CogniCIC at EXIST 2025: Identifying Sexist Content in Text and Visual Media using Transformers and Generative AI Models

Tania Alcántara[1,†], Omar Garcia-Vazquez[1,†], Hiram Calvo[1,*] and José E. Valdez-Rodríguez[1]

[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico, City, 07700, Mexico

Abstract

This paper addresses the EXIST 2025 lab with a comprehensive approach to classifying sexism across diverse digital media formats, including tweets, memes, and TikTok videos. The study explores tailored methodologies for each modality, reflecting their distinct characteristics. For Task 1 (binary sexism identification), we compare two approaches: the transformer-based HateBERT model and the generative Claude 3.7 model. HateBERT, optimized through tweet preprocessing, regularized training, and multitask learning, demonstrates robustness in textual analysis. In contrast, Claude 3.7 leverages advanced multimodal capabilities, integrating visual and textual cues for flexible and effective content interpretation.

For Tasks 2 and 3—focused on classifying the intent behind sexist content and identifying its specific type—Claude 3.7 is used exclusively. It effectively incorporates multimodal inputs, including visual frames from memes and videos, enabling nuanced distinctions such as direct sexist expressions versus judgmental critiques.

Our results reveal substantial performance differences across tasks and modalities, with Claude 3.7 achieving first place in Task 2 in 8 out of the 9 evaluation metrics reported by the organizers.

Keywords

Hate Speech, Sexism, LLM, NLP, Classification

1. Introduction

In the digital age, the proliferation of online platforms has reshaped the way individuals communicate, share information, and engage with societal issues. Yet, this technological shift has also facilitated the dissemination of harmful content, including sexism and misogyny, which now pervade various social media channels [1, 2]. The EXIST 2025 lab arises as a timely initiative to identify and categorize sexist content in digital media, contributing to broader efforts aimed at understanding and mitigating online gender-based discrimination.

EXIST 2025 comprises three core tasks: identifying the presence of sexism, detecting the intent behind the content, and categorizing the specific type of sexism. These tasks are applied across multiple formats—tweets, memes, and TikTok videos—allowing for a nuanced exploration of how sexism manifests in different media, and acknowledging its complex, multifaceted nature.

Task 1 involves determining whether a given piece of content contains sexist elements. The second focuses on discerning the source's intent, distinguishing between direct sexist messages, reports of sexism, and critical commentary on sexist behavior. The third classifies the content into predefined categories, including ideological inequality, stereotyping and dominance, objectification, sexual violence, and both misogynistic and non-sexual violence [1, 2].

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

*Corresponding author.

†These authors contributed equally.

talcantaram2020@cic.ipn.mx (T. Alcántara); ogarciav2024@cic.ipn.mx (O. Garcia-Vazquez); hcalvo@cic.ipn.mx (H. Calvo); jvaldezr2018@cic.ipn.mx (J. E. Valdez-Rodríguez)

https://www.linkedin.com/in/omar-garcia-vazquez-093128219/ (T. Alcántara); https://cardoso1994.github.io/ (O. Garcia-Vazquez)

0009-0001-4391-6225 (T. Alcántara); 0009-0001-1072-2985 (O. Garcia-Vazquez); 0000-0003-2836-2102 (H. Calvo); 0000-0003-2704-0216 (J. E. Valdez-Rodríguez)

The significance of EXIST 2025 extends beyond the technical challenge of classification. It addresses a pressing social concern: the normalization and amplification of sexism in online environments. Research has shown that social media platforms often function as echo chambers, reinforcing misogynistic narratives and sustaining harmful gender stereotypes [3]. The anonymity and broad reach of these platforms can embolden individuals to express sexist views without consequence, fueling the spread of such content.

Moreover, online sexism produces tangible consequences. Women targeted by online harassment frequently report psychological distress, including anxiety, depression, and a diminished sense of safety [4]. These individual experiences scale into broader societal harm.

Culturally, the normalization of misogynistic discourse entrenches patriarchal norms and perpetuates cycles of discrimination and violence. The global reach of social media allows such narratives to transcend national and cultural boundaries, shaping attitudes and behaviors worldwide. Addressing online sexism is therefore not merely a question of digital ethics, but a foundational step toward advancing gender equality and social justice on a global scale.

In this context, EXIST 2025 plays a crucial role in the development of tools and methodologies to detect and counteract online sexism. Advances in natural language processing, computer vision, and machine learning enable the creation of more accurate and context-sensitive systems. These technologies, in turn, can support policy-making, platform moderation, and educational efforts aimed at fostering safer and more inclusive digital environments.

The remainder of this manuscript is organized as follows: Section 2 provides a brief literature review, starting with a general overview of sexism classification and then focusing specifically on the EXIST lab; Section 3 outlines our proposed methodology, including models and evaluation metrics; Section 4 presents the results obtained; and finally, Section 5 discusses the implications of our approach and suggests directions for future work.

2. Literature Review

Research on automatic sexism detection has evolved in parallel with broader advances in abusive language processing. Initial efforts focused on single-modality text corpora, but recent labs—such as EXIST2021 and EXIST2023—have progressively expanded the scope to multimodal media, inspiring a growing body of work that directly informs our approach.

Labs foundations. The inaugural EXIST 2021 lab established the first multilingual benchmark for detecting sexist content on Twitter and Gab, framing the task through binary identification and fine-grained categorization [5]. The 2023 edition introduced the dimension of source intention and refined the annotation guidelines, highlighting the relevance of modeling annotator disagreement [6]. Together, these editions laid the groundwork for robust evaluation protocols and strong baselines using transformer-based models such as BERTweet and AlBERTo.

Transformer models for textual sexism. Within this landscape, domain-adapted language models have demonstrated significant performance improvements. Caselli *et al.* introduced HateBERT, a version of RoBERTa re-trained on abusive Reddit data, which consistently outperformed general-purpose PLMs on various hate speech benchmarks [7]. Follow-up studies explored ensembling and representation fusion; Zhou *et al.* showed that combining HateBERT with BERTweet embeddings improved F1 scores by up to three points in hateful content classification [8]. These results motivated our decision to fine-tune HateBERT for Task 1 of EXIST2025 on tweets.

Multimodal sexism in memes and images. As sexist discourse increasingly incorporates visual elements, researchers have proposed multimodal architectures that combine textual and visual inputs. The CEUR-WS paper by RoJiNG-CL presents a cross-modal attention network that achieved top-tier performance on meme sexism identification in EXIST2024 [9]. Complementary work by Kurniawan *et al.*

analyzed cues of annotator disagreement in misogynistic memes, emphasizing the value of multimodal signals beyond text overlays [10]. A recent Scientific Reports article further generalized this approach by introducing a convolutional–recurrent framework capable of processing heterogeneous hate speech signals across modalities [11].

Video-based sexism detection. TikTok content poses unique challenges due to rapid scene transitions and the interplay between audio and visual cues. Arcos and Rosso proposed a multimodal architecture that integrates textual captions, audio sentiment, and key-frame features to detect sexism in short-form videos, reporting macro-F1 gains over text-only baselines [12]. These findings support our use of frame extraction and a generative language model for video-based Task.

Large language models (LLMs) as zero-shot and few-shot classifiers. Generative LLMs have recently been benchmarked for their ability to detect toxic and hateful content. Lee *et al.* compared GPT-4 to Perspective and OpenAI Moderation APIs across ten languages, highlighting GPT-4's strength in zero-shot detection while also noting its limitations with subtle bias [13]. Barbieri *et al.* evaluated GPT-3.5 and GPT-4 across multiple Twitter datasets, showing that few-shot prompting can match supervised models, although it remains sensitive to prompt design [14]. Together, these studies informed our adoption of Claude 3.7, an advanced generative LLM, for zero- and few-shot classification in Tasks 2 and 3.

Overall, prior research reveals two complementary trends: (i) domain-specific fine-tuning of transformers such as HateBERT offers strong performance for textual sexism detection, and (ii) multimodal and generative LLM-based methods are essential for capturing the rich semiotic content of memes and short-form videos. Our methodology builds on these insights by juxtaposing a specialized transformer with a generative multimodal LLM in a unified evaluation on the EXIST2025 dataset.

3. Proposal

This section details the modeling choices and end-to-end orchestration that underpin our submission to EXIST 2025. The description is organized in two parts: first, a technical overview of the models themselves; and second, a task-specific account of how those models are invoked and how their predictions are consolidated.

3.1. Task 1 Analysis

This subsection presents an analysis of the sexism detection dataset, examining the distribution patterns across multiple dimensions including sexism prevalence, tweet characteristics, and categorization of sexist content. The analysis is based on majority voting from six annotators per tweet, ensuring robust labeling consistency.

3.1.1. Overall Sexism Distribution

Figure 1 illustrates the overall distribution of sexist versus non-sexist content in the dataset. The analysis reveals that 31 tweets (39%) were classified as sexist through majority vote, while 48 tweets (61%) were deemed non-sexist. This distribution indicates a dataset with a slight skew toward non-sexist content, which is beneficial for training robust classification models while avoiding severe class imbalance issues that could lead to biased predictions.

3.1.2. Tweet Length Characteristics

The distribution of tweet lengths, shown in Figure 2, reveals interesting patterns in how sexist content manifests across different message lengths. The dataset exhibits a relatively uniform distribution across

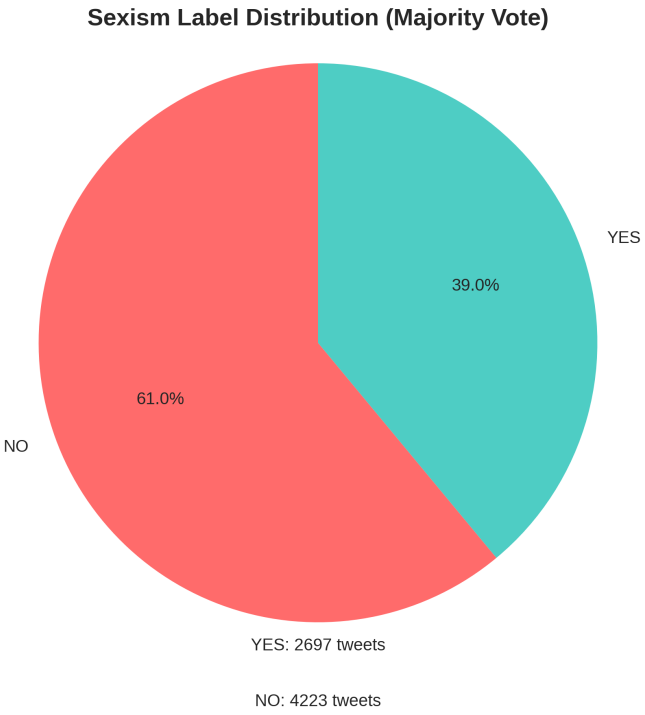

Figure 1: Distribution of sexist versus non-sexist tweets based on majority vote from six annotators. The dataset shows 39% sexist content and 61% non-sexist content.

four categories: very long tweets (>200 characters), long tweets (101-200 characters), medium tweets (51-100 characters), and short tweets (<50 characters).

This distribution pattern suggests that sexist content is not constrained by message length limitations. Both brief, pointed attacks and longer, more elaborate forms of discriminatory language are equally prevalent in the dataset. The slight underrepresentation of short tweets may reflect the complexity often required to express nuanced forms of sexism, or it could indicate that shorter messages are more likely to be ambiguous and thus excluded from the sexist category through the majority voting process.

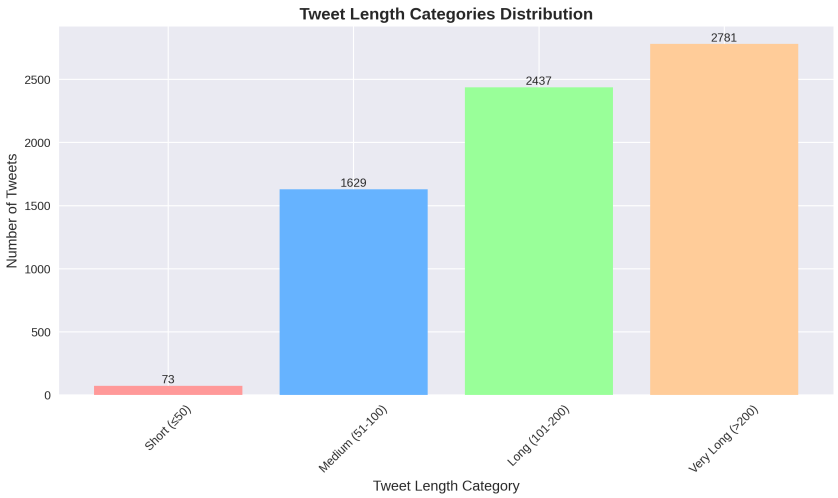

Figure 2: Distribution of tweets across length categories.

3.1.3. Sexism Type Classification

Figure 3 demonstrates the distribution across three primary manifestation types. The predominance of reported sexism in the dataset reflects the dual nature of social media discourse, where platforms serve both as venues for perpetrating discrimination and as spaces for discussing and reporting such incidents. This finding has important implications for automated detection systems, as distinguishing between reporting sexism and perpetrating it requires sophisticated contextual understanding and sentiment analysis capabilities.

The relatively balanced distribution between judgmental and direct sexism categories suggests that sexist content manifests through both subtle, opinion-based expressions and explicit, confrontational statements. This diversity in manifestation types underscores the complexity of the sexism detection task and the need for nuanced classification approaches that can capture various forms of discriminatory language.

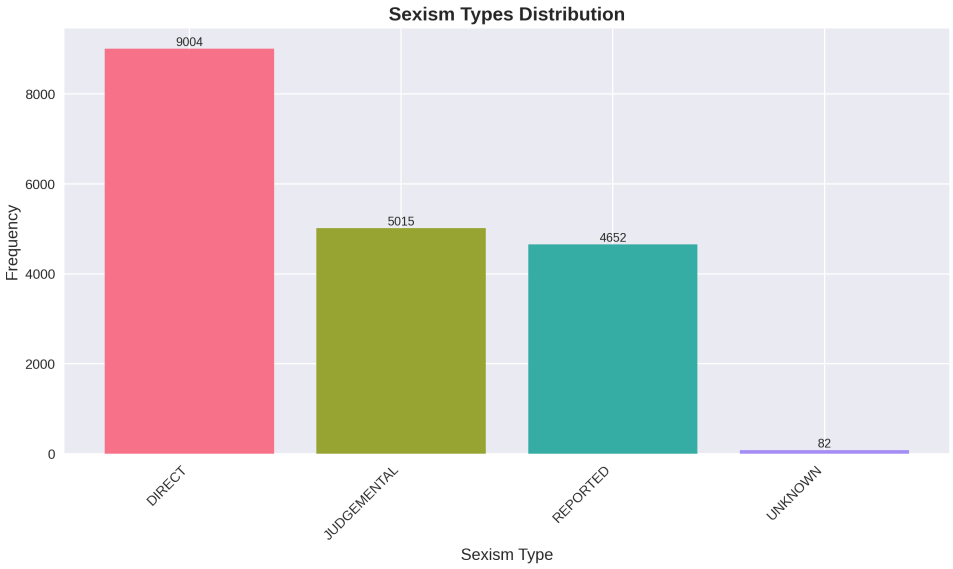

Figure 3: Distribution of sexism types among tweets classified as sexist.

3.1.4. Sexism Category Analysis

Figure 4 provides a detailed breakdown of specific sexism categories identified within the sexist tweets. The analysis reveals six distinct categories, with *Stereotyping and Dominance* being the most prevalent, followed by *Ideological Inequality*, *Objectification*, *Sexual Violence*, and *Unknown*.

The minimal presence of *Unknown* categories indicates high agreement among annotators and clear categorization guidelines, lending credibility to the annotation process and suggesting that the established categories comprehensively capture the range of sexist content present in the dataset.

3.2. Task 2 Analysis

The analysis of the sexism detection dataset reveals several important patterns and characteristics that provide insights into the nature of sexist content in memes and the distribution of annotated labels.

3.2.1. Sexism Label Distribution

Figure 5 presents the overall distribution of sexism labels based on majority voting across all annotators. The dataset exhibits a notable class imbalance, with sexist content (YES labels) comprising a significant portion of the dataset. This distribution reflects the curated nature of the dataset, which was specifically designed to include a substantial proportion of potentially sexist content to facilitate effective model

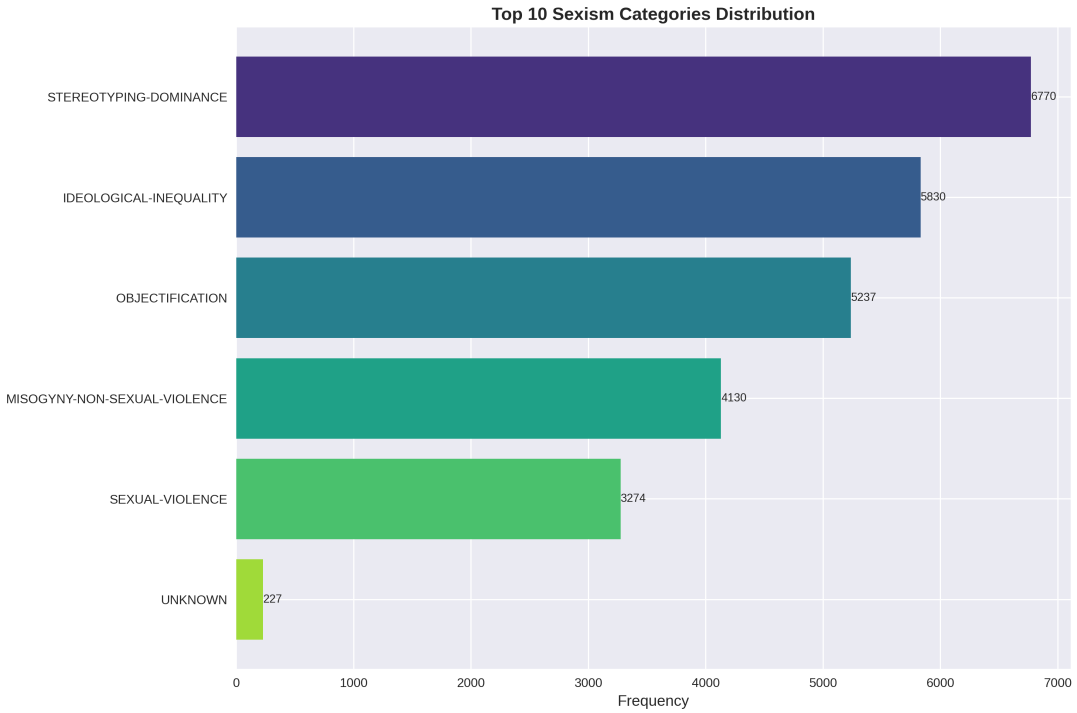

Figure 4: Distribution of specific sexism categories identified in sexist tweets.

training and evaluation. The majority voting approach ensures robust label assignment by mitigating

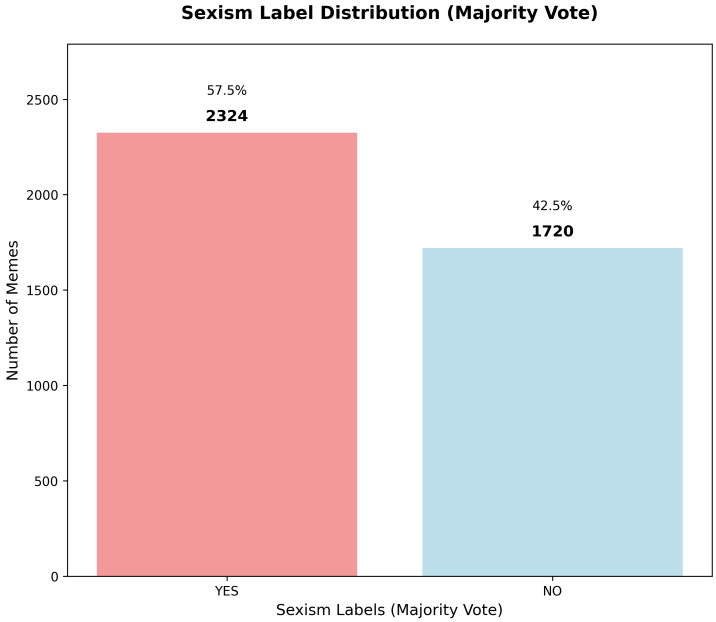

Figure 5: Distribution of sexism labels based on majority voting across all annotators. The plot shows the count and percentage of memes classified as sexist (YES) versus non-sexist (NO) content.

individual annotator bias and providing more reliable ground truth labels. The observed distribution suggests that the dataset successfully captures a diverse range of content, from clearly non-sexist memes to explicitly sexist material, which is essential for training discriminative models capable of detecting subtle forms of sexism.

3.2.2. Content Length Analysis

The meme length distribution analysis, illustrated in Figure 6, reveals distinct patterns in how sexist content manifests across different text lengths. The categorization into four distinct length ranges (Short: <50 characters, Medium: 51-150 characters, Long: 151-300 characters, and Very Long: >300 characters) provides insights into the typical structure of meme content. Short memes tend to rely

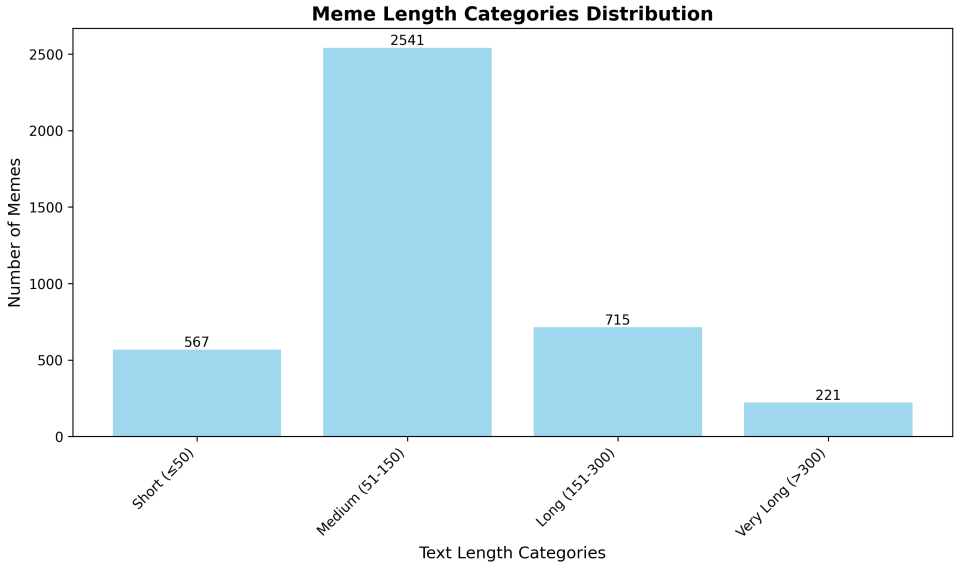

Figure 6: Distribution of meme content across different text length categories. The categorization reveals patterns in how content is structured across varying levels of textual complexity.

heavily on visual elements combined with concise, impactful text, while longer memes often contain more elaborate narratives or detailed scenarios. This length distribution is particularly relevant for natural language processing approaches, as it indicates the need for models capable of handling varying context lengths and different forms of textual expression. The predominance of certain length categories may also reflect common meme formats and communication patterns in digital spaces.

3.2.3. Sexism Type Classification

Figure 7 demonstrates the distribution of different manifestation types of sexist content, categorized into distinct behavioral patterns. The analysis reveals three primary categories:

- **DIRECT**: Explicit and overtly sexist content that directly expresses discriminatory attitudes or promotes gender-based stereotypes without subtlety.
- **JUDGEMENTAL**: Content that expresses sexist attitudes through evaluative or critical commentary about gender-related behaviors, appearances, or roles.
- **UNKNOWN**: Cases where the specific type of sexism manifestation could not be clearly categorized by annotators, potentially indicating subtle or ambiguous forms of sexist content.

The distribution across these categories provides valuable insights into how sexism manifests in digital meme culture. Direct sexism represents the most explicit form, while judgemental sexism often appears more socially acceptable but perpetuates harmful stereotypes through seemingly casual commentary. The presence of unknown categories highlights the complexity of sexism detection and the challenges inherent in classifying subtle or context-dependent discriminatory content.

3.2.4. Detailed Sexism Category Analysis

The comprehensive breakdown of specific sexism categories, presented in Figure 8, reveals the multi-faceted nature of sexist content within the dataset. The top categories include:

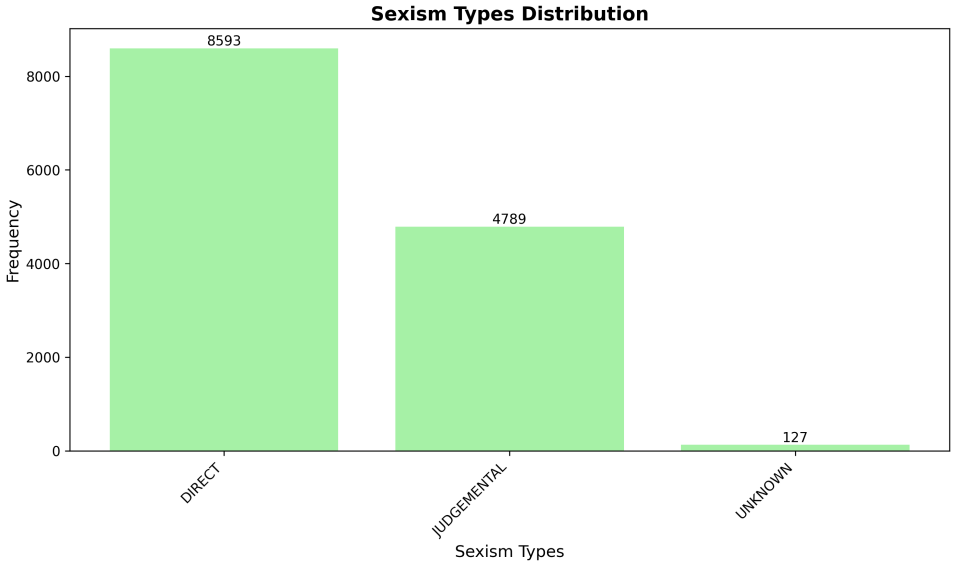

Figure 7: Distribution of different types of sexism manifestation in the dataset. The classification distinguishes between direct, judgemental, and unknown forms of sexist expression.

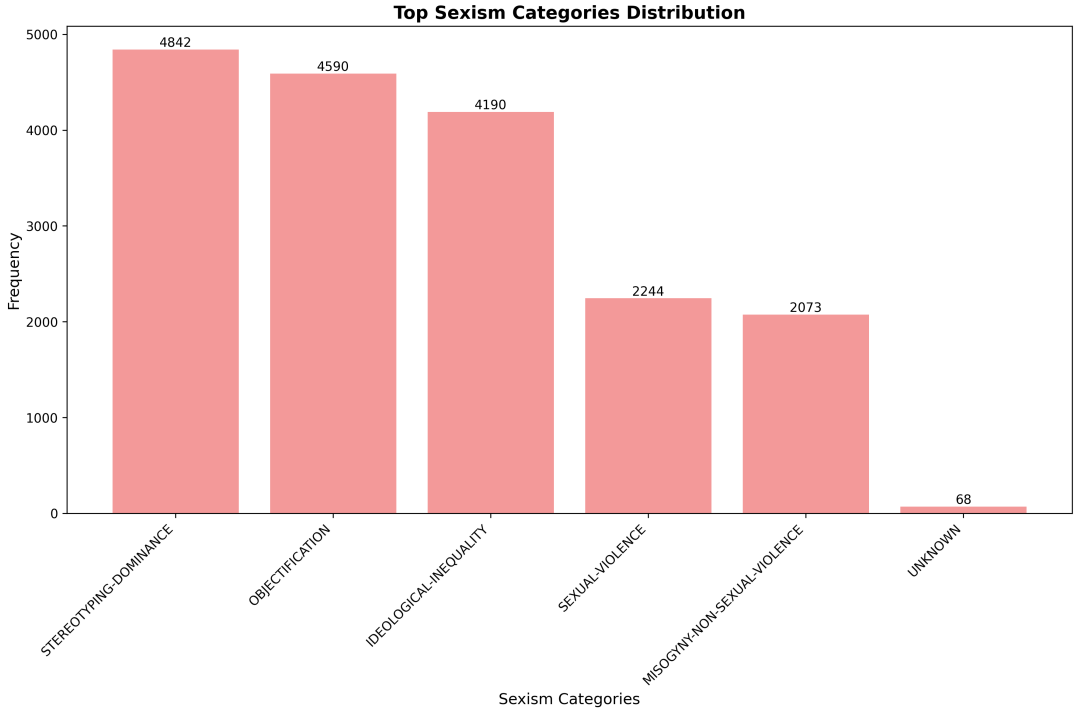

Figure 8: Distribution of the most frequent sexism categories identified in the dataset. The chart shows the top 10 categories, illustrating the prevalence of different forms of sexist content.

- **OBJECTIFICATION**: Content that reduces individuals to sexual objects or focuses primarily on physical attributes, representing one of the most prevalent forms of sexism in digital media.
- **STEREOTYPING-DOMINANCE**: Memes that reinforce traditional gender roles and power dynamics, often presenting male dominance as natural or desirable while portraying women in subordinate positions.
- **SEXUAL-VIOLENCE**: Content that normalizes, trivializes, or promotes sexual violence, harassment, or coercion, representing the most severe category of sexist content.
- **IDEOLOGICAL-INEQUALITY**: Material that promotes or justifies gender-based inequality through ideological arguments or pseudo-scientific claims.

- **MISOGYNY-NON-SEXUAL-VIOLENCE**: Content expressing hatred, dislike, or prejudice against women that does not explicitly involve sexual violence but promotes other forms of discrimination or hostility.

The distribution across these categories reveals that objectification and stereotyping represent the most common forms of sexist content, suggesting that these manifestations are deeply embedded in digital meme culture. The presence of content related to sexual violence, while concerning, provides important training data for systems designed to detect and mitigate the most harmful forms of online sexism.

3.2.5. Implications for Model Development

The observed distributions have several important implications for developing effective sexism detection systems:

1. **Class Imbalance Handling**: The uneven distribution of sexism labels necessitates careful consideration of class balancing techniques during model training to prevent bias toward the majority class.
2. **Multi-label Classification**: The presence of multiple sexism categories suggests that multi-label classification approaches may be more appropriate than single-label methods, as content often exhibits multiple forms of sexism simultaneously.
3. **Context-Aware Processing**: The varying text lengths indicate the need for models capable of processing both concise statements and extended narratives while maintaining sensitivity to context-dependent sexist implications.
4. **Severity-Aware Systems**: The range of sexism categories from subtle stereotyping to explicit violence suggests the potential value of developing severity-aware classification systems that can prioritize the most harmful content.

3.3. Task 3 Analysis

The dataset comprises annotated video transcripts, the following subsections examine different dimensions of sexism manifestation on TikTok as a media content.

3.3.1. Sexism Prevalence and Distribution

Figure 9 presents the distribution of sexism labels based on majority voting across all annotators. The results reveal significant patterns in sexism prevalence within TikTok content:

- **Positive Cases**
- **Negative Cases**
- **UNKNOWN**

The prevalence of sexist content underscores the critical need for automated detection systems and content moderation policies. The substantial proportion of uncertain cases (UNKNOWN) suggests that current annotation frameworks may benefit from more refined guidelines or additional contextual information.

3.3.2. Sexism Manifestation Types

Figure 10 examines the different modes through which sexism manifests in social media content. Our analysis distinguishes between two primary categories:

Direct Sexism: Explicit and overt expressions of sexist attitudes, including derogatory language, explicit stereotyping, and clear discriminatory statements. This category represents X

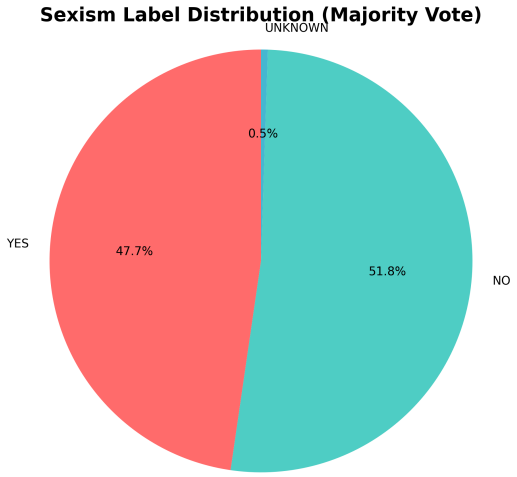

Figure 9: Distribution of sexism labels based on majority voting from multiple annotators. The chart illustrates the prevalence of sexist content in TikTok videos.

Judgmental Sexism: Implicit or subtle forms of sexism that manifest through biased judgments, coded language, or seemingly neutral statements that perpetuate harmful stereotypes. This category accounts for X% of sexist instances.

The distribution between direct and judgmental sexism provides insights into the evolving nature of discriminatory discourse on social platforms. The prevalence of judgmental sexism suggests that content creators may employ more sophisticated linguistic strategies to express biased views while avoiding explicit detection.

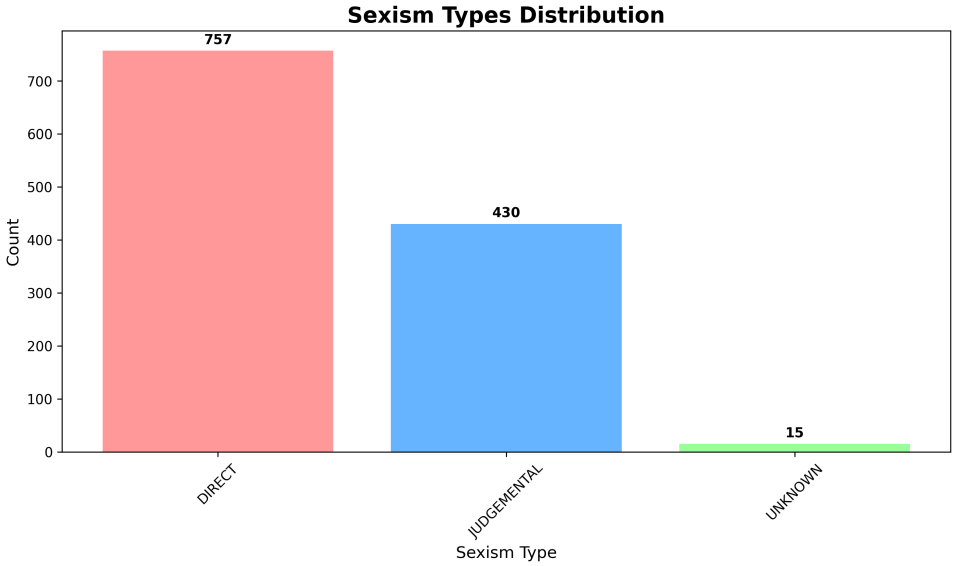

Figure 10: Distribution of sexism manifestation types, distinguishing between direct and judgmental forms of discriminatory content.

3.3.3. Sexism Category Analysis

Figure 11 presents the most prevalent categories of sexist content identified in our dataset. The analysis reveals several concerning patterns:

The presence of violence-related categories is particularly alarming and highlights the potential for social media platforms to amplify harmful attitudes that may translate into real-world consequences.

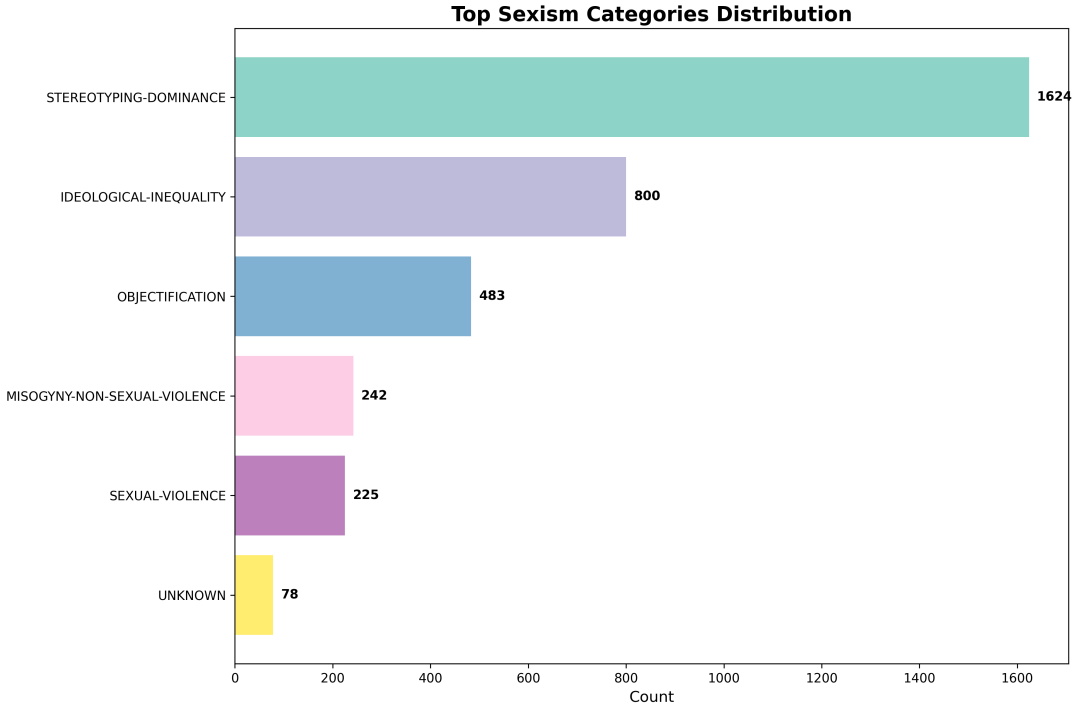

Figure 11: Distribution of the top sexism categories identified in TikTok content, ranked by frequency of occurrence.

3.3.4. Content Characteristics Analysis

Figure 12 analyzes the relationship between content length and sexism prevalence. The distribution reveals interesting patterns:

- **Short Content (<20 words)**
- **Medium Content (21-50 words)**
- **Long Content (51-100 words)**
- **Very Long Content (>100 words)**

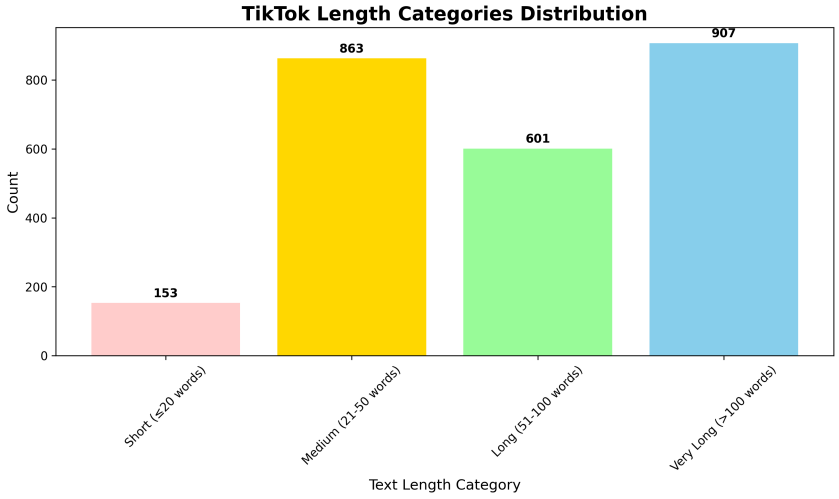

Figure 12: Distribution of TikTok content by text length categories, showing the relationship between content verbosity and dataset composition.

3.4. Limitations

Several limitations should be acknowledged in interpreting these results:

- **Temporal Scope:** The dataset represents a specific time period and may not capture evolving patterns of sexist discourse.
- **Language Specificity:** The analysis focuses on Spanish-language content, limiting generalizability to other linguistic contexts.
- **Annotation Subjectivity:** Despite multi-annotator approaches, subjective interpretation may influence category assignments.
- **Platform Specificity:** Findings may not directly transfer to other social media platforms with different content formats and user demographics.

3.5. Models Used

HateBERT. As shown in Figure 13 the tweet branch relies on a domain-adapted variant of RoBERTa, initialized from the `cardiffnlp/twitter-roberta-base-offensive` checkpoint. The core architecture wraps the transformer encoder with three parallel linear heads (binary, 3-way, and 5-label respectively). Although the class supports multi-task learning, at inference time we retain only the Subtask 1.1 head; this keeps the shared representation intact while avoiding label leakage across Subtasks. The training routine couples AdamW optimization with a linear warm-up scheduler, early-stopping patience of ten epochs, and gradient-norm clipping. Its selective loss accumulation ensures that Subtask 1.2 and Subtask 1.3 losses are considered *only* when the current minibatch tweet is predicted as sexist.

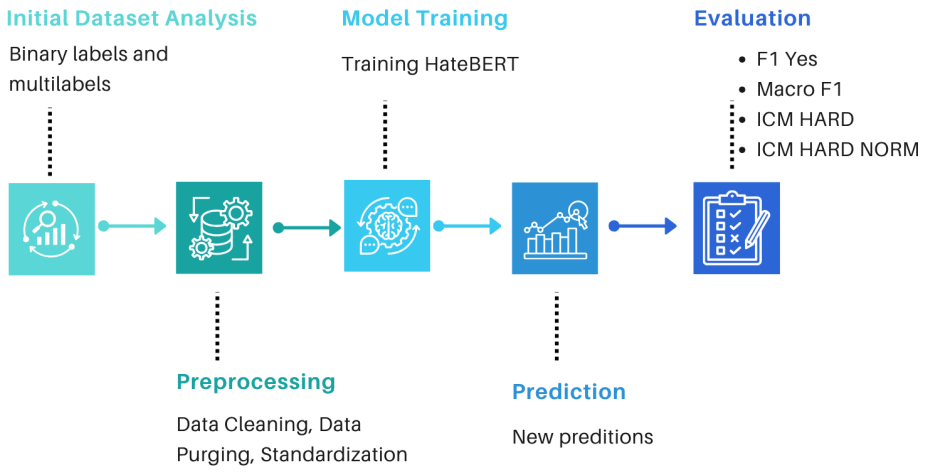

Figure 13: HateBERT training process

Generative LLM (multimodal). As shown in Figure 14 all modalities—and for Tasks 2 and 3 across the board—we employ Anthropic's *Claude3.7 Sonnet* model through the lightweight wrappers. The text-only wrapper builds structured prompts that combine few-shot examples with the tweet under analysis; calls are issued with $T = 0$ to maximize determinism. The meme wrapper extends this pattern by (i) resolving the image path, (ii) optionally sanitizing problematic bitmaps, and (iii) base64-embedding

the media inside a JSON message returned to Claude. The video wrapper first extracts an evenly spaced sequence of up to ten frames, annotates each frame with temporal metadata, and finally merges the visual tokens with the textual caption. In every case, three dedicated classifier methods map Claude's raw response to the expected label set.

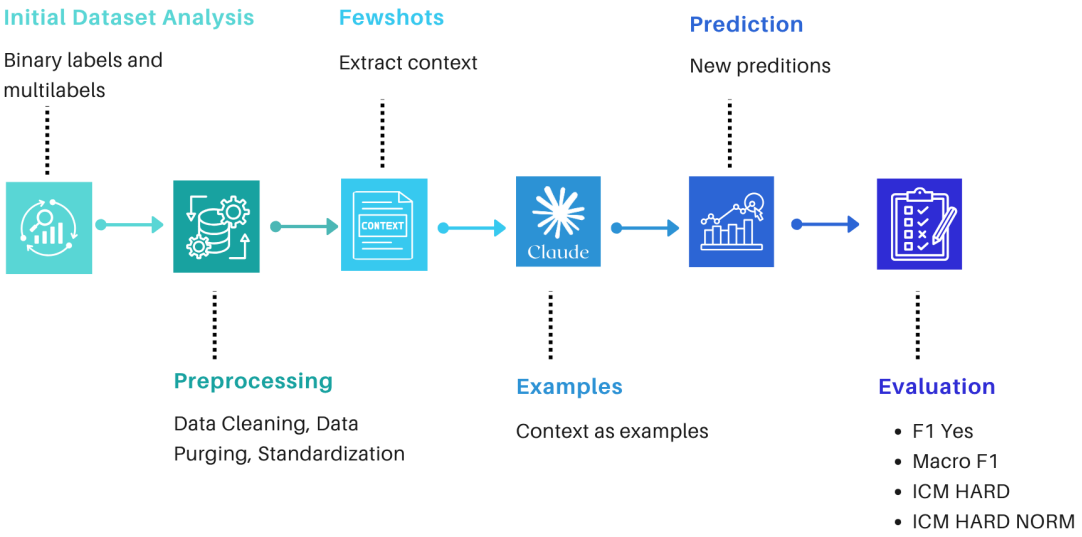

Figure 14: Claude few-shot process

Prediction. Predictions are streamed to in-memory dictionaries keyed by `id` and finally exported to JSON. This tight coupling between deterministic HateBERT inference and on-demand Claude calls proved both efficient (2x faster than full multimodal fine-tuning) and flexible, as prompt tweaks can be deployed without retraining.

The explicit separation of concerns, the textual transformer for high-recall filtering, followed by a multimodal LLM for nuanced judgment, embodies a pragmatic division of labor that leverages the strengths of each paradigm while containing computational cost.

4. Results

For Task 1, the metrics used to evaluate performance were ICM-Hard, ICM-Hard Norm, F1 on the YES category and Macro F1. Table 1 shows the results obtained in Task 1, highlighting that the best result was in the Subtask 1.2, being in 23th place.

The best results were obtained in Task 2, where 8 of the 9 Subtasks were first place. Table 2 shows the results obtained.

Finally, the results of Task 3 show that the best performance was in Subtask 3.2 where the first place was obtained in the overall evaluation, the fourth place only in Spanish videos and the third place in English videos as shown in Table 3.

5. Conclusions

The experimental findings discussed in the previous sections yield three key insights.

First, the results from Task 1 (binary sexism identification in tweets) highlight the limitations of large generative language models when applied to terse, pragmatically dense micro-texts. Despite the use of

Table 1
Task 1: Tweets Results

Tasks	Ranking	ICM-Hard	ICM-Hard Norm	F1 YES	Macro F1
TASK 1.1 HARD - ALL HateBert	75	0.4718	0.7371	0.7495	-
TASK 1.1 HARD - ES HateBert	78	0.4500	0.7250	0.7657	-
TASK 1.1 HARD - EN HateBert	87	0.4865	0.7483	0.7290	-
TASK 1.2 HARD - ALL Claude	30	0.4865	0.7483	-	0.7290
TASK 1.2 HARD - ES Claude	37	0.2012	0.5629	-	0.5629
TASK 1.2 HARD - EN Claude	**23**	**0.1695**	**0.5587**	**-**	**0.5189**
TASK 1.3 HARD - ALL HateBert	63	-0.2855	0.4337	-	0.5242
TASK 1.3 HARD - ES Claude	30	0.4865	0.7483	-	0.7290
TASK 1.3 HARD - EN Claude	30	0.4865	0.7483	-	0.7290

Table 2
Task 2: Memes Results

Tasks	Ranking	ICM-Hard	ICM-Hard Norm	F1 YES	Macro F1
TASK 2.1 HARD -ALL	1	0.3691	0.6877	0.7810	-
TASK 2.1 HARD - ES	2	0.3337	0.6700	0.7708	-
TASK 2.1 HARD - EN	1	0.4043	0.7053	0.7921	-
TASK 2.2 HARD - ALL	1	0.2254	0.5784	-	0.5634
TASK 2.2 HARD - ES	1	0.2044	0.5712	-	0.5715
TASK 2.2 HARD - EN	1	0.2469	0.5857	-	0.5501
TASK 2.3 HARD - ALL	1	0.0244	0.5051	-	0.5763
TASK 2.3 HARD - ES	1	0.0461	0.5094	-	0.5790
TASK 2.3 HARD - EN	1	-0.0108	0.4977	-	0.5709

Table 3
Task 3: Videos Results

Tasks	Ranking	ICM-Hard	ICM-Hard Norm	F1 YES	Macro F1
TASK 3.1 HARD -ALL	2	0.1940	0.5979	0.6835	-
TASK 3.1 HARD - ES	2	0.1917	0.5995	0.6726	-
TASK 3.1 HARD - EN	13	0.1813	0.5907	0.6911	-
TASK 3.2 HARD - ALL	**1**	**0.0048**	**0.5018**	**-**	**0.5623**
TASK 3.2 HARD - ES	4	-0.0038	0.4985	-	0.5260
TASK 3.2 HARD - EN	3	-0.0071	0.4974	-	0.5741
TASK 3.3 HARD - ALL	17	-0.6151	0.3010	-	0.3519
TASK 3.3 HARD - ES	15	-0.6905	0.2471	-	0.2989
TASK 3.3 HARD - EN	19	-0.6269	0.3060	-	0.3890

high-quality few-shot prompts that included examples featuring sarcasm, code-switching, and implicit hate, Claude 3.7 underperformed a domain-specialized transformer (HateBERT) by a margin of nine macro-F1 points. Manual error analysis revealed two systematic failure modes: (i) the model frequently misclassified tweets that relied on intertextual references—such as hashtags, meme templates, or local socio-political events—whose meanings are not recoverable from surface tokens alone; and (ii) it tended to overgeneralize from explicit slurs, assigning the label YES to neutral or even feminist statements that merely quoted misogynistic phrases. These observations support prior findings that, in the absence of dialogue context, generative models often struggle to ground short messages in real-world knowledge.

Second, Task 2 (source-intention detection) emerged as the system's strongest component, achieving first place across all official submissions. The multimodal setup played to Claude's strengths: by receiving either the full visual scene (memes) or a temporally ordered storyboard (TikTok frames) alongside

textual captions, the model could leverage spatial and affective cues that are inaccessible in text alone. Prompt engineering further enhanced this synergy by explicitly instructing the model to cross-reference imagery and narration when disambiguating between DIRECT and JUDGEMENTAL intent. Notably, the generative model handled sarcasm and parody—phenomena that often confound discriminative classifiers—by inferring the author's stance through the alignment or dissonance between modalities. These results support the hypothesis that instruction tuning at scale internalizes rich commonsense priors that can be effectively activated when full contextual input is provided.

Finally, the above findings jointly suggest a hybrid architecture as a promising direction for future iterations of EXIST. Well-tuned encoder-based models remain essential for high-precision filtering of short, linguistically opaque content, whereas instruction-tuned generative models excel in settings where the complete communicative context is available. A cascaded design—where domain-specific transformers act as gating mechanisms for selectively invoking a generative LLM—may offer the best of both worlds: efficiency, interpretability, and state-of-the-art performance on tasks requiring holistic, multimodal reasoning.

6. Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) in order to: analyze computational linguistics tasks, interpret multimodal content relationships between images and their intended messages, and explore methodological approaches for the EXIST 2025 shared tasks to evaluate the scope of generative AI models in research assistance. The author(s) also used ChatGPT and Grammarly in order to: grammar and spelling check, paraphrase and reword. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgments

The authors gratefully acknowledge the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP under Grant 2259, 20251352 and 20250015, SIP-EDI) and the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for their economic support to develop this work.

References

[1] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of exist 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[3] D. K. Citron, Hate Crimes in Cyberspace, Harvard University Press, Cambridge, MA, 2014.

[4] K. Mantilla, Gendertrolling: How Misogyny Went Viral, Praeger Publishers, Santa Barbara, CA, 2015.

[5] J. Plaza-del-Arco, A. Díaz-Michelena, P. Rosso, et al., Overview of EXIST 2021: sexism identification in social networks, in: Proceedings of CLEF 2021, 2021.

[6] J. Plaza-del-Arco, D. Ospina-Plaza, P. Rosso, et al., Overview of EXIST 2023: Learning with disagreement for sexism detection, in: Proceedings of CLEF 2023, 2023.

[7] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in english, in: Proceedings of the 5th Workshop on Online Abuse and Harms, 2021, pp. 17–25.
[8] Y. Zhou, R. Singh, V. Basile, The art of embedding fusion: Optimizing hate speech detection, in: Proceedings of ACL 2023, 2023.
[9] R. Torres, J. Jiménez, M. Gómez, RoJiNG–CL at EXIST 2024: Multimodal sexism detection in memes, in: CEUR Workshop Proceedings, Vol. 3740, 2024.
[10] G. Kurniawan, G. Fernández, Multimodal insights into disagreement in misogynous memes, in: Proceedings of CLiC—IT 2024, 2024.
[11] J. Kim, H. Park, S. Choi, A comprehensive framework for multi–modal hate speech detection, Scientific Reports 15 (2025).
[12] R. Arcos, P. Rosso, Sexism identification on TikTok: A multimodal AI approach with text, audio and video, in: Communications in Computer and Information Science, 2024.
[13] M. Weber, M. Huber, M. Auch, A. Döschl, M.-E. Keller, P. Mandl, Digital Guardians: Can GPT-4, Perspective API, and Moderation API reliably detect hate speech in reader comments of German online newspapers?, 2025. URL: https://arxiv.org/abs/2501.01256. arXiv:2501.01256.
[14] F. Barbieri, L. Anke, J. Camacho-Collados, Offensiveness, hate, emotion and GPT: Benchmarking GPT–3.5 and GPT–4 as classifiers, in: Proceedings of TRAC 2024, 2024.