outperform xlm_RoBERTa trained on cross-translated and AEDA-augmented data. The Metadata incorporation has promised improvement in scores for teams such as ABCD in 2024 [13]. In our case, the less favorable scores could be linked to the lack of AEDA augmentation while training component models for each language, and using the xlm_RoBERTa base model instead of xlm_RoBERTa Large. Training with augmented data and a bigger model posed an excessive computational overhead and hence was skipped.

Table 9
Task 2: Soft-soft Evaluation on Dev Set

Metric	Base Model	mBERT Cross Translation	xlm_RoBERTa Cross Translation	mBERT + AEDA	xlm_RoBERTa + AEDA	xlm_RoBERTa + Annotators' Metadata
ICM-Soft	-5.8599	-1.7699	-1.2154	-1.5778	**-1.2165**	-1.7906
ICM-SoftNorm	0.0537	0.3652	0.4074	0.3833	**0.4103**	0.4006
Cross Entropy	4.8106	1.8722	1.8057	1.8823	**1.8220**	1.8906

Table 10
Task 2: Hard-Hard Evaluation on Dev Set

Metric	Base Model	mBERT Cross Translation	xlm_RoBERTa Cross Translation	mBERT + AEDA	xlm_RoBERTa + AEDA	xlm_RoBERTa + Annotators' Metadata
ICM	-1.0234	-0.3568	-0.1606	-0.3555	**-0.1126**	-0.2092
ICMNorm	-1.0234	0.3884	0.4497	0.3888	**0.4647**	0.4345
FMeasure	0.1579	0.4469	0.5013	0.4677	**0.5330**	0.5102

5.3.2. Test Results & Ranking

Soft-soft evaluations We submitted three runs on the best identified model, xlm_RoBERTa trained on Cross-translation and AEDA-augmented datasets, split into ALL-EN+ES (Run 3), ES (Run 2), and EN (Run 1) tweets. The splits helped gauge the model's performance in English and Spanish together and separately.

Our runs performed exceptionally well on soft-soft evaluations as shown in Table 11. ALL achieved the rank 6th, ES was also ranked 6th, and EN achieved the rank of 7th out of 192 submissions [1]. The test scores are comparable to the dev scores (Table 9) and indicate that the model was well-equipped to generalize and predict soft probabilities for unseen tweets. It is to be noted that the model did not show any significant bias either for English or Spanish tweets while predicting soft probabilities.

Table 11
Task 2: Soft-soft Evaluation on Test Set & Ranking

Run	Language	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
3	ALL	6	-1.0572	0.4148	2.0396
2	ES	7	-0.9479	0.4241	2.0352
1	EN	6	-1.3180	0.3923	2.0446

Hard-hard evaluations The model was trained on the gold soft dataset instead of the gold hard dataset, which played a substantial role in the model's predictions. While the model predicted the relative difference of probabilities in soft predictions well, as indicated by the ranks and scores in soft evaluations, the trend understandably did not hold for hard evaluations.

The hard predictions were calculated by selecting the most probable class from soft predictions. Here, a curious disparity is found. Overall, the ranks are lower than the soft evaluations, which is expected because the training was performed with soft vectors. However, instead of a unanimous worse performance, the model performed especially worse for English tweets with Run 1 at rank 65, while Run 2 and 3 are ranked 21 and 26, respectively (Table 12). This hints at a bias in the xlm_RoBERTa base model for Spanish, but could also be linked to other participants' focus on English tweets specifically, which would explain the similar disparity in the ranks of Task 1.

Table 12
Task 2: Hard-Hard Evaluation on Test Set & Ranking

Run	Language	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
3	ALL	26	0.1987	0.5646	0.5101
2	ES	21	0.3022	0.5944	0.5439
1	EN	65	0.0488	0.5169	0.4566

5.4. Task 3: Soft Models

5.4.1. Dev Pre-testing & Experimentation

For the multi-label classification task, we evaluated the performance of several augmentation strategies on the soft models. The baseline results, derived from using unprocessed soft labels, were notably poor, with an ICM-Soft of −8.7 and a normalized ICM-Soft of 0.0. These metrics highlighted the need for improved label estimation strategies.

Our first attempt, which involved predicting hard labels using the base model and then applying softmax to reconstruct soft probabilities, yielded slightly better results (−7.0, 0.0), but ultimately proved conceptually flawed. This approach ignored the inherent distributional nature of soft labels and introduced error by artificially re-softening hard decisions. A substantial improvement was observed with cross-translation, where the dataset was expanded by translating tweets between English and Spanish. Training language-specific models on these extended corpora and then merging predictions led to a major boost in performance (−2.59 ICM-Soft, 0.36 ICM-Soft norm). This demonstrated that increasing dataset diversity can significantly enhance model understanding. However, further augmentations using EASE-S (synonym replacement) showed a slight decline in performance. Despite increasing the dataset with an additional 6,000 tweets, the model's ICM-Soft dropped to −2.79 with a normalized score of 0.35, suggesting that excessive or noisy augmentation can introduce semantic drift and reduce effectiveness.

EASE-AEDA, which combined EASE-S and AEDA, offered marginally better performance than EASE-S alone (−2.78,0.35), but still underperformed compared to cross-translation. Interestingly, AEDA alone yielded the best results. By selectively augmenting only the underrepresented categories with minor random noise, the model achieved its highest accuracy with −2.51 ICM-Soft and 0.37 ICM-Soft norm. This highlights that AEDA was the best technique for augmentation for this task. This can be summarized in the Table 13.

Table 13
Task 3: Soft-soft Evaluation on Dev Set

Metric	Baseline	Base Model	Cross-Translated	EASE-S	EASE-AEDA	AEDA
ICM-Soft	-8.7	-7.0	-2.59	-2.79	-2.78	**-2.51**
ICM-Soft norm	0.0	0.0	0.36	0.35	0.35	**0.37**

5.4.2. Test Results & Ranking

We submitted three runs on the best-performing augmentation strategy identified during development: AEDA. As shown in Table 14, our rankings were 5th for ALL, 6th for ES, and 6th for EN out of 181 submissions, highlighting the robustness and competitiveness of our models in a multilingual, multi-label classification task. These results reaffirm the strength of the AEDA-based augmentation approach, which is well generalized from the development set to the testing set.

Table 14
Task 3: Soft-soft Evaluation on Test Set & Ranking

Run	Language	Rank	ICM-Soft	ICM-Soft Norm
3	All	5	-2.4632	0.3699
2	ES	6	-2.6287	0.3632
1	EN	6	-2.3548	0.3710

5.5. Task 3: Hard Models

5.5.1. Dev Testing & Experimentation

For the hard classification model, the baseline performance was relatively poor, with an ICM of −1.72, ICM-Norm of 0.11, and an F1-score of 0.10. This served as a reference point for evaluating the impact of various augmentation strategies. Our initial base approach using only the original dataset without augmentation achieved -0.51 ICM, 0.26 ICM-Norm, and 0.27 F1. This was better than the baseline; however, the model was still limited due to insufficient diversity in training data.

Significant improvement was observed with cross-translation, where tweets were translated between English and Spanish to enlarge and diversify the dataset. This resulted in a noticeable performance jump, with an ICM of −0.30, ICM-Norm of 0.43, and F1-score of 0.467. On top of this, AEDA on underrepresented tweets further improved performance to −0.268, 0.44, and 0.478, respectively.

Interestingly, combining both EASE and AEDA (EASE-AEDA) led to a slight drop in performance (−0.29, 0.44, 0.47). While still better than using cross-translation alone, the combination did not yield the expected gains (likely due to noise or redundancy introduced by combining both augmentation techniques). The best results were achieved using EASE-S alone. This method yielded the top scores across all metrics: −0.24 ICM, 0.445 ICM-Norm, and 0.482 F1 (Table 15). This suggests that strategically chosen augmentations can significantly enhance hard multi-label model performance by increasing generalization without overwhelming the semantic integrity of the tweets.

Table 15
Task 3: Hard-Hard Evaluation on Dev Set

Metric	Baseline	Base Approach	Cross-Translated	EASE-S	EASE-AEDA	AEDA
ICM	-1.72	-0.51	-0.3	**-0.24**	-0.29	-0.268
ICM Norm	0.11	0.26	0.43	**0.445**	0.44	0.44
F1	0.10	0.27	0.467	**0.482**	0.47	0.478

5.5.2. Test Results & Ranking

We submitted three test runs for the ICM-Hard evaluation, again with AEDA identified as most effective during development. In contrast to the Soft evaluations, the results for the hard evaluations are not as impressive. The EN model, submitted as Run 1, achieved a 69th-place ranking, slightly lower than our ES model in Run 2, which placed 67th. Despite this ranking difference, both models achieved comparable Macro F1 scores (0.5184 for EN and 0.5210 for ES), suggesting relatively balanced classification performance across both languages.

Run 3, which combined predictions from both EN and ES models, ranked 65th overall, showing a slight improvement in the ICM-Hard Norm (0.4188) and a better Macro F1 of 0.5205. These results indicate that, although our models performed better under soft evaluations, they remained moderate in the hard setting. Since this was a multi-label task, it is highly possible for the model to have been overconfident and predicted more than one label (when there wasn't) or have higher confidence in the wrong classes. Any value above the threshold we set resulted in the label being included, whereas our soft models just compared how near the values were to the original values (not considering the label as absolute if it was). However, the consistency across evaluation types once again reinforces the effectiveness of our AEDA-driven modeling pipeline. The balanced performance in ICM-Hard Norm and Macro F1 further highlights our model's capacity for multilingual generalization in both Spanish and English in a multi-label environment. This can be summarized in Table 16.

Table 16
Task 3: Hard-Hard Evaluation on Test Set & Ranking

Run	Language	Rank	ICM-Hard	ICM-Hard Norm	Macro F1
3	All	65	-0.3499	0.4188	0.5205
2	ES	67	-0.4361	0.4026	0.5210
1	EN	69	-0.2708	0.4336	0.5184

6. Conclusion

This paper presents our approach for the EXIST 2025 Tasks 1.1, 1.2, and 1.3 involving sexism identification and classification in tweets. We employed ensemble architectures and augmentation techniques to generate both soft and hard predictions. Our methodology included cross-translation between English and Spanish to enhance data diversity, along with the data augmentation techniques EASE-S and AEDA. These augmented datasets were utilized across all three tasks.

For Task 1.1, an ensemble approach was adopted, comprising two or three English and Spanish models, which delivered improved scores and demonstrated the benefit of combining multiple models to mitigate individual weaknesses and enhance robustness. In Task 1.2, xlm_RoBERTa trained on a cross-translated and AEDA class-targeted augmented dataset significantly outperformed other approaches, indicating that xlm_RoBERTa variants are well-suited for multilingual sexism-related classification tasks and that augmenting underrepresented classes can help address class imbalance. For Task 1.3, EASE-S performed best on the hard sub-task, whereas AEDA achieved superior results on the soft sub-task.

These findings underscore the importance of carefully selecting models and employing augmentation and ensemble strategies to address linguistic and contextual challenges in sexism classification. We intend to experiment with zero-shot and few-shot learning to improve our hard sub-task models, given their demonstrated performance in the literature. Additionally, we aim to incorporate annotator-specific contextual embeddings and evaluate how different annotator data affect the results. This will help large language models (LLMs) learn the vast diversity of human perspectives and socio-cultural nuances, and integrate them into their predictions.

Acknowledgments

The authors would like to acknowledge the support provided by the Office of Research (OoR) at Habib. University, Karachi, Pakistan, for funding this project through the internal research grant IRG-2235.

Declaration on Generative AI

While preparing this work, the authors used ChatGPT and Grammarly to check grammar and spelling, reword, and paraphrase for clarity. All content was reviewed and edited by the authors, who take full

responsibility for the final manuscript.

References

[1] L. Plaza, J. Carrillo-De-Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025 - learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: Notebook for the EXIST Lab at CLEF 2025, 2025. URL: https://www.damianospina.com/publication/plaza-2025-overview/.

[2] 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 tweets and memes (extended overview), in: Notebook for the EXIST Lab at CLEF 2024, 2024. URL: https://ceur-ws.org/Vol-3740/paper-87.pdf.

[3] A. Azadi, B. Ansari, S. Zamani, Bilingual sexism classification: Fine-tuned xlm-roberta and gpt-3.5 few-shot learning, in: Working Notes of CLEF 2024– Conference and Labs of the Evaluation Forum, 2024.

[4] V. Ruiz, J. C. de Albornoz, L. Plaza, Concatenated transformer models based on levels of agreements for sexism detection, in: Working Notes of CLEF 2024– Conference and Labs of the Evaluation Forum, 2024.

[5] N. Maqbool, Sexism identification in social networks: Advances in automated detection - a report on the exist task at clef, in: Conference and Labs of the Evaluation Forum, 2024.

[6] A. Naebzadeh, M. Nobakhtian, S. Eetemadi, Nica at exist clef tasks 2024: Notebook for the nica group at exist lab at clef 2024, 2024.

[7] M. Usmani, R. Siddiqui, S. Rizwan, F. Khan, F. Alvi, A. Samad, Sexism identification in tweets using bert and xlm-roberta, in: Working Notes of CLEF 2024– Conference and Labs of the Evaluation Forum, 2024.

[8] R. Pan, J. Antonio García-Díaz, T. Bernal-Beltrán, R. Valencia-García, Umuteam at exist 2024: Multi-modal identification and categorization of sexism by feature integration, in: Notebook for the EXIST Lab at CLEF 2024, 2024. URL: https://ceur-ws.org/Vol-3740/paper-106.pdf, retrieved March 1, 2025.

[9] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based mixture of transformers for exist2024, in: Working Notes of CLEF 2024– Conference and Labs of the Evaluation Forum, 2024.

[10] J. Tavarez-Rodríguez, F. Sánchez-Vega, A. Rosales-Pérez, A. P. López-Monroy, Better together: Llm and neural classification transformers to detect sexism, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[11] G. Aru, N. Emmolo, S. Marzeddu, A. Piras, J. Raffi, L. Passaro, Robexedda: Sexism detection in tweets, in: Notebook for the EXIST Lab at CLEF 2024, 2024. URL: https://ceur-ws.org/Vol-3740/paper-88.pdf, retrieved March 1, 2025.

[12] F. Shifat, F. Haider, S. Ul, R. Sourove, D. Dutta Barua, F. Ishmam, M. Fahim, F. Bhuiyan, Penta-nlp at exist 2024 task 1-3: Sexism identification, source intention, sexism categorization in tweets, in: Notebook for the EXIST Lab at CLEF 2024, 2024. URL: https://ceur-ws.org/Vol-3740/paper-114.pdf, retrieved March 1, 2025.

[13] L. M. Quan, D. V. Thin, Sexism identification in social networks with generation-based approach, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[14] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, NYCU-NLP at EXIST 2024 – Leveraging Transformers with Diverse Annotations for Sexism Identification in Social Networks, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[15] A. M. M. Rahman, W. Yin, G. Wang, Data augmentation for text classification with EASE, in: M. Abbas, A. A. Freihat (Eds.), Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023), Association for Computational Linguistics, Online, 2023, pp. 324–332. URL: https://aclanthology.org/2023.icnlsp-1.35/.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[17] 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, arXiv preprint arXiv:1911.02116 (2019).