Comparative Analysis of Transformer-Based Models for Sexism Detection in Text

Ruiz Beatriz[1],†

[1] *National University of Distance Education (UNED), Malaga*

Abstract

This study presents a comprehensive evaluation of three transformer-based models—DistilBERT, XLM-RoBERTa, and DistilGPT-2—applied to the task of detecting and classifying sexist content across multiple subtasks within the EXIST 2025 challenge. The models were assessed on binary classification (sexist vs. non-sexist), intent classification, and multilabel categorization of sexist types. Results reveal distinct behavioral patterns and biases: while all models tend to overpredict sexist content and underutilize the non-sexist class in complex subtasks, DistilBERT demonstrated the most balanced performance in binary classification, XLM-RoBERTa showed robustness but a propensity for overgeneralization, and DistilGPT-2 exhibited greater flexibility in multilabel assignments despite its generative architecture. The findings underscore the challenges of fine-grained sexism detection, particularly the need for improved calibration and thresholding to enhance specificity without compromising sensitivity. Future work should focus on developing hybrid or hierarchical models, incorporating better data balancing strategies, and refining decision pipelines to more accurately discern subtle and varied sexist language. This work contributes valuable insights into the strengths and limitations of current NLP architectures in addressing socially sensitive content.

Keywords

Sexism detection, Natural Language Processing (NLP), Transformer models, Binary classification, Multilabel classification, EXIST 2025 challenge, Text classification, Social media analysis, Offensive language detection

1. Introduction

This work presents a technical approach for detecting and categorizing sexist content on social media, developed in the context of the EXIST 2025 challenge. The adopted strategy is based on fine-tuning pretrained multilingual Transformer models available through the Hugging Face Transformers library. Three subtasks are addressed: (1) binary classification between sexist and non-sexist content, (2) identification of sexist intent, and (3) categorization of the type of sexist discourse through multi-label classification. The system was trained for a single epoch to maximize computational efficiency and avoid overfitting. Three main models were evaluated —DistilBERT, XLM-RoBERTa, and DistilGPT-2— and their comparative performance was analyzed. The results show significant differences in sensitivity, balance, class bias, and overlabeling, highlighting the inherent challenges of automatically processing sexist discourse in multilingual environments.

2. Main Objectives

The main goal of this project is to explore the capabilities of modern NLP systems in detecting and classifying sexist content on social media, a task that presents both social relevance and technical complexity. The problem is challenging not only because of the implicit and nuanced ways sexism can manifest, but also due to the multilingual nature of social media platforms, where language use is highly varied and informal. The EXIST 2025 challenge provides a structured benchmark for this task, organizing

[1] *CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, SpainCorresponding author.*
Email: bearv4165@gmail.com (B. Ruiz)

it into three subtasks that reflect increasing levels of granularity: simple binary classification, identification of intent behind sexist messages, and fine-grained categorization by discourse type.

This project aims to assess how well current pretrained Transformer-based language models can be adapted to these subtasks, using the EXIST dataset as a real-world testbed. The intention is not only to measure raw classification performance but also to understand the comparative strengths and limitations of different modeling strategies—particularly the contrast between discriminative and generative paradigms. The experimental design is centered on comparing how these approaches respond to the demands of the different subtasks, especially in contexts where input signals may be ambiguous, subtle, or underrepresented.

Beyond performance evaluation, another key objective is to implement a robust and extensible pipeline that handles data preprocessing, model training, prediction formatting, and evaluation in a coherent and replicable manner. This includes adapting label representations to model-friendly formats, designing inference strategies appropriate to the nature of each task (e.g., single-label vs. multi-label), and aligning outputs with the structure expected by the competition. Ultimately, the project seeks to provide insight into how pretrained language models, when minimally adapted, can contribute to the automatic detection of online sexism. The comparison across modeling paradigms also aims to open a discussion on whether generative approaches, despite their current limitations, may offer complementary advantages in scenarios involving nuanced social language or low supervision.

3. Approach(es) used and progress beyond state-of-the-art

Recent advancements in the automatic detection of sexism in text have demonstrated clear progress beyond traditional fine-tuning of transformer-based models. While early solutions—such as those using DistilBERT, XLM-RoBERTa, or DistilGPT-2—offered strong baselines through straightforward supervised training, the field has now evolved toward more sophisticated, layered approaches that integrate architectural innovation with semantic and data-centric strategies. These newer methods not only surpass baseline performance but also address some of the key limitations revealed in previous evaluations, particularly regarding label imbalance, overgeneralization, and contextual misinterpretation.

One of the most significant developments is the integration of definition-based and context-aware data augmentation techniques. For example, recent systems have employed Definition-based Data Augmentation (DDA), which synthesizes examples that are semantically aligned with specific label definitions. This strategy introduces synthetically generated but linguistically diverse samples into the training pipeline, helping the model internalize what each label truly represents. Alongside this, Contextual Semantic Expansion (CSE) has been used to extend the model's input with supporting semantic material drawn from large knowledge bases or contextual prompts. These augmentations have been shown to improve both macro and micro F1-scores, particularly in binary and multi-label settings. Crucially, they help models better distinguish between truly sexist content and benign or ambiguous expressions—a recurring weakness in simpler models like DistilGPT-2 and DistilBERT when used with default hyperparameters.

In parallel, ensemble learning has emerged as a dominant paradigm for pushing performance beyond that of any single architecture. Top systems in recent shared tasks—such as EXIST 2023 and 2025—have leveraged combinations of transformer models with differing pretraining objectives, such as BERT, RoBERTa, XLM-RoBERTa, and even newer large language models like Mistral-7B. These ensembles often rely on voting or confidence-weighted averaging mechanisms to aggregate predictions, thereby mitigating the overconfident misclassifications that individual models frequently produce. For instance, while a model like DistilGPT-2 might demonstrate sensitivity to subtle evaluative language, it may also exhibit noise in label assignment; pairing it with a more conservative but stable model like XLM-RoBERTa can strike a more effective balance. In some architectures, fallback systems are also employed—wherein one model's output is passed to another when confidence is low—introducing a form of decision-level redundancy that increases robustness.

Model architecture itself has seen notable evolution. The move from purely discriminative classifiers to hybrid or generative-informed systems reflects a growing appreciation for the complexity of sexist language. Generative models, despite being traditionally designed for text production rather than classification, have shown potential in capturing implicit cues and indirect speech acts, which are often missed by pattern-dependent discriminative models. Moreover, the incorporation of prompt engineering and zero-shot learning techniques—especially in larger LLMs—has allowed for more flexible and dynamic label interpretation. For instance, rather than rigidly classifying a tweet into predefined labels, a prompt-informed system might assess whether the tone is judgmental or ideological based on semantic proximity to a label description. This strategy has also paved the way for prompt-tuned multi-task learning, wherein each classification task is reframed with label definitions embedded into the model input, encouraging the model to reason more carefully about each category.

Reinforcement learning from human feedback (RLHF) represents another breakthrough, especially in models like Mistral-7B and LLaMA-3, which have been fine-tuned to better reflect user expectations and contextual nuance. This process, involving iterative alignment of model outputs with human evaluators' judgments, enhances the interpretability and precision of classification decisions. In the context of sexism detection, RLHF-trained models have shown improved performance in detecting not only overt misogyny but also subtler forms of harm, such as stereotyping and dismissive language. More importantly, these models are capable of expressing uncertainty, an essential feature when working with socially sensitive content where misclassification can have real-world consequences.

Alongside architectural innovations, the task formulations themselves are being reconsidered to reflect the layered nature of sexism in discourse. Rather than treating label assignment as a flat classification problem, newer approaches embrace hierarchical task structures. These begin with a coarse-grained determination of whether sexist intent is present, followed by finer classification into specific subtypes such as objectification or ideological inequality. Such pipelines allow for error containment: a tweet misclassified as sexist in the first stage does not automatically receive a misleading subtype label. Furthermore, these multi-stage models are better equipped to handle overlapping categories and ambiguous cases, which were especially problematic in earlier multi-label frameworks where models like XLM-RoBERTa tended to over-predict and assign nearly all labels to each example.

Another key trend in recent work is the attention to cross-lingual robustness and cultural contextualization. While early models were primarily trained and evaluated in English or other high-resource languages, current state-of-the-art systems apply domain adaptation strategies to function effectively across linguistic boundaries. Domain Adaptive Pretraining (DAP), for example, involves retraining models on domain-specific corpora—such as social media posts from particular regions or cultural contexts—before fine-tuning on the task-specific dataset. This process allows models to better recognize localized patterns of sexist expression. Similarly, cross-lingual embeddings and multilingual tokenizers enable better generalization across languages, as seen in the adaptation of XLM-RoBERTa or multilingual LLaMA variants to sexism detection in Spanish, French, and Arabic.

Finally, a more subtle but equally important direction in the progression beyond the state of the art is the growing awareness of the need for interpretability and ethical responsiveness. As sexism detection systems are increasingly deployed in moderation tools and public-facing applications, it is no longer sufficient to maximize classification performance alone. Recent research stresses the importance of explanation mechanisms—such as saliency maps, attention visualization, and natural language rationales—to allow users and moderators to understand why a piece of content was flagged. Additionally, socially aware modeling practices—like fairness-aware training, bias auditing, and inclusion of community-derived label definitions—are now considered best practices in the field. These ensure that systems are not only technically robust but also aligned with the diverse expectations and values of the communities they aim to serve.

In conclusion, while the use of pretrained transformers like DistilBERT and XLM-RoBERTa marked a significant advancement in the early stages of sexism detection, the field has rapidly evolved beyond such baselines. The current state of the art combines data augmentation, ensemble modeling, prompt-aware architectures, reinforcement learning, and hierarchical task design into unified systems that far exceed the capabilities of any single fine-tuned model. This holistic progress underscores a growing recognition that sexism detection is a complex linguistic and social task—one that demands not just computational power, but also careful engineering, interpretive nuance, and ethical foresight.

4. Resources employed

Three pretrained language models were employed as the backbone for classification: DistilBERT-base-multilingual-cased, XLM-RoBERTa-base, and DistilGPT-2. These models were chosen primarily for their multilingual support and their direct availability via the Hugging Face Transformers library. The first two models, DistilBERT and XLM-RoBERTa, served as discriminative classifiers, while DistilGPT-2 was used to evaluate a generative approach to textual understanding and label generation.

Each discriminative model was fine-tuned separately for the three subtasks. For Subtask 1, a binary classifier was implemented by adding a softmax layer over two output units (sexist vs. non-sexist). Subtask 2 employed a similar structure but with a softmax layer over four classes to capture different types of sexist intent. For Subtask 3, a multilabel classification head was added with a sigmoid activation, enabling the model to independently predict the presence of multiple categories per tweet. A fixed threshold of 0.5 was used to determine label presence in Subtask 3. In cases where no class exceeded this threshold, the class with the highest probability was selected as a fallback to ensure at least one label prediction per instance.

The preprocessing pipeline included several steps: parsing the JSON files into pandas DataFrames, extracting relevant fields, handling edge cases such as missing or ambiguous labels, and converting categorical annotations into numerical indices. Each tweet was tokenized using the respective tokenizer associated with its model—either DistilBERT, XLM-RoBERTa, or GPT-2—and padded dynamically using Hugging Face's 'DataCollatorWithPadding', which ensures uniform batch shapes without truncating valuable context.

Training and inference were conducted using Hugging Face's 'Trainer' API, which streamlined the fine-tuning process while providing support for GPU acceleration, mixed-precision training (FP16), and reproducible logging. For each model–task combination (totaling nine distinct experiments), we adopted a lightweight training setup: a single epoch of training ('num_train_epochs=1') with a batch size of 8 samples per device, using default learning rate, optimizer, and scheduler settings. No additional optimization strategies—such as learning rate tuning, dropout adjustment, early stopping, or data augmentation—were applied, in order to limit computational cost and assess out-of-the-box model effectiveness.

Evaluation was conducted using a reserved validation set from the EXIST 2024 dataset, and focused on four standard metrics: accuracy, precision, recall, and macro-averaged F1-score. Although the official test set was unlabeled and used strictly for inference, the validation set allowed for internal performance monitoring. Both hard predictions (final label choices) and soft predictions (class probability distributions) were generated and saved in a JSON format compatible with the competition's requirements, allowing for future benchmarking or submission.

The generative model, DistilGPT-2, was employed differently. Instead of learning classification boundaries directly, it was prompted to generate textual outputs from which labels were inferred using rule-based extraction and approximate string matching. Custom prompts were crafted for each subtask, emulating natural instructions for classification or annotation. Generated responses were parsed to extract label mentions, which were then mapped to known class labels. Although this method is more fragile than classification, it provides insights into model interpretability and performance in zero-shot or low-resource scenarios. Like its discriminative counterparts, DistilGPT-2's outputs were evaluated against ground truth annotations on the validation set.

5. Results obtained

Based on the predictions and validation metrics obtained for Subtask 1.1 (binary classification of sexist vs. non-sexist content) using the DistilBERT model, we can provide a more comprehensive analysis of its performance and behavior. The hard prediction outputs on the unlabeled test set show that DistilBERT predicted 1,274 tweets as "NO" (non-sexist) and 802 tweets as "YES" (sexist). This indicates a tendency toward the negative class, but not to an extreme degree. Roughly 38.6% of predictions are classified as sexist, which reflects a level of sensitivity to discriminatory content while still maintaining caution—a desirable balance for a socially sensitive task where false positives can be problematic, but false negatives (overlooking harmful content) are equally critical.In terms of soft predictions, the accompanying file, 'distilbert_soft.json', contains probability distributions for each class ("YES" and "NO"). An inspection of these values reveals that the model often makes high-confidence predictions. For instance, when predicting "YES", the associated probabilities are frequently above 0.8 or even 0.9, indicating that the model is not relying on uncertain margins for classification but rather differentiates the two classes with relatively strong internal confidence.

Importantly, the validation metrics on labeled development data provide insight into the generalization capacity of the model. The reported F1-score is 0.7604, with precision at 0.7656 and recall at 0.7576. These values demonstrate that the classifier is well-balanced in identifying both positive and negative cases. The slightly higher precision suggests the model is conservative when flagging tweets as sexist, which aligns with its prediction behavior on the test set, where a greater proportion of tweets were labeled as non-sexist. This could help mitigate false accusations of harmful intent, which is ethically important in sensitive applications.

Overall, DistilBERT's performance in Subtask 1.1 can be considered strong and reliable, especially considering that the model was fine-tuned for only a single epoch with default hyperparameters and without advanced optimization techniques like early stopping or learning rate scheduling. The results show that even with minimal training, pretrained transformer-based models can effectively capture the underlying patterns associated with sexist language, particularly when the task is framed as binary classification. The model exhibits both reasonable discrimination and calibrated output probabilities, making it a suitable baseline or production-ready component for real-world systems focused on harmful content detection.

In Subtask 2.2, whose goal is to identify the intent behind sexist content through a four-class classification (NO, DIRECT, REPORTED, and JUDGEMENTAL), the DistilBERT model showed a strong preference for the DIRECT class, assigning it to 1,411 examples. The REPORTED and JUDGEMENTAL classes were much less frequent in the predictions, with 399 and 266 cases respectively. Notably, the NO class did not appear in any predictions, suggesting that the model did not identify any instances as lacking sexist intent.This behavior reflects a certain rigidity or bias in the model's output, which can be explained by several factors. It is possible that the model learned to associate most problematic expressions in the training set with direct intent, which makes sense if the corpus contains a substantial number of explicit examples. However, the complete omission of the "NO" class presents a serious issue of overprediction of intentional sexism, as it suggests a lack of sensitivity in detecting neutral or non-offensive messages within the context of this subtask.The skewed distribution may indicate that the model has captured certain aggressive or explicit linguistic patterns and generalized them broadly. This could result in a high number of false positives in the "DIRECT" or "REPORTED" classes if the original test content includes ambiguous, ironic, or merely critical tweets without sexist connotation. Moreover, the underrepresentation of the "JUDGEMENTAL" class may be due to the model's difficulty in identifying the indirect evaluative tone that characterizes this category.

In Subtask 3.3, focused on multi-label classification of the types of sexism present in texts, the DistilBERT model displayed a clear tendency toward over-assigning certain categories. Each tweet can receive multiple labels, and the results indicate that IDEOLOGICAL-INEQUALITY and STEREOTYPING-DOMINANCE were assigned to all examples (2,076 times each), revealing a lack of discrimination between classes and a structural bias in the model's output. These categories, while relevant, should not cover the entire dataset if the model were capturing nuances more accurately.Other classes, such as MISOGYNY-NON-SEXUAL-VIOLENCE (1,644 assignments), OBJECTIFICATION (1,472), and SEXUAL-VIOLENCE (1,004), were also detected at significant but lower frequencies. Notably, the NO class—which would indicate an absence of sexism in the tweet—does not appear in any prediction, which aligns with the patterns observed in Subtask 2: the model systematically avoids or ignores this label. This suggests a tendency of the system to assume the presence of sexism in every test case, likely driven by a training process lacking a sufficient balance of well-represented neutral or negative examples.The observed behavior indicates that DistilBERT has overgeneralized certain categories, particularly those encompassing ideological or dominance-related discourse—possibly because such language tends to be more explicit or frequent in the training set. However, this overgeneration reduces the model's utility in precise discrimination, as it loses the ability to identify only the relevant types of sexism for each case. In practical terms, this can lead to label inflation per tweet, reducing the specificity of the predictions.

In Subtask 1.1, the XLM-RoBERTa model showed a prediction distribution similar to DistilBERT's, although with an even stronger bias toward the "NO" class. Of the 2,076 processed examples, 1,274 were classified as non-sexist and only 802 as sexist. This trend reflects a greater inclination of the model to view messages as non-problematic, suggesting a conservative decision policy in detecting sexist content.The imbalance in predictions may be due to multiple factors. The model might be reacting cautiously to ambiguous linguistic patterns, classifying as "NO" those cases without explicit evidence of sexism, or it may have learned to overvalue neutral linguistic features as indicators of non-sexism. This behavior could help reduce false positives but also poses a significant risk of overlooking subtle manifestations of sexism, especially if they do not follow prototypical linguistic patterns.Despite this bias, the number of cases detected as sexist (802) is still significant, indicating that the model is capable of recognizing problematic expressions when they are clear. However, without reference labels in the test set, it is not possible to determine definitively whether this distribution reflects high precision or a loss of sensitivity.

In Subtask 2.2, focused on identifying the intent behind sexist content, the XLM-RoBERTa model showed a strong tendency to classify examples as DIRECT, which received 1,379 of the 2,076 predictions. The second most frequent category was REPORTED, with 695 cases, while JUDGEMENTAL was used sparingly, with only 2 examples. Once again, the NO category—corresponding to the absence of sexist intent—was not assigned to any case, indicating that the model assumed all texts presented some form of sexist intent.This pattern suggests that XLM-RoBERTa tends to view sexism in binary and explicit terms, prioritizing the most direct and easily recognizable forms of offensive expression. The high volume of predictions in the DIRECT class suggests that the model has strongly internalized certain explicit textual cues during training, but at the cost of a more nuanced understanding of indirect forms of sexism, such as implicit judgments or subtle value statements. The near absence of the JUDGEMENTAL category and the complete omission of NO indicate that the model lacks adequate sensitivity to identify neutral content or messages with subtle intent.Compared to DistilBERT, XLM-RoBERTa appears somewhat more flexible in assigning the REPORTED class, which could suggest a better ability to recognize when a message refers to third-party sexism rather than expressing it directly. However, the distribution remains highly imbalanced, limiting the model's value as an analytical tool in contexts where fine-grained detection of intent is crucial.

In Subtask 3.3, which requires classifying the specific types of sexism present in texts through a multi-label strategy, the XLM-RoBERTa model showed an extreme tendency to tag each entry with multiple categories, reaching nearly full coverage in five out of six available classes. The labels STEREOTYPING-DOMINANCE and MISOGYNY-NON-SEXUAL-VIOLENCE were assigned in all cases (2,076 times), while OBJECTIFICATION appeared in almost all (2,068 instances), and IDEOLOGICAL-INEQUALITY and SEXUAL-VIOLENCE were also highly frequent (2,027 and 1,859 times, respectively). As with previous models, NO, the class representing the absence of sexism, was not used at all.This pattern suggests that XLM-RoBERTa is not effectively discriminating between the available categories, but instead tends to tag nearly all texts as containing multiple types of sexism simultaneously. Although overlap between categories is plausible in real-world contexts, such broad and uniform coverage points to a failure in model calibration or threshold tuning for deciding which labels to assign. Rather than distinguishing between different patterns of sexism, the model seems to be maximizing sensitivity at the expense of specificity, leading to an inflation of labels per example.Such behavior may be driven by a bias in the training set, where label co-occurrences were common, or by prediction thresholds that were set too low, causing the model to assign labels even with minimal probabilities. The consequence is a loss of utility in analytical or intervention tasks, since the predictions do not clearly indicate which type of sexism predominates in each message.Ultimately, while XLM-RoBERTa shows high capability in detecting signs of sexism in general, its performance in Task 3 reveals a lack of selectivity that compromises its semantic precision. To improve, it would be essential to apply class-wise calibration techniques, dynamic thresholds, or even hierarchical strategies that first assess the presence of sexism and then refine classification among the categories.

As for DistilGPT-2, in Subtask 1.1, it produced results reflecting a strong inclination toward the "NO" class, with 1,314 predictions versus only 762 classified as "YES". This marks the most skewed distribution among the three model variants tested for this task and suggests that DistilGPT-2 is the most conservative when labeling text as sexist.This behavior may be linked to the nature and training of the model itself. Unlike DistilBERT and XLM-RoBERTa—explicitly pretrained for text understanding—DistilGPT-2 derives from an autoregressive generative model, originally designed for text generation, not classification. Although adapted for supervised tasks through fine-tuning, its architecture may struggle more to capture the subtle discriminative signals between binary classes without more intensive or task-specific training.The imbalance in predictions suggests that DistilGPT-2 was less sensitive to indicators of sexism, which could result in a higher false negative rate if a significant proportion of problematic messages exists in the test set. This makes it a less suitable option in contexts where detecting offensive content is a priority. However, its high rate of "NO" classifications may also imply a low false positive rate, which could be beneficial in applications seeking to avoid over-censorship or mislabeling.

In Subtask 2.2, focused on classifying the intent of sexist content, DistilGPT-2 showed a prediction distribution similar to the other models, but with slightly more diversity in its results. The DIRECT class was by far the most assigned, with 1,368 cases, followed by REPORTED (431 cases) and JUDGEMENTAL (277 cases). As with the other models, the NO class—indicating absence of sexist intent—was not assigned in any prediction.This behavior suggests that DistilGPT-2 has a strong preference for labeling sexist intent as direct, possibly because explicit linguistic patterns are more easily recognized by a generative language model like this. However, compared to XLM-RoBERTa and DistilBERT, there is a greater presence of the JUDGEMENTAL class, which may indicate that DistilGPT-2 is better at detecting subtle or evaluative nuances in texts—albeit still on a limited scale.The slightly more balanced representation of the REPORTED and JUDGEMENTAL classes could stem from the fact that, being pretrained as a generative model, DistilGPT-2 tends to capture broader and more diffuse semantic relationships and may be somewhat more "creative" in label assignment when adapted to classification tasks. Nevertheless, a strong bias toward the presence of sexism remains, as no instance was labeled as having no intent, reinforcing the hypothesis of systematic overprediction.

In Subtask 3.3, designed to identify multiple types of sexism present in texts, DistilGPT-2 generated predictions that, while still following a label over-assignment pattern like the other models, showed greater variability in class assignment. The labels IDEOLOGICAL-INEQUALITY and STEREOTYPING-DOMINANCE dominated predictions, with 2,073 and 2,014 assignments, respectively, but other categories like OBJECTIFICATION (1,655), MISOGYNY-NON-SEXUAL-VIOLENCE (603), and SEXUAL-VIOLENCE (431) were used less frequently. This suggests a more differentiated distribution of classes compared to the results from XLM-RoBERTa or DistilBERT.This greater variability may be related to the generative architecture of DistilGPT-2, which—though not optimized for classification—tends to capture broader and less rigid semantic relationships. As a result, the model seems less prone to tagging all examples with the full set of available classes, though it still shows a general tendency to assume the presence of multiple forms of sexism in nearly every case. Like the other models, the NO class was not assigned to any text, reinforcing the idea of a systematic overprediction of the phenomenon.The lower frequency of categories like MISOGYNY-NON-SEXUAL-VIOLENCE and SEXUAL-VIOLENCE may also indicate that the model struggles more to identify sexist discourse that doesn't involve explicit or physical violence, or that is expressed indirectly. This behavior may reflect biases in the training data or the need for more tailored thresholding for each class in multi-label tasks.

6. Analysis of the results

The comparison between the three models used across the three subtasks of the EXIST 2025 challenge reveals notable differences in behavior, sensitivity, and prediction balance. Although all models share certain general trends—such as the absence of predictions for the "NO" class in subtasks 2 and 3—there are nuances that distinguish their generalization ability and approach to sexist content.

In Subtask 1.1 (binary classification between sexist and non-sexist), all three models showed a bias toward the "NO" class, albeit with varying intensity. DistilGPT-2 was the most conservative model, with only 762 predictions of sexism versus 1,314 non-sexist predictions. XLM-RoBERTa occupied a middle ground (802 "YES" vs. 1,274 "NO"), while DistilBERT showed the most balanced distribution (864 "YES" and 1,212 "NO"). This difference suggests that DistilBERT may be more sensitive to detecting explicit sexism, while DistilGPT-2 tends to avoid labeling messages as sexist unless there are very clear signals.

In Subtask 2.2 (classification of the intent behind sexist content), all three models overemphasized the DIRECT category, though there were variations in how they treated less frequent classes. XLM-RoBERTa was the most extreme, almost completely ignoring the JUDGEMENTAL category (only 2 cases) and not using the NO class at all. DistilGPT-2, though also biased toward DIRECT (1,368 times), showed relatively higher sensitivity to REPORTED (431) and JUDGEMENTAL (277) categories, suggesting a better capacity to capture more nuanced intentions. DistilBERT, for its part, fell in between: it used DIRECT as the main class, but assigned a moderate number of examples to the other categories, though not achieving a true balance.

In Subtask 3.3 (multi-label classification of sexist categories), all models displayed a clear trend toward over-labeling, but differed in their label distribution. XLM-RoBERTa was the most extreme, assigning virtually all labels to all examples and completely omitting the "NO" class. DistilBERT also showed a strong tendency toward mass labeling, though with slightly more variability. DistilGPT-2 produced more differentiated predictions: while it still favored dominant labels (IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE), it used categories like SEXUAL-VIOLENCE and MISOGYNY-NON-SEXUAL-VIOLENCE less frequently, suggesting a greater ability to discern among categories in the multi-label setting, despite not being originally optimized for such tasks.

In summary, DistilBERT stands out for its balanced performance in binary classification and reasonable sensitivity in supervised tasks. XLM-RoBERTa excels in robustness but tends to overgeneralize, especially in multi-label tasks. DistilGPT-2, while less conventional for classification, provides a more nuanced response in the intent and category tasks, showing less rigid behavior and possibly greater adaptability with proper tuning. The optimal model choice would therefore depend on the specific objective: conservative precision, broad detection, or finer, more differentiated analysis of sexist discourse.

The results show that no single model dominates across all subtasks, and each contributes distinct advantages depending on the classification type. Architecture significantly influences predictive behavior: while discriminative models tend toward more decisive outputs, generative models offer flexibility that could be better leveraged through calibration. The widespread absence of the "NO" class in the more complex subtasks highlights an urgent need to improve balance and sensitivity toward non-sexist content.

7. Perspective for future works

The analysis of the three models' performance across the EXIST 2025 subtasks reveals several opportunities and imperatives for future work in the automated detection and classification of sexist content. While transformer-based architectures such as DistilBERT, XLM-RoBERTa, and DistilGPT-2 demonstrate considerable power in handling textual classification tasks, their outputs also reflect systemic limitations that must be addressed to improve both accuracy and ethical robustness. One of the most consistent issues across subtasks—particularly evident in intent classification (Subtask 2.2) and multi-label categorization (Subtask 3.3)—was the near-total absence of the "NO" class in the models' predictions. This lack of sensitivity to non-sexist or neutral content highlights a major shortcoming in the current model pipelines: the failure to account for the *absence* of harmful content as a meaningful and distinct outcome.

Improving this dimension of classifier sensitivity should be a primary objective in future iterations. A system that cannot reliably distinguish between offensive and inoffensive content is not only less useful but also risks reinforcing harmful or biased moderation practices in real-world deployments. To mitigate this, future work should incorporate better-balanced datasets that explicitly include a wider variety of neutral, ambiguous, and non-problematic content. Techniques such as counterexample mining, adversarial data augmentation, and semantic contrastive learning could enhance the model's discrimination between subtle linguistic signals. These methods would help ensure that a model can correctly withhold a label when none is appropriate, a capacity just as important as identifying problematic discourse.

A closely related issue is the problem of *overlabeling*, especially prevalent in the multi-label classification task. In Subtask 3.3, the models, particularly XLM-RoBERTa, demonstrated a tendency to assign almost all available categories to every instance, regardless of the actual semantic nuance present. This overgeneration of labels dilutes the specificity of the predictions and undermines the practical value of the system for tasks that require focused intervention—such as identifying whether a tweet involves objectification versus ideological inequality. The overlabeling phenomenon suggests that the default thresholding and scoring mechanisms used during prediction were insufficiently tuned. Rather than applying static thresholds across all categories, future systems should explore *class-specific calibration* strategies or *dynamic thresholding*, where decisions are adjusted based on the confidence distribution or the expected co-occurrence patterns among classes. A two-stage or *hierarchical* pipeline, in which the presence of sexism is first confirmed before secondary classification into subtypes, could also reduce noise and better reflect the natural hierarchy in sexist discourse.

Another important perspective for future work involves model architecture. The comparative behavior of DistilBERT and XLM-RoBERTa—both discriminative models—with DistilGPT-2, a generative model, demonstrates that different modeling strategies offer distinct strengths. The discriminative models tended to provide consistent, bounded predictions, favoring stability and precision, but sometimes at the cost of nuance. In contrast, the generative model, while more erratic in some tasks, exhibited greater variability and a tendency to detect more subtle, judgmental language. This suggests a compelling opportunity for the development of *ensemble or hybrid models* that combine the structured decision-making of discriminative classifiers with the contextual fluidity of generative ones. For instance, an ensemble could use a discriminative model to produce high-confidence core predictions, while a generative model refines these outputs in edge cases or adds interpretive layers around intent. Such systems could help balance caution with contextual awareness, offering a more robust solution for real-world content moderation tasks.

From a technical development standpoint, there is also a clear need to move beyond minimal training regimens. Many of the tested models were fine-tuned using default parameters, without advanced optimization strategies such as early stopping, learning rate scheduling, or adaptive loss functions. While this was useful for establishing baselines, more rigorous training—especially with techniques like curriculum learning or targeted fine-tuning on challenging subsets—could significantly enhance the model's ability to generalize to difficult or ambiguous examples. Furthermore, integrating *uncertainty* *estimation* during inference (e.g., using Monte Carlo dropout or ensemble variance) would help differentiate between high-confidence and low-confidence decisions, enabling safer deployment in sensitive contexts.

Beyond model-specific interventions, the overall design of the classification framework also warrants reconsideration. The binary, intent, and multi-label structures used in EXIST 2025 offer useful scaffolding, but more sophisticated task formulations could improve clarity and effectiveness. One promising direction is the use of *hierarchical taxonomies* of sexism, which would acknowledge that some forms of harmful content are nested within broader communicative patterns or intentions. A tweet might be judgmental in tone, ideological in substance, and simultaneously objectifying in expression—each layer providing a different insight into its harmfulness. Rather than forcing a flat multi-label output, a more structured hierarchy could allow systems to progressively refine predictions, mirroring human reasoning more closely.

Moreover, integrating *pragmatic and contextual metadata*—such as author identity, reply chains, or engagement metrics—could greatly enrich the model's understanding of intent and tone. Current models operate largely in isolation, treating each tweet as a standalone unit, which limits their ability to assess sarcasm, irony, or indirect speech acts. In real-world moderation scenarios, context is often critical to distinguish between problematic and benign content. Developing models capable of leveraging this surrounding context could dramatically improve performance, especially on the subtasks that require interpretive depth.

Ultimately, the evaluation of the EXIST 2025 results reinforces that addressing linguistic phenomena like sexism is not purely a technical challenge but also a question of system design and value alignment. While large language models offer powerful foundations, their effectiveness depends on thoughtful tuning, robust evaluation, and sensitivity to the social stakes of their predictions. Future work in this space should therefore pursue a multi-dimensional strategy: refining architectures, improving training regimes, rebalancing data, and rethinking task structures. In doing so, we move closer to building models that are not only accurate but also responsible, fair, and aligned with the complex realities of language and social meaning.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019). URL: https://arxiv.org/abs/1910.01108.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, et al., Unsupervised cross-lingual representation learning at scale, in: Proceedings of ACL, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Technical Report (2019). URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[4] R. Rodríguez-Sánchez, F. Rangel, P. Rosso, M. Wiegand, Overview of EXIST 2021: Sexism identification in social networks, in: Proceedings of IberLEF 2021, CEUR Workshop Proceedings, 2021. URL: http://ceur-ws.org/Vol-2943/exist_overview.pdf.

[5] F. Rangel, R. Rodríguez-Sánchez, P. Rosso, EXIST 2023 at IberLEF: Sexism Identification in Social Networks, in: CEUR Workshop Proceedings, Vol. 3461, 2023. URL: http://ceur-ws.org/Vol-3461/paper-exist-overview.pdf.

[6] S. Sharifirad, S. Matwin, When a tweet is actually sexist, in: Proceedings of the Workshop on Abusive Language Online (ALW), ACL, 2019, pp. 22–26. doi:10.18653/v1/W19-3503.

[7] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of abusive language: the problem of biased datasets, in: Proceedings of NAACL, 2019, pp. 602–608. doi:10.18653/v1/N19-1063.

[8] V. Kumar, D. Ghosal, A. Ekbal, P. Bhattacharyya, Data augmentation using pre-trained transformer models for hate speech detection, in: Proceedings of the Workshop on Combating Online Hostile Posts, EACL, 2021, pp. 139–148. doi:10.18653/v1/2021.constraint-1.11.

[9] H. Zhang, M. Cissé, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations (ICLR), 2018. URL: https://openreview.net/forum?id=r1Ddp1-Rb.

[10] J. Nam, J. Kim, I. Gurevych, S. Loos, Large-scale multi-label text classification—Revisiting neural networks, in: Proceedings of ECML-PKDD, 2017, pp. 437–452. doi:10.1007/978-3-319-71249-9_26.

[11] C. N. Silla Jr., A. A. Freitas, A survey of hierarchical classification across different application domains, Data Min. Knowl. Disc. 22 (2011) 31–72. doi:10.1007/s10618-010-0175-9.

[12] H. Zhang, Y. Zhao, Y. LeCun, Ensemble learning methods for deep neural networks: A survey, arXiv preprint arXiv:2005.13590 (2020). URL: https://arxiv.org/abs/2005.13590.

[13] Z. H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC Press, Boca Raton, FL, 2012. ISBN: 9781439830031.

[14] L. Ouyang, J. Wu, X. Jiang, et al., Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155 (2022). doi:10.48550/arXiv.2203.02155.

[15] J. Wei, X. Wang, D. Schuurmans, et al., Chain-of-thought prompting elicits reasoning in large language models, arXiv preprint arXiv:2201.11903 (2022). doi:10.48550/arXiv.2201.11903.

[16] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of NAACL, 2016, pp. 88–93. doi:10.18653/v1/N16-2013.

[17] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. URL: https://www.R-project.org.

[18] Hugging Face, Transformers library, Version 4.x, 2023. URL: https://huggingface.co/transformers/.

[19] P. Liu, W. Xu, N. He, et al., The role of thresholding strategies in multi-label classification: empirical insights and practical guidelines, Pattern Recogn. 120 (2021). doi:10.1016/j.patcog.2021.108148.

[20] T. Wolf, L. Debut, V. Sanh, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of EMNLP: System Demonstrations, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.