Sexism Detection in Multilingual Tweets*

Notebook for the EXIST Lab at CLEF 2025

Ahmed Gamal Ibrahim[1,*,†], Rui Pedro Lopes[1,†]

[1]*Research Center in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal*

Abstract

Online sexism is a widespread social issue, found around all social media platforms, in multiple languages. In this paper, we present our submission to the EXIST 2025 task 1 on sexism detection in English and Spanish tweets. To enhance model reliability and deal with data sparsity, we used a combination of multiple text augmentation strategies, including AEDA (punctuation-based), synonym replacement, back-translation, and light code-switching via round-trip translation. These augmentations were applied to diversify training samples and better capture patterns.

Our architecture builds on XLM-RoBERTa-large, fine-tuned for three subtasks: binary sexism detection, source classification, and sexism categorization. We incorporated both soft and hard label strategies to account for annotation disagreement and applied label smoothing and class-weighted loss functions to manage class imbalance. The system was trained and evaluated using official splits, with results showing promising results in multi-label classification (mainly Task 1.3), especially under soft label settings. However, its performance in binary classification (Task 1.1) showed limitations in generalization. Overall, our approach focuses on the importance of multilingual-aware data augmentation and training strategies in building fairer content moderation systems.

Keywords

Sexism detection, Data augmentation, Back translation, Code switching

1. Introduction

Due to the spread of social media use, the dynamics of communication have been altered significantly, which has introduced both opportunities and challenges. Among the latter, the proliferation of sexism in digital discourse, ranging from clear hostility to subtle bias, has become a pressing concern. Detecting and mitigating such content is essential to establishing inclusive digital environments and reducing the real-world harm associated with online gender-based discrimination.

Natural Language Processing (NLP) has consistently been a growing field for many years by now, primarily driven by transformer-based architectures such as BERT [1] and RoBERTa [2]. These models have demonstrated impressive capabilities in handling difficult linguistic tasks, including multilingual and low-resource scenarios, which are particularly relevant in the context of online social media, where code-switching and informal language are prevalent.

The EXIST 2025 task 1 focuses on sexism detection in tweets that are in English or Spanish through three hierarchical subtasks: binary classification, source intention, and sexism categorization. These subtasks present several challenges, including language diversity, class imbalance, annotation uncertainty, and semantic nuances.

In this paper, we present a unified multilingual pipeline built upon XLM-RoBERTa-large to address all three subtasks. Our contributions include an augmentation strategy to use for multilingual data,

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

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

ahmed@ipb.pt (A. G. Ibrahim); rlopes@ipb.pt (R. P. Lopes)

0009-0005-3495-2055 (A. G. Ibrahim); 0000-0002-9170-5078 (R. P. Lopes)

soft-label modeling to incorporate annotator uncertainty, and task-specific loss function design to handle label imbalance. By evaluating our models across Tasks 1.1, 1.2, and 1.3, we provide insights into the effectiveness and limitations of our approach and outline pathways for improving performance in the future.

2. State of the art

Recent advancements in NLP have shed light on the development and extensive use of transformer-based models such as XLM-RoBERTa. Transformer architectures completely changed the game in NLP tasks due to their capacity for modeling long-range dependencies through self-attention mechanisms [3]. XLM-RoBERTa, specifically, displayed significant performance improvements in cross-lingual representation learning, allowing for effective multilingual understanding, particularly beneficial for low-resource languages and complex linguistic scenarios involving code-switching [4, 5]. This model outperformed earlier multilingual models such as mBERT by taking advantage of massive multilingual datasets and complex training techniques [4].

Addressing class imbalance remains a difficult obstacle in machine learning and in NLP. Johnson and Khoshgoftaar [6] conducted a survey about deep learning approaches to manage class imbalance, indicating the necessity of developing effective classification techniques specifically designed for imbalanced datasets prevalent in real-world scenarios like fraud detection, anomaly detection, and medical diagnosis. Buda et al. [7] examined the effects of class imbalance on Convolutional Neural Networks (CNNs), concluding that oversampling consistently outperforms undersampling as it helps maintain balanced class representations, avoiding the risk of overfitting, which often affects smaller training sets. Their experimental analysis provided important guidelines for handling imbalance in diverse neural network architectures.

Data augmentation strategies have greatly amplified NLP models' reliability and generalization capabilities. Back-translation, first systematically introduced by Sennrich et al. [8], capitalizes on monolingual data to augment bilingual training sets, thereby improving translation quality and downstream NLP task performance. This approach has become a cornerstone in augmentation for multilingual models. Moreover, recent explorations into adversarial perturbations and contextual synonym replacement illustrate the importance of data augmentation [9]. They demonstrated how well-designed perturbations could effectively bolster model resilience against adversarial attacks and improve generalization to unseen data distributions, which is an important aspect for deployment in real-world scenarios.

Furthermore, methods such as label smoothing and knowledge distillation have shown major improvements in neural network training and generalization. Label smoothing, examined by Müller et al. [10], reduces model overconfidence, leading to better-calibrated predictions and improved generalization across different classification tasks. However, Müller et al. also noted potential drawbacks, indicating that label smoothing might negatively affect the performance of subsequent knowledge distillation processes. Knowledge distillation, detailed by Hinton et al. [11], simplifies transferring insights from large-scale intensive models into smaller, more efficient architectures, achieving high performance without high computational demands, which comes in handy in practical deployment.

Multilingual NLP tasks greatly benefit from translation frameworks, such as OPUS-MT, which was developed by Tiedemann and Thottingal [12]. OPUS-MT provides open-source translation models and pretrained multilingual resources, allowing for cross-lingual communication and addressing linguistic inequalities, particularly benefiting low-resource languages. Additionally, standardized benchmarks such as LinCE [5] have been important in enabling rigorous evaluation and facilitating improvements in multilingual NLP models. LinCE's datasets and metrics target linguistic code-switching scenarios, supporting the development and validation of multilingual models.

Our research builds upon these foundations, integrating multilingual representation techniques, data augmentation methods, and class imbalance handling to enhance sexism detection capabilities across multilingual and code-switched datasets. The systematic application of these methodologies aim at achieving better accuracy and generalization, especially within complex linguistic contexts.

3. Dataset

The dataset used in this work is released within the EXIST 2025 task 1, which addresses sexism detection in multilingual and code-switched social media text. The dataset consists of Spanish and English tweets, annotated for three hierarchical subtasks: binary sexism identification (Task 1.1), multi-class sexism type classification (Task 1.2), and categorization (Task 1.3). Each tweet is labeled by six annotators, and annotations are complemented with detailed demographic metadata such as gender, age range, country, study level, and ethnicity, ensuring representation diversity across perspectives.

The three subtasks are organized as follows. In Task 1.1, each tweet is labeled as YES (sexist) or NO. Task 1.2 assigns a label among DIRECT, REPORTED, JUDGEMENTAL, or UNKNOWN if the tweet is sexist. Task 1.3 adds finer granularity with categories such as OBJECTIFICATION, STEREOTYPING-DOMINANCE, SEXUAL-VIOLENCE, and IDEOLOGICAL-INEQUALITY.

The dataset is split into training, development, and test partitions, comprising over 6,000 tweets for training, approximately 1000 for development, and over 2000 instances for testing (Table 1). While labels are provided for both training and development sets, the test set is unlabeled and intended for final system evaluation. The tweets are in English and Spanish, which allows for multilingual and cross-lingual modeling.

Table 1
Dataset splits used for training, validation, and final evaluation.

Split	Train	Dev	Test
# Instances	6,920	1,038	2,076

All tweets were preprocessed by removing user mentions, URLs, hashtags, emojis, special characters, and numbers, and by converting text to lowercase. To address class imbalance and improve model generalization, we used several data augmentation techniques including AEDA (insertion of punctuation), contextual synonym replacement using masked language modeling (`xlm-roberta-large`), and back-translation via OPUS-MT pipelines. We also simulated code-switching behavior based on the tweet's language label using round-trip translation. Augmented data was cached and reused across training runs to reduce computational overhead.

Annotation reliability is addressed through majority voting for Tasks 1.1 and 1.2. For Task 1.3, we retained the union of all non-null category labels provided by annotators, enabling training while preserving label diversity. Incomplete or ambiguous labels (e.g., marked as `"-"`) were excluded from loss computation but retained in the dataset for potential analysis or augmentation.

4. Methodology and Implementation

Our approach to the EXIST 2025 task 1 is grounded in a modular training pipeline designed to handle multilingual data, label noise, class imbalance, and cross-task generalization. We implement a unified system that supports Tasks 1.1 (binary classification), 1.2 (multi-class classification), and 1.3 (multi-label classification) using XLM-RoBERTa-large as the backbone architecture. This section provides a walkthrough of our training pipeline, model configuration, augmentation strategy, and optimization procedure.

4.1. Overall Architecture and Data Flow

The core architecture employs XLM-RoBERTa-large, a multilingual transformer pretrained on over 100 languages. We use the HuggingFace Transformers API to load both the tokenizer and the model. A custom PyTorch training loop wraps the forward and backward passes, optimizer steps, and evaluation metrics.

Each tweet is preprocessed to remove mentions, URLs, hashtags, digits, and punctuation. Cleaned tweets are augmented with four variants using AEDA, contextual synonym replacement, code-switch

simulation, and back-translation. The augmented data is stored and cached for efficient reuse.

4.2. Task-Specific Handling

We process each task (1.1, 1.2, 1.3) separately, with both hard and soft labeling strategies:

- **Task 1.1 (Binary)**: Labels are binarized using majority voting. For soft labels, the proportion of "YES" votes among six annotators is used as a float target.
- **Task 1.2 (Multi-class)**: Votes across classes (DIRECT, REPORTED, JUDGEMENTAL) are tallied and converted into either a hard label or a soft distribution (normalized counts).
- **Task 1.3 (Multi-label)**: We use a fixed taxonomy of 7 categories (including 'NO' category for unlabeled instances). Each category is assigned a binary indicator for hard mode or a proportional weight in soft mode.

All tasks share a unified dataset wrapper with dynamic label encoding based on task and mode.

4.3. Augmentation Pipeline

We put four data augmentation techniques into use:

1. **AEDA (An Easier Data Augmentation)**: This technique injects punctuation marks (such as commas, periods, semicolons) at random positions within the token sequence. It was originally proposed to provide syntactic variety with minimal semantic distortion. For instance, the sentence "She thinks women should stay home" might be transformed into "She, thinks women. should stay; home". AEDA is language-agnostic and efficient, and its simplicity helps diversify sentence structure without requiring any additional language resources.
2. **Contextual Synonym Replacement**: Tokens are randomly masked and replaced using predictions from a masked language model (XLM-RoBERTa). This method provides contextual paraphrasing that is semantically aligned with the original sentence. For example, "girls cannot code" might yield "girls cannot program". Unlike naive synonym substitution, this method considers sentence-wide context, leading to fluent and accurate rewrites. The top prediction is selected to ensure semantic coherence.
3. **Code-Switching Simulation**: This strategy randomly selects tokens from a source language (e.g., English) and translates them into a target language (e.g., Spanish), followed by a reverse translation. The goal is to simulate real-world multilingual input where users switch languages mid-sentence. For example, "He is such a macho man" could be transformed to "He is such a hombre macho" and then retranslated to "He is such a macho man" or a variant. This approach reflects authentic usage patterns in bilingual communities and trains the model to handle intra-sentence code-switching.
4. **Back-Translation**: This method uses OPUS-MT neural translation models to perform round-trip translation, i.e., translating from the original language to the other (English to Spanish or vice versa), and then back. This process generates paraphrastic variants that preserve meaning while altering lexical and syntactic structure. For example, "She should not talk so loud" could become "Ella no debería hablar tan fuerte" and back to "She shouldn't speak so loudly". This augmentation enriches the dataset with diverse linguistic constructions and improves generalization to unseen phrasings.

Each original training tweet produces five total versions (1 original + 4 augmentations).

4.4. Model Configuration

Each task is trained using XLM-RoBERTa with a classification head. For Tasks 1.1 and 1.3, the output layer is a single-unit or multi-sigmoid layer respectively, while Task 1.2 uses a softmax head with class weights.

The model uses dropout in both the attention and feed-forward layers with a probability of 0.1. We adopt the AdamW optimizer with a learning rate of 3e-5, weight decay of 0.01, and a linear warmup over 10% of total steps.

4.5. Handling Imbalance and Label Smoothing

To mitigate label imbalance and improve generalization, we applied task-specific strategies for loss functions and target smoothing:

- **Task 1.1 (Binary Classification)**: In soft mode, labels were constructed by computing the proportion of "YES" annotations among six annotators. We used `SmoothBCEWithLogits`, a custom smoothed binary cross-entropy loss, which regularizes model confidence by slightly adjusting soft targets. In hard mode, we applied `BCEWithLogitsLoss` with a positive class weight inversely proportional to the frequency of the "YES" class to address the strong class imbalance.
- **Task 1.2 (Multi-Class Classification)**: For soft mode, we converted annotator votes into a normalized probability distribution and trained using KL-divergence loss (`SoftKLDivLoss`) to align model outputs with partial annotator consensus. In hard mode, we used `CrossEntropyLoss` with class weights computed from label frequencies to prevent dominant categories like `DIRECT` from overshadowing less frequent ones such as `REPORTED` and `JUDGEMENTAL`.
- **Task 1.3 (Multi-Label Classification)**: In soft mode, each category label was assigned a real-valued target based on the proportion of annotators selecting it. Unlike Task 1.2, we did not apply class weighting in the multi-label task and relied solely on smoothing to mitigate label imbalance. In hard mode, we used standard `BCEWithLogitsLoss` without additional class reweighting.

4.6. Training and Evaluation

Training was conducted for 10 epochs with early stopping based on the F1 score on the validation set. We compute micro-F1 and accuracy after each epoch. The best checkpoints were saved and reloaded if they surpass previous performance.

Batch sizes were set to 16. A learning rate scheduler adjusts the rate linearly with warmup. All computations were performed on a CUDA-enabled GPU when available.

5. Results and Discussion

5.1. Task 1.1: Binary Sexism Classification

For Task 1.1, which involved classifying tweets as sexist or not, our approach integrated label smoothing, class weighting, and five-version data augmentation (original + 4 augmented variants). Despite these techniques, our model did not appear in the top ranks of the leaderboard. Upon inspecting the results, we confirm that our system underperformed relative to other performers that managed to achieve higher normalized scores.

Given the training logic in our implementation, including soft-target binarization based on annotator agreement and smoothed binary cross-entropy loss, this performance gap may be attributable to either:

- Limited generalization due to overfitting on augmented variants.
- Conservative thresholding strategies during sigmoid post-processing.

5.2. Task 1.2: Multi-Class Sexism Type Classification

In Task 1.2, which required identifying the type of sexism (`DIRECT`, `REPORTED`, `JUDGEMENTAL`), our system performed better, especially in soft labeling. The leaderboard shows that `exist@Cedri` obtained

a Cross Entropy score of **3.7432** and an ICM-Soft Norm score of **0.3279**, ranking better in comparison to the 1.1 task.

Our architecture used soft targets computed from annotator votes and employed KL-divergence loss to model partial agreement. This allowed the model to maintain sensitivity toward ambiguous or multi-class examples. However, performance might have been negatively affected by:

- A skewed class distribution, with a dominance of DIRECT labels.

Class reweighting was applied during training to balance this skew, but additional techniques could have further improved the performance.

5.3. Task 1.3: Multi-Label Fine-Grained Categorization

Task 1.3 required the identification of multiple sexism-related labels per tweet. Our approach applied sigmoid classification with soft targets reflecting label agreement across annotators. The system performed well in soft labeling, getting a normalized score of **0.3193** in ICM-Soft Norm, placing within the top ten of submissions.

We attribute this relative success to:

- The application of `SmoothBCEWithLogits` to prevent overconfidence in rare category predictions.

However, the system still suffered from:

- Overfitting on majority labels due to their prevalence.
- Underrepresentation of specific categories, which limited recall despite good precision.

5.4. Overall Assessment and Future Directions

Our architecture showed reliable performance in some cases (such as task 1.3 soft labels), while it heavily underperformed in other cases (such as task 1.1 soft labels). That mainly tells us that there is a decent basis but also a huge room for improvement in the future.

Nonetheless, future work could benefit from:

- Exploring transformer variants optimized for multi-label settings.
- Conducting more extensive hyperparameter sweeps and specific augmentation removal studies to quantify the marginal utility of each augmentation method.

6. Conclusion

This paper introduced a multilingual sexism detection system based on the XLM-RoBERTa-large architecture, evaluated across all three subtasks of task 1 in the EXIST 2025 challenge. Our approach incorporated a modular training pipeline with 4 different data augmentation techniques, soft-label strategies to capture annotator uncertainty, and class imbalance-aware optimization procedures.

While the system demonstrated promising results in multi-label classification (mainly Task 1.3), especially under soft label settings, its performance in binary classification (Task 1.1) revealed limitations in generalization. These outcomes show the complexity of building reliable classifiers for nuanced social phenomena, especially when labels are inherently subjective or sparsely represented.

Through the analysis of performance across tasks, we identified that methods such as SmoothBCE and KL-divergence-based loss functions provided tangible benefits in modeling partial consensus among annotators. However, further improvements may rely on more aggressive regularization, tuning of augmentation techniques, and the exploration of task-specialized transformer variants.

Future work will focus on incorporating uncertainty-aware threshold calibration, the use of multilingual large language models in a few-shot or instruction-tuned setting, and conducting studies to isolate the contributions of each augmentation and loss formulation. Our findings contribute to the broader field of bias detection in NLP, offering insights into how multilingual architectures can be adapted to address socially impactful classification tasks.

7. Acknowledgments

The authors are grateful to the Foundation for Science and Technology (FCT, Portugal) for financial support through national funds FCT/MCTES (PIDDAC) to CeDRI, UIDB/05757/2020 (DOI: 10.54499/UIDB/05757/2020) and UIDP/05757/2020 (DOI: 10.54499/UIDP/05757/2020) and SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

Declaration on Generative AI

During the preparation of this work, the authors used AI tools for grammar and spelling checks. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: http://arxiv.org/abs/1706.03762. doi:10.48550/arXiv.1706.03762. arXiv:1706.03762 [cs].

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[5] G. Aguilar, S. Kar, T. Solorio, LinCE: A centralized benchmark for linguistic code-switching evaluation, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 1803–1813. URL: https://aclanthology.org/2020.lrec-1.223/.

[6] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance 6 (2019) 27. URL: https://doi.org/10.1186/s40537-019-0192-5. doi:10.1186/s40537-019-0192-5.

[7] M. Buda, A. Maki, M. A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks 106 (2018) 249–259. URL: https://www.sciencedirect.com/science/article/pii/S0893608018302107. doi:10.1016/j.neunet.2018.07.011.

[8] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, 2016. URL: http://arxiv.org/abs/1511.06709. doi:10.48550/arXiv.1511.06709. arXiv:1511.06709 [cs].

[9] T. Gokhale, S. Mishra, M. Luo, B. Sachdeva, C. Baral, *Generalized but not Robust?* comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, 2020, pp. 2705–2718. URL: https://aclanthology.org/2022.findings-acl.213/. doi:10.18653/v1/2022.findings-acl.213.

[10] R. Müller, S. Kornblith, G. Hinton, When does label smoothing help?, 2020. URL: http://arxiv.org/abs/1906.02629. doi:10.48550/arXiv.1906.02629. arXiv:1906.02629 [cs].

[11] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015. URL: http://arxiv.org/abs/1503.02531. doi:10.48550/arXiv.1503.02531. arXiv:1503.02531 [stat].

[12] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada (Eds.), Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61/.