Mario at EXIST 2025: A Simple Gateway to Effective Multilingual Sexism Detection

Notebook for the EXIST Lab at CLEF 2025

Lin Tian[1], Johanne R. Trippas[2] and Marian-Andrei Rizoiu[1]

[1]University of Technology Sydney, Sydney, Australia
[2]RMIT University, Melbourne, Australia

Abstract

This paper presents our approach to EXIST 2025 Task 1, addressing text-based sexism detection in English and Spanish tweets through hierarchical Low-Rank Adaptation (LoRA) of Llama 3.1 8B. Our method introduces conditional adapter routing that explicitly models label dependencies across three hierarchically structured subtasks: binary sexism identification, source intention detection, and multilabel sexism categorization. Unlike conventional LoRA applications that target only attention layers, we apply adaptation to all linear transformations, enhancing the model's capacity to capture task-specific patterns. In contrast to complex data processing and ensemble approaches, we show that straightforward parameter-efficient fine-tuning achieves strong performance. We train separate LoRA adapters (rank=16, QLoRA 4-bit) for each subtask using unified multilingual training that leverages Llama 3.1's native bilingual capabilities. The method requires minimal preprocessing and uses standard supervised learning. Our multilingual training strategy eliminates the need for separate language-specific models, achieving 1.7-2.4% F1 improvements through cross-lingual transfer. With only 1.67% trainable parameters compared to full fine-tuning, our approach reduces training time by 75% and model storage by 98%, while achieving competitive performance across all subtasks (ICM-Hard: 0.6774 for binary classification, 0.4991 for intention detection, 0.6519 for multilabel categorization).

Keywords

Sexism Detection, Low-Rank Adaptation, Hierarchical Classification, Social Media Analysis

1. Introduction

Everyday sexism – ranging from overt misogyny to subtle and implicit forms of gendered microaggressions – undermines women's psychological well-being, silences their voices, and perpetuates structural inequality in digital spaces [1, 2]. Social networks, while instrumental in mobilizing feminist activism through movements like #MeToo, #8M, and #Time'sUp, are also vehicles for the large-scale dissemination of harmful stereotypes and normalized discrimination. Recent research has demonstrated the concerning rise of harmful discourse during crisis events [3], highlighting the urgent need for robust detection systems that can identify not only explicit sexism but also the subtle ways gender-based discrimination infiltrates mainstream online discussions [4].

The scale of this problem demands automated solutions. Manual content moderation cannot keep pace with the billions of posts generated daily [5], yet existing detection systems often fail to capture the nuanced ways sexism manifests online. Context matters: a tweet reporting sexist experiences differs fundamentally from one perpetrating sexism, though both may contain similar language. Cultural and linguistic variations further complicate detection, as sexist expressions evolve rapidly and differ across communities. Research on behavioral homophily has shown that users can exhibit similar engagement patterns when discussing different topics [6], suggesting that understanding user intent and content categorization requires modeling hierarchical label dependencies. These challenges necessitate sophisticated approaches that can distinguish whether content is sexist, understand its intent, categorize its specific manifestation, and operate effectively across languages.

The EXIST 2025 shared task [7] provides a comprehensive framework for advancing sexism detection

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
Lin.Tian-3@uts.edu.au (L. Tian); j.trippas@rmit.edu.au (J. R. Trippas); Marian-Andrei.Rizoiu@uts.edu.au (M. Rizoiu)


research. For the first time, the task spans three modalities (text, images, and videos) and two languages (English and Spanish), reflecting the multimodal and multilingual nature of contemporary social media. We focus on Task 1, which addresses text-based sexism detection through three hierarchically structured subtasks: *(i)* binary sexism identification – determining whether content contains sexism; *(ii)* source intention classification – distinguishing between direct sexism, reported experiences, and judgmental commentary; and *(iii)* sexism type categorization – classifying content into specific categories such as ideological inequality, stereotyping, objectification, sexual violence, and misogyny.

Traditional approaches to these tasks have relied on task-specific models, often struggling with the hierarchical dependencies between subtasks and requiring separate systems for each language. Building on advances in multi-task learning for hate speech detection [8, 9], we present a unified framework leveraging Low-Rank Adaptation (LoRA) [10] of Llama 3.1 8B [11] that addresses all three subtasks simultaneously while maintaining computational efficiency. Our key innovation lies in hierarchical label-aware routing, where LoRA adapters are conditionally activated based on parent-task predictions, explicitly modeling the structured relationships between tasks.

2. Related Work

The automatic detection of sexism on social media has gained increasing attention within the natural language processing (NLP) community, motivated by the need to mitigate online harassment and promote equitable digital discourse. This section reviews existing work in four parts: *(i)* evolution of sexism detection tasks and methodologies, *(ii)* advances in text-based classification approaches, *(iii)* recent developments in large language model (LLM) adaptation for this domain, and *(iv)* harmful content detection and moderation research that informs our understanding of sexism as part of the broader landscape of online harmful content.

2.1. Evolution of Sexism Detection Tasks

Early work in online sexism detection focused primarily on binary classification of overtly hateful content. Talat and Hovy [1] proposed with their Twitter dataset distinguishing sexist, racist, and neutral content. However, researchers quickly recognized that sexism manifests across a spectrum from explicit misogyny to subtle linguistic biases, necessitating more nuanced approaches.

The EXIST shared tasks have been instrumental in advancing the field since 2021 [12, 13, 14]. These tasks progressively introduced hierarchical classification schemes, distinguishing between sexism identification, intention categorization, and fine-grained typing. Similarly, the SemEval-2023 Task 10 on Explainable Detection of Online Sexism (EDOS) [15] focused on interpretability alongside detection accuracy. In addition, the field has evolved from feature-engineered approaches using lexicons and n-grams [16] to neural architectures. Early deep learning approaches used CNNs and LSTMs [17, 18], achieving strong improvements over traditional classifiers. The introduction of transformer-based models marked another paradigm shift, with BERT [19] and its variants becoming the de facto standard for sexism detection tasks [20].

Recent work has explored multi-task learning frameworks to jointly model related tasks. For example, Samory et al. [21] showed that jointly learning sexism and racism detection improves performance on both tasks. Chiril et al. [22] extended this to emotion and target identification, showing the benefits of auxiliary task learning for sexism detection. Building on this foundation, Yuan and Rizoiu [9] demonstrated that multi-task learning across multiple hate speech datasets substantially improves generalization to previously unseen datasets, achieving consistent improvements in cross-domain scenarios through their leave-one-out evaluation scheme.

2.2. Text-based Classification for Sexism Detection

Text remains the primary modality for sexism detection, given its prevalence in social media discourse. The unique challenges of social media text – including informal language, code-switching, and platform-

specific conventions – have shaped methodological developments in this area.

The evolution of representation learning has been particularly influential in capturing subtle sexist language. Early approaches relied on static word embeddings such as Word2Vec [23] and GloVe [24], which provided limited context sensitivity. The transition to contextual representations marked a significant advancement, with transformer-based encoders pre-trained on social media data demonstrating superior performance. Models like BERTweet [25] and TweetEval [26] excel at modeling platform-specific linguistic patterns, including hashtags, mentions, and abbreviated expressions common in online discourse.

Cross-lingual sexism detection has emerged as a research direction, particularly through the development of multilingual models. While mBERT [19] and XLM-RoBERTa [27] have enabled approaches across languages, their effectiveness varies. The EXIST shared tasks have consistently featured Spanish-English tracks, with top-performing systems leveraging language-agnostic representations. However, Nozza [28] revealed persistent performance disparities across languages, with multilingual models often underperforming on low-resource languages despite their theoretical universality.

The importance of domain-specific adaptation has been reported through multiple studies. Chiril et al. [29] demonstrated that models trained on general offensive language datasets exhibit performance degradation when applied to sexism-specific tasks, highlighting the unique linguistic characteristics of gender-based harassment. This finding motivated the creation of specialized resources, such as the expert-annotated datasets introduced by Guest et al. [30], designed to capture implicit forms of sexism that automated systems frequently miss. Yuan et al. [8] further advanced this area by proposing transfer learning techniques that leverage multiple independent datasets jointly, constructing unified hate speech representations that enable effective cross-dataset knowledge transfer while reducing annotation requirements.

2.3. Large Language Model Adaptation

LLMs have introduced new possibilities for automated sexism detection [31]. However, this application remains comparatively underexplored in NLP, compared to more extensively studied tasks such as sentiment analysis, summarization, or machine translation.

Initial investigations into prompt-based approaches revealed both promise and limitations. For example, Chiu et al. [32] showed that carefully engineered prompts enable models like GPT-3 [33] to achieve competitive zero-shot performance on hate speech detection tasks. However, subsequent work by Yin and Zubiaga [34] identified critical weaknesses: prompt-based methods struggle with implicit sexism and exhibit high sensitivity to prompt formulation, resulting in inconsistent predictions across semantically equivalent queries. However, recent advances in parameter-efficient fine-tuning methods present alternatives to full model adaptation. While techniques like LoRA [10] have succeeded in various domains, their application to hierarchical sexism detection remains underexplored. Our work addresses this gap by demonstrating that comprehensive LoRA adaptation with hierarchical routing can effectively model the multi-level nature of sexism categorization, achieving strong performance while maintaining computational efficiency.

2.4. Harmful Content Detection and Moderation

Understanding sexism detection requires broader context about harmful content dynamics and moderation effectiveness. Sexism represents a significant category within the wider ecosystem of harmful online content, and methodological advances in general harmful content detection can be leveraged for gender-based harassment identification. Recent research has revealed concerning patterns in how various forms of harmful content spread across social media platforms. Kong et al. [35] demonstrated that coordinated harmful content campaigns can be detected through the social system reactions they elicit, using interval-censored transformer approaches to identify coordinated behavior patterns with high accuracy. This work has important implications for sexism detection, as it shows how temporal patterns and user engagement can reveal coordinated campaigns of gender-based harassment, sug-

gesting that sexism detection systems can benefit from approaches originally developed for broader harmful content identification.

The importance of early detection in harmful content mitigation has been further emphasized by recent advances in engagement prediction models. Tian et al. [36] developed IC-Mamba, a state space model that excels at forecasting social media engagement within the crucial first 15-30 minutes of posting, enabling rapid assessment of content reach and early identification of potentially problematic content. Their approach to modeling interval-censored data with integrated temporal embeddings provides valuable insights for sexism detection systems, as early engagement patterns could signal the viral potential of sexist content, allowing for more timely intervention strategies.

Understanding the causal mechanisms underlying harmful content spread is equally critical for effective detection and intervention. Tian and Rizoiu [37] introduced a novel joint treatment-outcome framework that distinguishes correlation from causation in social media influence analysis, particularly for misinformation spread. Their approach adapts causal inference techniques to estimate Average Treatment Effects within the sequential nature of social media interactions, addressing challenges from external confounding signals. This work has important implications for sexism detection, as understanding the true causal influence of sexist content on user engagement can inform more targeted intervention strategies and help distinguish organic spread from coordinated amplification campaigns. In addition, Schneider and Rizoiu [5] showed that faster content moderation reduces harm from the most severe content, even on high-traffic platforms like Twitter. Using self-exciting point processes, the study highlights the urgent need for timely responses, an insight directly applicable to real-world sexism detection systems targeting gender-based harassment.

Research on opinion dynamics and intervention strategies offers additional perspectives relevant to sexism detection as part of broader harmful content mitigation. Calderon et al. [38] introduced the Opinion Market Model to evaluate positive interventions for stemming harmful opinion spread, demonstrating how media coverage can modulate the dissemination of problematic content. This framework provides valuable insights for understanding how sexist discourse spreads and how detection systems might be integrated with intervention strategies targeting various forms of harmful content.

Studies of extreme opinion infiltration have revealed the pathways through which harmful discourse enters mainstream conversations. Kong et al. [4] employed mixed-method approaches to show how extreme opinions gradually infiltrate online discussions, with their human-in-the-loop methodology providing insights into the dynamics of problematic speech evolution from conservative to extreme viewpoints. These findings are particularly relevant for sexism detection, as they highlight the importance of capturing subtle shifts in discourse that may not be immediately apparent through traditional classification approaches, and demonstrate how techniques developed for general harmful content can be adapted for gender-specific harassment detection.

The study of harmful discourse during crisis events provides additional context for understanding sexist content dynamics within broader patterns of problematic online behavior. Bailo et al. [3] analyzed the performance of far-right Twitter users during the Australian bushfires and COVID-19 pandemic, revealing how accounts promoting harmful content moved from peripheral to central positions in disaster-driven conversations. Their work demonstrates the importance of monitoring evolving discourse patterns, as the association between information disorder and overperformance of accounts spreading harmful content suggests systematic coordination that may include gender-based harassment campaigns.

Recent work on ideology detection has also informed approaches to sexism identification within the broader harmful content landscape. Ram et al. [39] presented an end-to-end ideology detection pipeline that constructs context-agnostic ideological signals from media slant data, demonstrating effective detection of extreme ideologies alongside psychosocial profiling. Their approach offers valuable methodological insights for sexism detection, particularly in terms of developing automatic signal generation that reduces dependence on manual annotation while maintaining detection accuracy across different types of harmful content.

3. Methodology

We present our methodology for EXIST 2025 Task 1, which uses parameter-efficient fine-tuning of Llama 3.1 8B [11] using LoRA with hierarchical label-aware routing to address all three subtasks across English and Spanish. Our approach leverages a unified multilingual model with task-specific adapters and conditional specialization based on the hierarchical label structure.

3.1. Task Formulation

We formulate the three EXIST 2025 Task 1 subtasks as follows:

Subtask 1.1 - Binary Sexism Identification: A binary classification problem where the model determines whether a given tweet contains sexist content (SEXIST vs. NOT_SEXIST).

Subtask 1.2 - Source Intention Detection: A multiclass classification task that categorizes the intention behind sexist tweets into three categories:

- DIRECT: Messages that are inherently sexist or incite sexist behavior
- REPORTED: Messages that report sexist situations experienced by women
- JUDGEMENTAL: Messages that judge or criticize sexist behavior

Subtask 1.3 - Sexism Categorization: A multilabel classification task that categorizes sexist content according to five types:

- IDEOLOGICAL_AND_INEQUALITY
- STEREOTYPING_AND_DOMINANCE
- OBJECTIFICATION
- SEXUAL_VIOLENCE
- MISOGYNY_AND_NON_SEXUAL_VIOLENCE

3.2. LoRA Configuration and Target Module Selection

We used LoRA for parameter-efficient finetuning, with attention to target module selection. While conventional approaches often restrict LoRA adaptation to attention weight matrices, our experiments showed that module targeting yields better performance.

Specifically, we applied LoRA decomposition to all linear transformation layers in the model architecture: the attention mechanism components (`q_proj`, `k_proj`, `v_proj`, `o_proj`), the feed-forward network layers (`gate_proj`, `up_proj`, `down_proj`), and the language modeling head (`lm_head`). For each target module, we introduced trainable low-rank matrices with rank $r = 16$, following the parameterization:

$$W = W_0 + BA,$$

where W_0 represents the frozen pretrained weights, and $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times d}$ are the trainable adaptation matrices initialized with $B \sim \mathcal{N}(0, \sigma^2)$ and A as zeros.

Table 1 summarizes our complete configuration for both LoRA hyperparameters and training settings. We selected rank $r = 16$ to balance adaptation capacity with memory efficiency, while maintaining computational feasibility through 4-bit quantization and gradient checkpointing. The use of Flash Attention v2 [40] further accelerates training without compromising model quality.

This configuration enables efficient fine-tuning while preserving the model's multilingual capabilities, which is crucial for our unified approach to handling English and Spanish data. The comprehensive targeting of all linear layers, rather than just attention matrices, provides the flexibility to capture task-specific patterns across the hierarchical sexism detection tasks.

Table 1
LoRA hyperparameters and training configuration

Configuration	Value/Setting
LoRA Hyperparameters	
Rank (r)	16
Alpha (α)	16
Target Modules	All linear layers (`all-linear`)
Dropout	0.1
Modules to Save	`lm_head`, `embed_tokens`
Training Configuration	
Quantization	4-bit QLoRA
Attention	Flash Attention v2
Learning Rate	2×10^{-4}
LR Schedule	Constant with 10% warmup
Batch Size	8 per device
Gradient Accumulation	2 steps
Gradient Checkpointing	Enabled

3.3. Hierarchical Label-Aware Adaptation with LoRA

To model the hierarchical structure of the subtasks, we implement a level-specific LoRA routing mechanism. The hierarchy includes three levels: *(i)* binary sexism detection, *(ii)* source intention classification, and *(iii)* sexism type categorization. For each level $\ell \in \{1, 2, 3\}$, we define a dedicated LoRA module $\Delta^{(\ell)}$ that adapts the shared language model f_θ.

During training, adapter routing is conditioned on the gold parent labels. At inference, predictions proceed sequentially from the top level, with the model using the predicted label $\hat{y}^{(\ell-1)}$ to activate the corresponding LoRA module $\Delta^{(\ell)}$ for the current level. This design supports conditional specialization while maintaining parameter efficiency.

The hidden representation at level ℓ is computed as:

$$\mathbf{h}^{(\ell)} = f_\theta(x) + \Delta^{(\ell)}_{\hat{y}^{(\ell-1)}}(x).$$

In addition to standard task-specific losses (e.g., cross-entropy for classification and binary cross-entropy for multi-label prediction), we introduce a soft constraint that penalizes invalid parent-child label transitions, thereby encouraging structured coherence across the hierarchy in:

$$\mathcal{L}_{\text{hierarchy}} = \lambda \sum_{i=1}^{N} \sum_{\ell=2}^{3} \mathbb{I}[\hat{y}_i^{(\ell-1)} = \text{NOT\_SEXIST}] \cdot \max_{c \in \mathcal{C}^{(\ell)}} p_i^{(\ell)}(c),$$

where λ is the hierarchy constraint weight, N is the number of instances, $\hat{y}_i^{(\ell-1)}$ is the predicted label at level $\ell - 1$ for instance i, $\mathcal{C}^{(\ell)}$ is the set of valid classes at level ℓ, $p_i^{(\ell)}(c)$ is the predicted probability for class c at level ℓ, and $\mathbb{I}[\cdot]$ is the indicator function. This constraint specifically penalizes cases where a non-sexist prediction at the binary level is followed by high-confidence predictions at subsequent hierarchical levels.

The total training objective combines task-specific losses with the hierarchical consistency constraint:

$$\mathcal{L}_{\text{total}} = \sum_{\ell=1}^{3} \mathcal{L}_{\text{task}}^{(\ell)} + \mathcal{L}_{\text{hierarchy}},$$

where $\mathcal{L}_{\text{task}}^{(\ell)}$ represents the standard loss at each level: cross-entropy for binary and multiclass classification tasks, and binary cross-entropy for the multilabel categorization task.

3.4. Data Processing and Multilingual Strategy

Our approach uses a straightforward supervised learning methodology with gold standard labels for training:

Label Processing: We use the provided gold standard labels for each subtask, treating each tweet-label pair as a standard supervised learning instance. While the EXIST 2025 dataset includes multiple annotations per instance under the Learning with Disagreement paradigm, our methodology focuses on the gold labels for efficient and direct optimization.

Input Formatting: Tweets are formatted using Llama 3.1's instruction template structure:

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
[Task-specific system prompt]<|eot_id|>
<|start_header_id|>user<|end_header_id|>
[Tweet text]<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
[Classification output]<|eot_id|>
```

Text Preprocessing: Minimal preprocessing is applied to preserve the authentic social media language characteristics. We maintain original tweet formatting, including hashtags, mentions, and emoji, as these elements often carry semantic significance for sexism detection.

Multilingual Strategy: We apply a unified multilingual approach, training a single model on both English and Spanish data simultaneously, leveraging Llama 3.1's native bilingual capabilities. Rather than training separate language-specific models, we hypothesize that joint bilingual training enhances cross-lingual transfer learning and improves overall performance by exposing the model to diverse linguistic expressions of sexism across both languages. This approach is inspired by recent work showing that transfer learning across hate speech datasets can achieve substantial improvements in generalization [9], and our multilingual strategy extends this principle to cross-lingual knowledge transfer.

Training Strategy: We fine-tune separate LoRA adapters for each subtask–binary classification for sexism identification (1.1), multiclass for source intention detection (1.2), and multilabel for sexism categorization (1.3)–optimizing each for its specific classification requirements. All adapters are trained using standard supervised learning with gold standard labels on the combined English-Spanish dataset, leveraging cross-lingual transfer to ensure robust bilingual performance without requiring separate language-specific models. Training continues until convergence while monitoring validation performance, with early stopping applied when validation loss plateaus to prevent overfitting.

4. Experimental Results

4.1. Cross-lingual Training Analysis

To validate our unified multilingual training strategy, we conducted ablation studies comparing joint bilingual training against separate language-specific models. Table 2 presents performance comparisons across all three subtasks.

Joint bilingual training consistently outperforms language-specific models across all subtasks, with F1 improvements ranging from $1.7 - 2.4$ percentage points. These gains validate our hypothesis that cross-lingual transfer enhances sexism detection by leveraging shared semantic patterns across languages, consistent with findings from transfer learning research in hate speech detection [8]. The improvements are particularly obvious for the multilabel categorization task ($+2.4$), suggesting that complex semantic distinctions benefit most from exposure to diverse linguistic expressions of sexism.

The bidirectional nature of cross-lingual transfer manifests itself in consistent improvements for both languages. Spanish, despite typically having fewer training instances in multilingual datasets, achieves comparable or slightly higher gains than English across all subtasks. This symmetric improvement pattern indicates effective knowledge sharing, where English contributes richer training signal while Spanish provides complementary linguistic patterns and cultural-specific expressions of sexism.

Table 2
Performance Comparison: Joint vs. Separate Language Training (F1 Scores on validation set)

Subtask	Training Strategy	English	Spanish	Average
1.1 (Binary)	Separate	0.847	0.831	0.839
	Joint Training	**0.863** (+0.016)	**0.851** (+0.020)	**0.857** (+0.018)
1.2 (Intention)	Separate	0.742	0.728	0.735
	Joint Training	**0.758** (+0.016)	**0.745** (+0.017)	**0.752** (+0.017)
1.3 (Categorization)	Separate	0.681	0.665	0.673
	Joint Training	**0.704** (+0.023)	**0.689** (+0.024)	**0.697** (+0.024)

These findings have important implications for multilingual sexism detection systems. Rather than maintaining separate models per language – which requires duplicated development effort and computational resources – our joint training approach achieves superior performance with a single unified model.

4.2. Task-Specific Performance Analysis

Table 3 presents our final results on the EXIST 2025 test set, evaluated using the Information Contrast Measure (ICM-Hard) [41], which provides a robust assessment of model performance under class imbalance and hierarchical label structures.

Table 3
Final Results on EXIST 2025 Test Set on Hard Label Evaluation on ICM-Hard

Subtask	English	Spanish	Overall	Overall Ranking
1.1 (Binary Sexism Detection)	0.6231	0.7124	0.6774	1
1.2 (Source Intention Detection)	0.3676	0.5937	0.4991	1
1.3 (Sexism Categorization)	0.5085	0.7650	0.6519	1

Our approach achieved first place across all three subtasks, showing the effectiveness of hierarchical LoRA adaptation with comprehensive module targeting. Spanish consistently outperforms English across all subtasks, with particularly pronounced differences in intention detection ($+0.23$ ICM-Hard) and sexism categorization ($+0.26$ ICM-Hard). This cross-lingual performance gap likely reflects both differences in training data distribution and linguistic characteristics of sexist expressions across languages, suggesting that sexist discourse may manifest through more identifiable patterns in Spanish social media text.

As expected, performance decreases with task complexity, from binary classification (0.6774) to the more nuanced intention detection task (0.4991). This progression aligns with the inherent difficulty of fine-grained semantic understanding required for distinguishing between direct sexism, reported experiences, and judgmental commentary. Interestingly, the multilabel categorization task achieves intermediate performance (0.6519), suggesting that our hierarchical approach effectively leverages parent-level predictions to guide more complex downstream classifications.

The strong performance on hierarchically dependent tasks validates our design choice of conditional LoRA routing. Despite the challenging nature of intention detection – which requires understanding pragmatic context and author stance – our model maintains competitive performance by conditioning adapter selection on binary sexism predictions. This shows that explicitly modeling label dependencies through hierarchical specialization provides tangible benefits for complex, structured classification scenarios in social media discourse analysis.

Table 4
Computational Efficiency Comparison

Metric	Full Fine-tuning	LoRA (Ours)
Trainable Parameters	8.03B (100%)	134M (1.67%)
GPU Memory (Training)	32GB	12GB
Training Time (per task)	24 hours	6 hours
Storage per Adapter	32GB	512MB

Table 5
LoRA Rank Ablation Study on Subtask 1.1 with performance evaluated on validation set.

Rank	Alpha	F1 Score	Parameters
8	8	0.851	67M
16	16	**0.868**	134M
32	32	0.869 (+0.1%)	268M
64	64	0.871 (+0.3%)	536M

4.3. Efficiency Analysis and Ablation Study

Our LoRA-based approach achieves substantial computational efficiency compared to full fine-tuning while maintaining competitive performance. Table 4 demonstrates that our method reduces trainable parameters by 98.33% (from 8.03B to 134M), enabling training on consumer-grade GPUs with only 12GB memory compared to the 32GB required for full fine-tuning. This efficiency translates to 4x faster training times and 64x smaller storage footprint per task-specific adapter, making the deployment practical and cost-effective.

To validate our hyperparameter selection, we conducted ablation studies examining the relationship between LoRA rank and model performance. Table 5 shows results on Subtask 1.1, revealing that rank 16 achieves optimal efficiency-performance trade-offs. While higher ranks (32, 64) yield marginal F1 improvements of less than 0.3%, they require 2-4x more parameters and proportionally increased memory and training time. Our chosen configuration thus maximizes accessibility for researchers with limited computational resources while achieving near-optimal performance across all subtasks.

These efficiency gains are particularly crucial for the hierarchical multi-task nature of the shared task, where separate adapters for each subtask would traditionally require 3x the storage and memory. Our approach enables deployment of all three task-specific models within the memory constraints of a single GPU.

5. Conclusion

This paper presents a simple yet highly effective approach to text-based sexism detection that achieved first-place performance in EXIST 2025 Task 1 across English and Spanish languages on hard label evaluations. Our methodology demonstrates that straightforward Low-Rank Adaptation (LoRA) fine-tuning of Llama 3.1 8B, combined with unified multilingual training, outperforms more complex approaches while maintaining computational efficiency.

Our key finding is that joint bilingual training consistently surpasses separate language-specific models, achieving 1.7-2.4% improvements across all subtasks. The bidirectional knowledge transfer between English and Spanish shows that shared semantic representations of sexist patterns can transcend language boundaries while preserving language-specific nuances. This finding has broader implications for multilingual classification tasks beyond sexism detection, extending previous work on cross-dataset transfer learning [9, 8] to the cross-lingual domain.

While our approach achieves strong performance through straightforward supervised learning, incorporating demographic information into the fine-tuning process presents an opportunity for im-

provement. The EXIST 2025 dataset includes rich annotator demographic information, including gender, age, education level, ethnicity, and country of residence. Rather than discarding this valuable information in favor of gold labels, future work could explore encoding these persona characteristics directly into the LoRA adaptation process. This could involve persona-specific adapters [42], demographic-aware attention mechanisms, or multi-task learning approaches that jointly optimize for sexism detection while modeling annotator perspectives. Such persona-aware fine-tuning could capture the subjective nature of sexism perception across different demographic groups, leading to more culturally-sensitive detection systems.

Furthermore, our hierarchical LoRA approach opens avenues for integration with broader harmful content moderation frameworks [5] and opinion market models [38] that could provide real-time intervention capabilities. Recent advances in early engagement prediction [36] suggest that combining our sexism detection approach with temporal engagement forecasting could enable proactive identification of potentially viral sexist content within the critical first minutes of posting. Additionally, incorporating causal modeling approaches [37] could help distinguish between organic engagement and coordinated amplification of sexist content, providing deeper insights into the true influence mechanisms underlying gender-based harassment campaigns. Future work could explore how our sexism detection system might be combined with positive intervention strategies to not only identify sexist content but also guide counter-narrative generation or targeted educational interventions, contributing to more comprehensive approaches for fostering equitable online discourse within the broader ecosystem of harmful content mitigation.

Declaration on Generative AI

This work was created, reviewed, and edited by human authors. AI tools were used in two specific capacities: (1) debugging the syntax errors and optimizations in the code components of the LoRA framework, and (2) writing assistance to improve conciseness and readability of manuscript sections.

For writing assistance, we used **Claude** (Anthropic) exclusively for grammatical error correction and sentence-level clarity improvements. Example prompts included: "Identify grammatical errors in this sentence" and "Keep concise, and improve reading flow. Match style. Highlight changes. Break down complex and long sentences and make more concise.". All AI-generated suggestions were critically reviewed, modified, and integrated by human authors. The original conceptual content, technical contributions, experimental design, analysis, and final editorial decisions remain entirely human-authored. AI tools did not contribute to the research methodology, data analysis, or scientific conclusions.

References

[1] Z. Talat, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[2] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: Proceedings of the international AAAI conference on web and social media, volume 11, 2017, pp. 512–515.

[3] F. Bailo, A. Johns, M.-A. Rizoiu, Riding information crises: the performance of far-right twitter users in australia during the 2019–2020 bushfires and the covid-19 pandemic, Information, Communication & Society (2023) 1–19. URL: https://www.tandfonline.com/doi/full/10.1080/1369118X.2023.2205479. doi:10.1080/1369118X.2023.2205479.

[4] Q. Kong, E. Booth, F. Bailo, A. Johns, M.-A. Rizoiu, Slipping to the extreme: A mixed method to explain how extreme opinions infiltrate online discussions, Proceedings of the International AAAI Conference on Web and Social Media 16 (2022) 524–535. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/19312. doi:10.1609/icwsm.v16i1.19312.

[5] P. J. Schneider, M.-A. Rizoiu, The effectiveness of moderating harmful online content, Proceedings of the National Academy of Sciences 120 (2023) 1–3. URL: https://www.pnas.org/doi/10.1073/pnas.2307360120https://pnas.org/doi/10.1073/pnas.2307360120. doi:10.1073/pnas.2307360120.

[6] L. Yuan, P. J. Schneider, M.-A. Rizoiu, Behavioral homophily in social media via inverse reinforcement learning: A reddit case study, Proceedings of the International Web Conference (WWW) (2025). URL: http://arxiv.org/abs/2502.02943http://dx.doi.org/10.1145/3696410.3714618. doi:10.1145/3696410.3714618.

[7] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: European Conference on Information Retrieval, Springer, 2025, pp. 442–449.

[8] L. Yuan, T. Wang, G. Ferraro, H. Suominen, M.-A. Rizoiu, Transfer learning for hate speech detection in social media, Journal of Computational Social Science 6 (2023) 1081–1101. URL: https://link.springer.com/10.1007/s42001-023-00224-9. doi:10.1007/s42001-023-00224-9.

[9] L. Yuan, M.-A. Rizoiu, Generalizing hate speech detection using multi-task learning: A case study of political public figures, Computer Speech & Language 89 (2025) 101690. URL: https://linkinghub.elsevier.com/retrieve/pii/S0885230824000731. doi:10.1016/j.csl.2024.101690.

[10] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[11] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[12] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576.

[13] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[14] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: European Conference on Information Retrieval, Springer, 2023, pp. 593–599.

[15] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2193–2210.

[16] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: Proceedings of the 25th international conference on world wide web, 2016, pp. 145–153.

[17] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[18] Z. Zhang, L. Luo, Hate speech detection: A solved problem? the challenging case of long tail on twitter, Semantic Web 10 (2019) 925–945.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[20] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information processing & management 57 (2020) 102360.

[21] M. Samory, I. Sen, J. Kohne, F. Flöck, C. Wagner, "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples, in: Proceedings of the international AAAI conference on web and social media, volume 15, 2021, pp. 573–584.

[22] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: a multi-target perspective, Cognitive Computation (2022) 1–31.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).
[24] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[25] D. Q. Nguyen, T. Vu, A.-T. Nguyen, Bertweet: A pre-trained language model for english tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[26] F. Barbieri, J. Camacho-Collados, L. E. Anke, L. Neves, Tweeteval: Unified benchmark and comparative evaluation for tweet classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1644–1650.
[27] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
[28] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 907–914.
[29] P. Chiril, V. Moriceau, F. Benamara, A. Mari, G. Origgi, M. Coulomb-Gully, He said "who's gonna take care of your children when you are at acl?": Reported sexist acts are not sexist, in: 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), ACL: Association for Computational Linguistics, 2020, pp. 4055–4066.
[30] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An expert annotated dataset for the detection of online misogyny, in: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume, 2021, pp. 1336–1350.
[31] T. K. Smith, H. R. Nie, J. R. Trippas, D. Spina, RMIT-IR at EXIST Lab at CLEF 2024, in: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.
[32] K.-L. Chiu, A. Collins, R. Alexander, Detecting hate speech with gpt-3, arXiv preprint arXiv:2103.12407 (2021).
[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[34] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Computer Science 7 (2021) e598.
[35] Q. Kong, P. Calderon, R. Ram, O. Boichak, M.-A. Rizoiu, Interval-censored transformer hawkes: Detecting information operations using the reaction of social systems, in: Proceedings of the ACM Web Conference 2023, ACM, 2023, pp. 1813–1821. URL: http://arxiv.org/abs/2211.14114https://dl.acm.org/doi/10.1145/3543507.3583481. doi:10.1145/3543507.3583481.
[36] L. Tian, E. Booth, F. Bailo, J. Droogan, M.-A. Rizoiu, Before it's too late: A state space model for the early prediction of misinformation and disinformation engagement, in: Proceedings of the International Web Conference (WWW), 2025. doi:10.1145/3696410.3714527.
[37] L. Tian, M.-A. Rizoiu, Estimating online influence needs causal modeling! counterfactual analysis of social media engagement (2025).
[38] P. Calderon, R. Ram, M.-A. Rizoiu, Opinion market model: Stemming far-right opinion spread using positive interventions, Proceedings of the International AAAI Conference on Web and Social Media 18 (2024) 177–190. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/31306. doi:10.1609/icwsm.v18i1.31306.
[39] R. Ram, E. Thomas, D. Kernot, M.-A. Rizoiu, Detecting extreme ideologies in shifting landscapes: an automatic & context-agnostic approach, in: International AAAI Conference on Web and Social Media (ICWSM), AAAI, 2025. URL: http://arxiv.org/abs/2208.04097.
[40] T. Dao, Flashattention-2: Faster attention with better parallelism and work partitioning, in: The

Twelfth International Conference on Learning Representations, 2024.
[41] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.
[42] A. F. Magnosão de Paula, J. S. Culpepper, A. Moffat, S. P. Cherumanal, F. Scholer, J. Trippas, The Effects of Demographic Instructions on LLM Personas, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25, 2025.