DUTH at EXIST 2025: Multilingual Sexism Detection with Soft Labels and Transformers

Georgios Arampatzis*[1], Vasileios Perifanis[1], Symeon Symeonidis [2] and Avi Arampatzis[1]

[1]Democritus University of Thrace, Department of Electrical and Computer Engineering, Xanthi, Greece
[2]Democritus University of Thrace, Department of Production and Management Engineering, Xanthi, Greece

Abstract

This paper presents the DUTH system for the EXIST 2025 shared task on multilingual sexism detection. The task comprises three subtasks applied to a multilingual tweet corpus annotated with both hard and soft labels: (i) binary classification of sexist vs. non-sexist content, (ii) single-label classification of the type of sexism, and (iii) multi-label classification of the intended sexism category. The proposed system employs a transformer-based multilingual architecture, fine-tuned using techniques such as oversampling, class weighting, and soft-label learning to address class imbalance and annotator disagreement.

Our system demonstrates robust performance in binary sexism detection, particularly on Spanish data, achieving competitive results under both hard and soft evaluation metrics. However, performance on the more nuanced subtasks—classifying the type and intent of sexist speech—remains limited, underscoring the difficulty of modeling implicit and context-sensitive expressions of sexism. We analyze these challenges and propose future directions, including discourse-aware modeling, hierarchical label representations, and multimodal learning.

Keywords

Sexism Detection, Transformer Models, Soft Labels, Multi-label Classification

1. Introduction

Sexism remains prevalent in online discourse, often disguised **through implicit or veiled expressions**, which complicates automated detection efforts [1]. Social media platforms frequently exacerbate this issue by amplifying such content [2]. Consequently, effective computational approaches are essential to meet the growing need for identifying gender-based discrimination.

Detecting sexism automatically is inherently challenging due to linguistic ambiguity, annotator subjectivity, and cultural variation in its expression [3, 4]. The EXIST 2025 shared task tackles these challenges by providing a multilingual benchmark dataset consisting of tweets, memes, and TikToks annotated along multiple sexism-related dimensions [5, 6].

In this paper, we describe our participation in Task 1, which focuses on tweets and includes three subtasks. We adopt a multi-model architecture based on transformer models fine-tuned for multilingual and multi-label classification.

Previous work in offensive language and toxic comment detection has evolved from rule-based systems to deep learning architectures [7]. Transformer models such as BERT and its variants have shown strong performance across NLP classification tasks, including sentiment analysis, stance detection, and toxicity recognition [8].

Multilingual transformers like mBERT and XLM-R are particularly effective in cross-lingual scenarios with limited annotated data [9]. The EXIST series highlights the complexities of modeling sexism, especially in the presence of annotator disagreement [4]. Recent approaches to address this include soft labeling, uncertainty modeling, and disagreement-aware learning [3].

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

*Corresponding author.

†These authors contributed equally.

geoaramp@ee.duth.gr (G. Arampatzis*); vperifan@ee.duth.gr (V. Perifanis); ssymeoni@pme.duth.gr (. S. S.); avi@ee.duth.gr (A. Arampatzis)

0009-0003-3840-4537 (G. Arampatzis*); 0000-0003-3915-9628 (V. Perifanis); 0000-0003-2415-4592 (. S. S.); 0000-0003-2415-4592 (A. Arampatzis)

Our prior work on multilingual affective analysis [10] informed our modeling strategy in this task, underlining the effectiveness of ensemble and hybrid models in tackling nuanced cross-lingual phenomena such as sexism. Building on these insights, we structure our study as follows: Section 2 describes the dataset, including annotation methodology and statistical distributions, followed by our implementation environment and modeling approach. Section 3 presents the experimental results, evaluation metrics, and a detailed subtask analysis. Finally, Section 4 summarizes our findings and outlines future research directions.

2. Approach

2.1. Dataset

The EXIST 2025 dataset is a multilingual corpus of tweets annotated for three subtasks: binary sexism detection, intention classification, and multi-label sexism type categorization. Annotations were collected from multiple annotators per instance, with soft labels derived from aggregated votes. Tweets are annotated in both English and Spanish, providing a realistic and culturally diverse corpus.

Table 1: Statistics for Task 1.1 (Sexist vs Non-Sexist)

	Sexist (Yes)	Non-Sexist (No)
Train	18,753	22,767
Dev	2,998	3,230

Table 1 presents the label distribution for Task 1.1, which involves binary classification of tweets as either sexist (Yes) or non-sexist (No). The training set contains 18,753 sexist and 22,767 non-sexist instances, while the development set includes 2,998 and 3,230 examples, respectively. The overall distribution is relatively balanced, with a slight majority of non-sexist examples.

Table 2: Statistics for Task 1.2 (Sexism Type)

	Reported	Judgmental	Direct	Unknown
Train	4,652	5,015	9,004	82
Dev	757	863	1,378	0

Table 2 shows the label distribution for Task 1.2, which focuses on classifying the type of sexism expressed in tweets. The categories include: Reported Speech, Judgemental, Direct, and Unknown. In the training set, most sexist tweets fall under the Direct category (9,004 instances), followed by Judgemental (5,015) and Reported Speech (4,652), with only a small number labeled as Unknown (82). The development set follows a similar pattern. This distribution reflects a prevalence of explicit sexist content in the dataset and indicates that the Direct category is the most dominant, which may influence model learning and performance.

Table 3: Statistics for Task 1.3 (Sexism Intentions)

	Intentional	Unintentional	Ideological	Non-Sexist
Train	10,587	8,391	8,778	22,767
Dev	1,643	1,408	1,419	3,230

Table 3 summarizes the label counts for Task 1.3, which involves multi-label classification of sexism intentions. Each tweet can be annotated with one or more of the following categories: *Intentional*, *Unintentional*, *Ideological*, and *Non-Sexist*. In the training set, *Intentional* sexism is the most common category (10,587 tweets), closely followed by *Ideological* (8,778) and *Unintentional* (8,391). The *Non-Sexist* class corresponds to tweets without sexist content (22,767 instances) and can co-occur with the others due to the soft-label nature of the task. This distribution suggests that intentionality and ideology are prominent aspects of sexist expression in the dataset.

The test set consists of 2,076 tweets annotated with soft labels for all three subtasks. For Task 1.1, each instance includes a probability distribution over the binary classes (Sexist vs. Non-Sexist). For Tasks 1.2 and 1.3, soft multi-label annotations are provided for each of the respective categories. These probabilistic labels capture annotator disagreement and are intended to support evaluation methods beyond traditional classification metrics.

2.2. Implementation and Environment

All experiments were conducted in `Python 3.10`, using the HuggingFace `Transformers` library and the `PyTorch` framework.

The core software stack included `transformers` (v4.38.1) for model loading and fine-tuning, `datasets` (v2.18.0) for data handling, `scikit-learn` (v1.4.2) for evaluation metrics, and `pandas` (v2.2.1) and numpy (v1.26.4) for data manipulation. We used `torch` (v2.2.0) as the main backend for deep learning operations. Auxiliary libraries such as `accelerate`, `evaluate`, `tqdm`, `json`, and `argparse` supported training and evaluation.

The implementation supports soft-label training, class reweighting, and multi-label classification where applicable. Annotator labels were manually preprocessed into hard or probabilistic targets according to the task requirements.

2.3. Methodology

Task 1.1 – Binary Sexism Detection

For Task 1.1, we formulated the problem as a binary classification task, aiming to distinguish between sexist and non-sexist tweets. We filtered the training instances to retain only those with annotations from at least three annotators, and assigned hard labels based on majority vote. To address class imbalance, we applied oversampling to the minority class (sexist instances). We employed the multilingual `xlm-roberta-large` transformer model and fine-tuned it using a custom training routine with class-weighted cross-entropy loss to mitigate bias toward the majority class. Hyperparameters were tuned using stratified training-validation splits and early stopping based on F1-score.

Task 1.2 – Sexism Type Classification

Task 1.2 involves single-label classification of sexist tweets into three categories: *Reported Speech*, *Judgmental*, and *Direct*. Instances were labeled according to the most frequently selected category among annotators. Due to skewed class distributions, we balanced the training data via oversampling to ensure equal representation across categories. The `cardiffnlp/twitter-xlm-roberta-base` model was fine-tuned using a custom training pipeline that dynamically computed class weights based on the frequency of each label in the training set. Optimization was guided by macro-averaged F1-score, and early stopping was applied to prevent overfitting.

Task 1.3 – Multi-label Sexism Intention Classification

For Task 1.3, we treated the classification of sexism intentions as a multi-label problem, where tweets could be associated with one or more categories from a predefined set. We performed label normalization to unify semantically overlapping tags and filtered out inconsistently annotated or ambiguous instances. To alleviate class imbalance, we applied targeted data augmentation using paraphrased versions of underrepresented instances. We fine-tuned a multilingual `xlm-roberta-base` model with sigmoid activation on the output layer and binary cross-entropy loss. The model was trained using stratified sampling and evaluated with micro-averaged F1-score.

2.4. Training Details

All systems were developed using the Hugging Face `Transformers` library with a PyTorch backend. Stratified training-validation splits (typically 80/20 or 90/10) were used to preserve label distributions. Early stopping was employed to prevent overfitting, with patience values ranging from 2 to 3 epochs. Below, we outline the specific hyperparameter settings and preprocessing strategies adopted per subtask.

Task 1.1 – Binary Sexism Detection. We employed the `xlm-roberta-large` transformer model, fine-tuned using the AdamW optimizer. Training data was filtered to retain examples with at least three annotators and binarized via majority vote. Minority class oversampling and a class-weighted cross-entropy loss were used to address label imbalance. *Learning rate*: 1e−5 *Batch size*: 4 *Epochs*: up to 10 (early stopping patience: 2) *Max sequence length*: 128 *Class weights*: [1.0, 1.3]

Task 1.2 – Sexism Type Classification. The `cardiffnlp/twitter-xlm-roberta-base` model was fine-tuned using class-balanced oversampling and a dynamically computed class-weighted cross-entropy loss. *Learning rate*: 1e−5 *Batch size*: 4 *Epochs*: 6 (early stopping patience: 2) *Max sequence length*: 128 *Loss weighting*: inverse label frequency (normalized)

Task 1.3 – Multi-label Intention Classification. We used `xlm-roberta-base` in a multi-label setup with sigmoid activation and binary cross-entropy loss. Label normalization and label-aware paraphrasing were applied to address semantic overlap and class imbalance. *Learning rate*: 2e−5 *Batch size*: 8 *Epochs*: 3 *Max sequence length*: 128 *Augmentation strategy*: paraphrasing underrepresented categories to at least 300 examples per class

All models were evaluated using macro- or micro-averaged F1-score depending on the task. Mixed-precision (FP16) training was enabled when supported by the hardware.

3. Results

3.1. Evaluation Metrics

The evaluation of submitted systems in EXIST 2025 relies on a diverse set of metrics tailored to the nature of each subtask.

Information Contrast Measure (ICM): ICM is a hierarchical-aware metric that compares predicted and gold labels by incorporating the semantic distances between hierarchical classes [11]. It is particularly suitable for hard-label classification tasks involving taxonomies.

ICM-soft extends ICM to the soft-label setting by evaluating predicted probability distributions against annotator consensus distributions. It rewards models that capture annotator uncertainty and disagreement, aligning with recent trends in disagreement-aware learning and probabilistic labeling [3].

F1-score is used in subtasks with binary or imbalanced classification. It is defined as the harmonic mean of precision and recall, and may be reported per class or for the positive class (*YES*) depending on the evaluation protocol [12].

Cross-Entropy measures the divergence between predicted and reference probability distributions, offering insight into the probabilistic calibration of classifiers. It is particularly relevant for soft-label and uncertainty-based modeling [13].

3.2. Experimental Results

To assess system performance, the EXIST 2025 organizers adopted metrics that reflect both accuracy and agreement with annotator uncertainty.

ICM-Soft (Inter-Class Matching – Soft) is a divergence-based metric that compares predicted distributions with soft gold labels representing annotator consensus. It rewards systems that approximate the degree of disagreement among annotators rather than enforcing a single hard label [14, 15].

Normalized ICM-Soft rescales ICM-Soft relative to a random baseline, producing values between 0 and 1 for easier interpretation. A higher score indicates stronger alignment with annotators.

Cross Entropy measures the average divergence between predicted and true soft label distributions. Lower values signify better probabilistic calibration and alignment with annotator judgments.

These metrics, drawn from recent research in learning with disagreement [14, 15], are particularly suited to tasks involving subjective or multi-annotator data such as sexism detection.

Table 4: Task 1.1 – Evaluation Summary for Team DUTH

Instance Set	**ICM-Hard**	**ICM-Hard Norm**	**F1 YES**	**ICM-Soft**	**ICM-Soft Norm**	**Cross Entropy**
ALL	0.4628	0.7326	0.7432	0.1960	0.5314	2.1029
ES	0.4720	0.7360	0.7656	0.2949	0.5473	2.0821
EN	0.4374	0.7232	0.7126	0.0078	0.5013	2.1263

Table 4 presents a comparative summary of the performance of the DUTH system across all, Spanish (ES), and English (EN) instances in Task 1.1, evaluated under both hard and soft settings.

For the **ALL instances**, the system demonstrates balanced performance. The normalized hard-label score (ICM-Hard Norm = 0.7326) and a relatively high F1 score for the YES class (0.7432) indicate consistent behavior in detecting positive instances across both languages. The soft-label performance (ICM-Soft = 0.1960; ICM-Soft Norm = 0.5314) and reasonably low cross-entropy (2.1029) reflect fair model calibration and uncertainty estimation.

On **Spanish (ES) instances**, the system performs best. It achieves the highest F1 YES score (0.7656), with both hard (ICM-Hard = 0.4720; Norm = 0.7360) and soft (ICM-Soft = 0.2949; Norm = 0.5473) metrics supporting its robustness. The lowest cross-entropy (2.0821) further underscores the model's confident and accurate predictions in Spanish.

In contrast, **English (EN) instances** show relatively weaker performance. Despite a very low ICM-Soft score (0.0078), suggesting high confidence, the corresponding F1 YES score (0.7126) and ICM-Hard (0.4374) are the lowest among subsets. This disparity may indicate overconfidence or miscalibration in English-language predictions.

In summary, while the system maintains overall stable performance, it exhibits notably stronger results on Spanish instances—highlighting possible language-specific biases or training data imbalances that warrant further attention.

Table 5: Task 1.2 – Evaluation on All Instances for Team DUTH

System	**ICM-Hard**	**M-Hard No**	**Macro F1**	**ICM-Soft**	**M-Soft Norm**	**Cross Entropy**
ALL	-1.8988	0.0000	0.1967	-18.5641	0.0000	7.3212

Our system exhibited limited effectiveness on Task 1.2, achieving a Macro F1 score of only **0.1967**, which reflects poor balance across the three target classes. This result suggests that the model struggled particularly with identifying minority categories, especially those involving subtle or non-explicit expressions of sexist intent. Additionally, the high Cross Entropy value (**7.3212**) indicates substantial uncertainty and miscalibration in the model's probabilistic outputs.

These outcomes are not unexpected given the inherent complexity of Task 1.2. Unlike binary classification, this task requires the ability to distinguish between nuanced forms of sexism—such as ideological versus unintentional intent—and to interpret implicit language cues embedded in varying cultural and social contexts. Such subtleties often challenge general-purpose text encoders, which may lack the inductive bias needed to generalize over pragmatic and contextual features.

Table 6: Task 1.3 – Evaluation on All Instances for Team DUTH

System	ICM-Hard	M-Hard No	Macro F1	ICM-Soft	ICM-Soft Norm	Cross Entropy
ALL	-1.5980	0.1289	0.3897	-25.9339	0.0000	–

In Task 1.3, our system obtained a **Macro F1 score of 0.3897**, indicating moderate performance in distinguishing between the multiple categories associated with sexist intent. While the model captures certain patterns in the data, it struggles to generalize across all intent types in a balanced manner.

This relatively low score is not unexpected, given the inherent difficulty of Task 1.3, which involves not only identifying the presence of sexist content but also inferring the underlying intention—a subjective and highly context-sensitive construct. Distinguishing between intentional, unintentional, ideological, and non-sexist statements requires sensitivity to pragmatic cues, cultural nuances, and discourse-level features that go beyond surface-level lexical signals.

Our models showed robustness in binary classification but underperformed in the nuanced distinctions required by Tasks 1.2 and 1.3. The use of standard transformers, without explicit modeling of label hierarchy or annotator disagreement, likely contributed to the poor handling of ambiguous or multi-intent tweets. Performance was particularly limited in cases with rare label co-occurrence or high inter-annotator variance.

We hypothesize that incorporating hierarchical label modeling, disagreement-aware loss functions, and graph-based representation learning could substantially improve performance. Error analysis also highlighted the need for pragmatic and discourse-level features, which were lacking in our current token-level input representations.

3.3 Results Analysis

The experimental results across the three subtasks reveal several insights regarding the capabilities and limitations of our system.

In **Task 1.1 (Binary Sexism Detection)**, the system demonstrated robust performance, particularly on the Spanish dataset. The highest normalized ICM-Hard and F1-YES scores across all language subsets suggest stronger alignment with the linguistic characteristics of Spanish sexist content. We hypothesize that greater consistency of lexical cues in Spanish tweets, coupled with the model's cross-lingual generalization capabilities, contributed to this outcome. In contrast, the relatively lower performance on English may reflect increased linguistic ambiguity or higher annotation noise.

In **Task 1.2 (Sexism Type Classification)**, the system performed considerably worse. The macro-averaged F1 score of **0.1967** indicates substantial class imbalance and difficulty in distinguishing between closely related categories such as *Judgemental* and *Reported Speech*. The high cross-entropy further suggests that the model was poorly calibrated, frequently producing overconfident but incorrect predictions. The lack of explicit contextual signals in short tweet texts likely impeded its ability to disambiguate intent-related expressions.

In **Task 1.3 (Sexism Intention Classification)**, the system achieved moderate performance (**Macro F1 = 0.3897**) but struggled with overlapping labels and fine-grained distinctions. The task's multi-label nature, which requires handling interdependent and co-occurring classes, posed significant challenges. The absence of structured label modeling may have further constrained performance. Moreover, the extremely low or negative ICM-Soft scores highlight a misalignment with annotator disagreement, underlining the complexity of learning from soft-label distributions in subjective contexts.

Overall, while our system was effective at identifying explicit forms of sexism, it underperformed when required to infer nuanced, context-dependent phenomena such as intention or ideological framing. These findings are consistent with prior observations that transformer-based models, though strong in binary classification, benefit from extensions such as discourse-aware architectures, hierarchical label modeling, and pragmatic signal integration when applied to subjective or multi-dimensional annotation schemes.

4. Conclusion and Future Work

In this paper, we presented our approach for the EXIST 2025 shared task on multilingual sexism detection, addressing three subtasks involving binary, single-label, and multi-label classification. Our architecture leveraged transformer-based multilingual models trained with both hard and soft labels to accommodate the subjectivity and annotator disagreement inherent in the dataset [3, 16].

Our system achieved robust performance in Task 1.1, particularly on Spanish instances, demonstrating strong alignment with annotator labels in both hard and soft evaluation metrics. However, results on Tasks 1.2 and 1.3 revealed significant challenges in modeling subtle forms of intent and distinguishing fine-grained classes under conditions of low inter-annotator agreement. These tasks require more than lexical matching—understanding pragmatic cues, intent, and socio-linguistic context is essential [17].

For future work, we aim to enhance the modeling of subjective and ambiguous instances by integrating hierarchical and graph-based label representations [18]. We also plan to incorporate discourse-aware and pragmatics-driven features, possibly through large language models with conversational grounding or attention to speaker roles and framing. Agreement-aware loss functions and uncertainty modeling will be further explored to better align model behavior with the soft-label structure of the dataset.

5. Acknowledgements

We thank the organizers of the EXIST 2025 shared task for providing multilingual data and a reliable, transparent evaluation framework that enables fair, meaningful, and reproducible comparisons. Their contribution has been instrumental in advancing research on computational modeling of sexism, subjectivity, and annotator agreement.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Gasparini, et al., Memes as carriers of sexist ideologies, in: Digital Platforms and Feminist Politics, Palgrave Macmillan, 2021.

[2] M. Zampieri, et al., Predicting the type and target of offensive posts in social media, in: Proceedings of NAACL, 2019.

[3] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110. doi:10.1162/tacl_a_00454.

[4] E. W. Pamungkas, et al., Exist 2023 at clef: Incorporating author's intention and learning with disagreements for sexism detection, in: CLEF Working Notes, 2023.

[5] 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: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[6] 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 (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[7] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Toxicity detection: Does context really matter?, in: Proceedings of the 59th ACL (Volume 1: Long Papers), 2021, pp. 3341–3353. doi:10.18653/v1/2021.acl-long.264.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT 2019, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[9] A. Conneau, et al., Unsupervised cross-lingual representation learning at scale (xlm-r), in: ACL 2020, 2020, pp. 8447–8461. doi:10.18653/v1/2020.acl-main.747.

[10] G. Arampatzis, V. Perifanis, S. Symeonidis, A. Arampatzis, DUTH at SemEval-2023 Task 9: An Ensemble Approach for Twitter Intimacy Analysis, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, 2023, pp. 1225–1230. URL: https://aclanthology.org/2023.semeval-1.170. doi:10.18653/v1/2023.semeval-1.170.

[11] G. Angulo, et al., Hierarchical evaluation of classifiers with the information contrast model, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021.

[12] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (2009) 427–437.

[13] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, 2017.

[14] A. Uma, J. Baan, B. Plank, Learning from disagreement: A survey, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, 2021.

[15] J. Baan, A. Uma, B. Plank, Learning from label disagreement in natural language processing, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022.

[16] M. Forbes, A. Zhang, Y. Choi, Limitations of modeling annotator disagreement: A case study in hate speech detection, in: Findings of EMNLP, 2021.

[17] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: Proceedings of ACL, 2020.

[18] Y. Choi, J. Lee, J. Choi, S.-g. Lee, Deepertag: Disentangling hierarchical classification with discriminative learning of label structures, in: Proceedings of AAAI, 2021.