NLPDame at EXIST: Sexism Categorization in Tweets via Multi-Head Multi-Task Models, LLM & RAG Voting Synergy

Notebook for the EXIST Lab at CLEF 2025

Christina Christodoulou[1,2,3,*]

[1]Department of Computer Science & Engineering, University of Ioannina, Ioannina, Greece
[2]"Archimedes" Research Unit, "Athena" Research Center, Maroussi, Attica, Greece
[3]Institute of Informatics & Telecommunications, National Centre for Scientific Research (N.C.S.R.) "Demokritos", Aghia Paraskevi, Attica, Greece

Abstract

This paper details the author's participation as *NLPDame* in the fifth edition of EXIST (*sEXism Identification in Social neTworks*) as Lab at CLEF 2025. It outlines an approach for the text partitioning of sub-task 1.3, which pertains to the sexism categorization of tweets in English and Spanish using hard labels. The methodology includes fine-tuning 12 Transformer language models within a tailored multi-head and multi-task model architecture that employs CLS, mean, and max pooling for multi-label text classification. The multi-head architecture effectively addresses the multilingual nature of the dataset, while the multi-task architecture incorporates sentiment analysis to enhance the multi-label classification process. The methodology also involves utilizing the open-source multilingual LLM Llama-3.2-3B-Instruct, employing prompt engineering for performing classification. Additionally, a method incorporating RAG, chain-of-thought reasoning and annotators' profiles was used to provide contextual information within the LLM prompt engineering framework. Ultimately, majority voting was applied to test submissions, including the predictions from (i) the 12 Transformer models with LLM prompt engineering, and (ii) the 12 Transformer models with LLM prompt engineering, including chain-of-thought and annotators' profiles, along with RAG. The experimental approach consisted of data analysis, baseline experiments, a multi-step pre-processing pipeline, the application of various loss functions and thresholds, as well as the use of class positive weights to tackle class imbalance. For the hard-hard evaluation in both languages, 3 runs were submitted that ranked 4^{th}, 5^{th}, and 6^{th} out of a total of 132 submissions. The highest-scoring run was based on majority voting from the predictions of the 12 Transformer models, utilizing LLM prompt engineering in conjunction with RAG.

Keywords

sexism categorization, tweets, sub-task 1.3, hard labels, multilingual multilabel text classification, custom multi-head multi-task architectures, LLM, prompt engineering, chain-of-thought, RAG, Transformers, majority voting

1. Introduction

The term *sexism* has emerged during the second-wave feminism movement in the 1960s-1980s and refers to the discrimination, prejudice, and stereotyping directed at individuals based on their sex, predominantly affecting women and girls [1]. Sexism against women remains a pervasive issue in modern society, manifesting in various forms such as workplace discrimination, wage disparity, misogyny, stereotyping and objectification. Incidents of sexual harassment against women, occurring daily, reveal the persistent challenges women face. Hence, the consequences of sexism are grave, adversely affecting women's physical health, mental health, career advancement, and overall quality of life.

The rise and vast reach of social media have brought a surge in the expression of sexist behavior online, exhibited through various types like gender inequality, stereotyping, objectification, sexual violence and misogyny. This behavior has become more prevalent in online environments rather than

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

[*] Corresponding author.

christinachristodoulou1997@gmail.com (C. Christodoulou)
https://github.com/christinacdl/ (C. Christodoulou)
0009-0009-5616-3937 (C. Christodoulou)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4038/paper_147.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

in person, largely due to the anonymity and invisibility offered by these environments. The anonymity and invisibility foster what is known as the *online disinhibition effect*, where people are more likely to act in ways they might not in face-to-face interactions without considering the repercussions [2],[3]. Many women, who have experienced or witnessed such behavior, report feeling discouraged from interacting and posting online, resulting in self-censorship and limited engagement on social media platforms, as they encounter constant criticism, hate, and discrimination, and even feel threatened not only for their safety, but also for their safety of their families [4] [5]. Content monitoring and moderation on social media are essential. Nevertheless, manual social media content moderation is an arduous and time-consuming task as well as susceptible to the moderator's subjective judgment [6]. The constant exposure of unpleasant, violent, and pornographic content during manual moderation has a negative impact on the moderators' mental health [7]. The significant challenges and severe consequences associated with manual moderation, in conjunction with the widespread increase in social media content, underscore the urgent necessity for developing automated systems designed to detect and eliminate sexist content effectively.

EXIST (s**EX**ism **I**dentification in **S**ocial ne**T**works) is a series of shared tasks since 2021, focusing on detecting and categorizing online sexism, including both explicit and implicit expressions of sexism. It comprises 3 tasks: **Sexism Identification** (binary classification), **Source Intention** (multi-class classification) and **Sexism Categorization** (multi-label classification). This year, the fifth edition of EXIST was held as a Lab in CLEF 2025. It consisted of these 3 tasks, which were further divided into 6 sub-tasks. The aim was to detect sexism not only across different formats, including text (tweets), images (memes from Google Search), and videos (TikToks), but also across different languages, meaning English and Spanish [8] [9].

The present paper outlines the system and results from the author's participation as *NLPDame* in this year's sub-task 1.3 of EXIST Lab at CLEF 2025, which focuses on **Sexism Categorization in tweets (hard labels) for both English and Spanish**. The approach involved fine-tuning 12 multilingual Transformer-based language models within a custom multi-head and multi-task architecture designed specifically for multilingual and multi-label text classification, as well as sentiment analysis. This approach aimed to capture linguistic nuances and enhance the accuracy of detecting categories of sexism by integrating sentiment analysis to boost classification performance. Furthermore, the open-source multilingual Large Language Model (LLM) Llama-3.2-3B-Instruct was leveraged with prompt engineering, including the definitions of the sexism categories and the sentiment of the input tweet, to perform multi-label sexism categorization. An additional method, incorporating chain-of-thought reasoning and annotators' demographic background in the prompt in conjunction with Retrieval-Augmented Generation (RAG), was implemented to provide contextual information within the LLM prompt engineering framework. Ultimately, majority voting was applied to evaluate submissions, merging predictions from (i) the 12 Transformer models using LLM prompt engineering and (ii) the 12 Transformer models using LLM prompt engineering (chain-of-thought and annotators' profiles) along with RAG. The experimental approach included comprehensive data analysis, baseline experiments, a multi-step pre-processing pipeline, the application of various loss functions and thresholds, as well as leveraging class positive weights to address class imbalance. The code for the presented approach is available on the provided GitHub link.[1]

The structure of this paper is as follows: Section 2 provides background information concerning sexism detection shared tasks. Section 3 presents various aspects of the data, including data analysis and pre-processing. Section 4 introduces an overview of the developed system, the experiments and the submissions. Section 5 presents and discusses the results from both the development and test sets. Finally, Section 6 concludes the paper by discussing the findings and potential future work, while Section 7 addresses the limitations of the presented approach.

[1] https://github.com/christinacdl/EXIST_2025_CLEF

2. Background

Developing automated systems using Natural Language Processing (NLP) and machine learning methods to detect sexism in social media has gained considerable popularity as a shared task in recent years. The EXIST task was the first to be organized at IberLEF 2021, concentrating on identifying sexism in English and Spanish texts. This involved both determining the presence of sexism (binary classification) and categorizing it (multi-label classification) in posts from Twitter and Gab [10]. The following year, sexism was approached in task 5 of SemEval-2022, titled *Multimedia Automatic Misogyny Identification (MAMI)*, for detecting multi-modal misogyny, particularly in English misogynous memes found online, by leveraging both texts and images. It was divided into 2 sub-tasks: sub-task A, which involved binary classification to determine whether a meme was misogynous, and sub-task B, which dealt with the identification of various types of misogyny (multi-label classification) [11]. Additionally, sexism detection was featured in task 10 of SemEval-2023, named *Explainable Detection of Online Sexism (EDOS)*, which focused on detecting English texts from Gab and Reddit. It was divided into 3 sub-tasks: sub-task A was focused on identifying whether a post is sexist or not (binary classification). Sub-tasks B and C aimed at multi-class classification by identifying 4 categories and 11 fine-grained vectors of sexism, respectively [12].

Since its inception in 2021, the EXIST task has been held annually, concentrating exclusively on detecting sexism in English and Spanish social media texts until 2024 [13], when it expanded to incorporate additional sub-tasks focusing on memes, and in 2025, it further extended to include videos from TikTok [8] [9]. The previous year, in the fourth version of EXIST, 31 systems were submitted for sub-task 3 (Sexism Categorization), the hard-hard evaluation, which is discussed in the present paper. 28 out of the 31 systems managed to surpass the majority class baseline (where all instances are labelled as "NO"), while all systems outperformed the minority class baseline (where all instances are labelled as "SEXUAL-VIOLENCE").

More particularly, the team ABCD achieved the first rank on the test set with an ICM of **0.3713**, ICM-NORM of **0.5862**, and an F_1 score of **0.6004**. They also secured the second rank with an ICM of 0.3540, ICM-NORM of 0.5862, and an F_1 score of 0.6042. The team utilized the xlm-RoBERTa small and large models, the multilingual-T5 version models, and Llama-2-7B-Instruct to attain these results. They divided the dataset into 6 subsets based on the number of annotators, retaining identical tweets while incorporating metadata specific to each annotator. This allowed them to fine-tune 6-component models for each subset and develop an ensemble approach. For the task of sexism categorization, they only considered the predictions for instances classified as sexist from the component models used in sub-task 1, which focused on sexism identification (binary classification). Their methods included prompt engineering and Low-Rank Adaptation (LoRA). Notably, Llama-2 achieved the first rank, while xlm-RoBERTa-large secured the second [14]. The NYCU-NLP team achieved impressive results, securing the third, fourth, and fifth positions in the hard-hard evaluation of sub-task 3. They applied extensive data pre-processing methods, such as deleting redundant elements, standardizing text formats, increasing data diversity by back-translation, and augmenting texts utilizing the AEDA approach. They also incorporated annotator demographics such as gender, age, and ethnicity into the DeBERTa-v3 and xlm-RoBERTa models. Moreover, they adapted the Round to Closest Value approach to deal with non-continuous annotation values and maintain precise probability distributions. Optimization of shared layers across tasks based on the hard parameter-sharing techniques was followed to enhance generalization and computational efficiency [15]. Other teams, like Awakened, fine-tuned pre-trained models only in English and multilingual models as well as domain-specific models, such as twitter-xlm-roberta-base-sentiment or roberta-hate-speech-dynabench-r4, while also leveraging ensembling methods [16].

3. Data

3.1. Dataset & Data Analysis

Established in 2021, the EXIST shared task is designed to advance research in sexism detection, initially concentrating on textual data such as tweets. The EXIST 2025 edition introduced a significant expansion by connecting multiple data sources and modalities, including TikTok videos, memes from Google Search, and tweets, for a more holistic approach to multimedia and multimodal sexism detection [8] [9]. The approach presented in this paper focuses solely on the textual component of the EXIST 2025 Dataset, with particular emphasis on sub-task 1.3, which pertains to hierarchical multi-label Sexism Categorization in tweets. The approach employs the official hard labels (gold standard) for training and evaluation purposes.

The EXIST 2025 Tweet Dataset consists of tweets in both Spanish and English, resulting in a balanced cross-lingual corpus of over 10,000 annotated tweets. Notably, adhering to the Learning with Disagreement (LeWiDi) paradigm, each tweet is evaluated by 6 annotators from various socio-demographic backgrounds. The background information of the annotators, including details such as gender, age, ethnicity, education level, and country of origin, was provided along with the tweets and their labels in JSON files. Each tweet was represented as a JSON object with the attributes shown in Figure 1. Additionally, an evaluation folder was provided to the participants to assess their system's outputs, which contained two sub-folders. One sub-folder included the official gold standards for all sub-tasks for hard and soft evaluation contexts in both training and development sets. The other sub-folder contained the official baselines for each sub-task.

```
  "200006": {
    "id_EXIST": "200006",
    "lang": "en",
    "tweet": "According to a customer I have plenty of time to go spent the Stirling
    ↪  coins he wants to pay me with, in Derry. \"Just like any other woman, I'm sure
    ↪  of it.\" #EveryDaySexism in retail.",
    "number_annotators": 6,
    "annotators": ["Annotator_409", "Annotator_410", "Annotator_411", "Annotator_412",
    ↪  "Annotator_413", "Annotator_414"],
    "gender_annotators": ["F", "F", "M", "M", "M", "F"],
    "age_annotators": ["18-22", "23-45", "18-22", "23-45", "46+", "46+"],
    "ethnicities_annotators": ["White or Caucasian", "White or Caucasian", "White or
    ↪  Caucasian", "White or Caucasian", "White or Caucasian", "White or Caucasian"],
    "study_levels_annotators": ["Bachelor's degree", "Master's degree", "High school
    ↪  degree or equivalent", "Bachelor's degree", "Doctorate", "High school degree or
    ↪  equivalent"],
    "countries_annotators": ["Estonia", "Romania", "Slovenia", "Greece", "Spain",
    ↪  "United Kingdom"],
    "labels_task1_1": ["YES", "YES", "YES", "YES", "YES", "YES"],
    "labels_task1_2": ["REPORTED", "REPORTED", "REPORTED", "REPORTED", "REPORTED",
    ↪  "JUDGEMENTAL"],
    "labels_task1_3": [
      ["STEREOTYPING-DOMINANCE", "OBJECTIFICATION"],
      ["STEREOTYPING-DOMINANCE"],
      ["OBJECTIFICATION"],
      ["IDEOLOGICAL-INEQUALITY"],
      ["STEREOTYPING-DOMINANCE", "MISOGYNY-NON-SEXUAL-VIOLENCE"],
      ["OBJECTIFICATION"]
    ],
    "split": "TRAIN_EN"
}
```

Figure 1: Raw JSON annotation example of an English tweet contained in the training set. The entry includes the tweet content, annotator demographics, and the sexism labels for all text classification sub-tasks.

For all experiments, the provided training and development sets with their hard labels were utilized for training and evaluation, respectively. Table 1 presents an overview of the hard labels of the sub-task as well as their distribution across the training and development sets, both overall and per language. Hence, from Table 1, it can be observed that:

- The dataset is significantly imbalanced, with a skewed distribution toward the "NO" label in both the training and development sets. This imbalance poses challenges for models, as they may be prone to overpredicting non-sexist outputs unless appropriate class weight balancing techniques or loss weighting strategies are implemented.
- The sexism categories "MISOGYNY-NON-SEXUAL-VIOLENCE" and "SEXUAL-VIOLENCE" are relatively under-represented compared to more prevalent categories like "STEREOTYPING-DOMINANCE" and "IDEOLOGICAL-INEQUALITY", especially in the development set. This under-representation makes it more difficult to learn and evaluate nuanced patterns associated with rarer, but concerning and significant forms of sexist expression.
- The dataset features two languages, English and Spanish, and maintains a relatively balanced distribution between them. Except for the "NO" category, the majority of examples in the other categories are slightly skewed towards Spanish. This cross-linguistic aspect highlights the need to integrate language-specific components into the model to effectively capture cultural and linguistic variations.
- The dataset is designed for hierarchical multi-label classification, which means that each tweet can first be identified as either sexist or non-sexist. If a tweet is identified as sexist, it can then be categorized into multiple overlapping types of sexism. This introduces considerable complexity to the learning task compared to simpler binary or multi-class problems, as models must not only accurately differentiate between sexist and non-sexist tweets but also identify various combinations of intertwined sexist behaviors.

Overall, the inherently multilingual nature of the dataset, which encompasses both English and Spanish tweets, was carefully considered throughout the stages of data pre-processing, model design, training, and evaluation. This approach ensured that the system remains adaptable and fair across different linguistic contexts. Additionally, the significant class imbalance within the data was tackled by implementing targeted strategies during training to reduce bias and enhance the model's sensitivity to these nuanced yet crucial expressions of sexism. The background information of the annotators was also incorporated into the prompt engineering process to provide context for the role of the LLM and facilitate accurate classification. This method effectively addresses both the multilingual and imbalanced characteristics of the dataset while taking advantage of the additional information given to develop a robust system.

3.2. Data Pre-processing

The pre-processing pipeline for both English and Spanish tweets followed a multi-step approach. Previous research conducted by the author has demonstrated that pre-processing significantly enhances system outcomes compared to utilizing raw texts for training purposes, particularly tweets, which often include usernames, URLs, non-ASCII characters, and excessive punctuation [17] [18] [19]. The pre-processed data were leveraged for all kinds of experiments, as described below:

1. **Loading Raw Datasets:** The provided JSON files for training, development, and test sets were read, and identifier fields were standardized (e.g., `id_EXIST`) and were merged with the JSON files containing the gold hard labels.
2. **Data Cleaning:** Duplicate tweets were removed based on text content from the training and development sets. Rows with empty or null tweet entries or missing labels were removed. No data cleaning was applied to the test set. As can be observed from Table 2, a considerable amount of text was removed from both sets. More specifically, 870 and 107 texts were deleted in total from the training and development sets. 396 English and 474 Spanish tweets were discarded from the training set, while 45 English and 62 Spanish tweets were removed from the development set.

Table 1
Hard label distribution in the training and development sets in both languages and per language after data cleaning.

Label	All	
	Training Set	**Development Set**
NO	3367	479
IDEOLOGICAL-INEQUALITY	1113	212
STEREOTYPING-DOMINANCE	1423	241
OBJECTIFICATION	1103	183
SEXUAL-VIOLENCE	675	123
MISOGYNY-NON-SEXUAL-VIOLENCE	856	158
Label	**English**	
	Training Set	**Development Set**
NO	1733	250
IDEOLOGICAL-INEQUALITY	481	95
STEREOTYPING-DOMINANCE	613	105
OBJECTIFICATION	492	95
SEXUAL-VIOLENCE	316	49
MISOGYNY-NON-SEXUAL-VIOLENCE	304	68
Label	**Spanish**	
	Training Set	**Development Set**
NO	1634	229
IDEOLOGICAL-INEQUALITY	632	117
STEREOTYPING-DOMINANCE	810	136
OBJECTIFICATION	611	88
SEXUAL-VIOLENCE	359	74
MISOGYNY-NON-SEXUAL-VIOLENCE	552	90

3. **Emoji Conversion:** Emojis were converted into their descriptive textual representations, using language-specific mappings for English or Spanish from the Emoji library[2].
4. **HTML and URL Cleaning:** HTML entities (e.g., `&amp;` to `and`) were unescaped, and all URLs were removed using regular expressions.
5. **Contraction Expansion:** Spanish tweets were expanded using a manually defined contractions dictionary (e.g., `pa'` to `para`, `toy` to `estoy`). The contractions of English tweets were expanded through the Ekphrasis library[3][20].
6. **Accent Normalization (Spanish Only):** Accented characters were normalized to their base forms (e.g., `acción` to `accion`) to ensure consistency.
7. **Punctuation Pattern Substitution:** Repeated punctuation patterns (multiple question marks, multiple exclamation marks and mixed question and exclamation marks) were replaced with standardized placeholders, "??", "!!" and "?!", respectively. Spanish-specific punctuation such as inverted question marks (¿) or exclamation marks (¡) was removed.
8. **Text Normalization with Ekphrasis library:** The Ekphrasis library used the Twitter segmenter and corrector to segment hashtags and correct common misspellings, respectively. The usernames, emails, numbers, dates, and phone numbers were converted into <user>, <email>, <number>, <date>, and <phone>, respectively, for anonymization[20]. These tokens were added as special tokens into the model tokenizers.
9. **Repetition Cleaning and Spacing Fixes:** Excessive repetitions of characters or words were

[2] https://pypi.org/project/emoji/
[3] https://github.com/cbaziotis/ekphrasis

cleaned, appropriate spaces were ensured after punctuation, and only a single `<user>` token was maintained if multiple were present, as they were redundant.

10. **Tokenization with spaCy:** Tweets were tokenized using spaCy's language-specific models (`en_core_web_sm` for English and `es_core_news_sm` for Spanish).
11. **Hard Label Processing:** Hard labels were converted into multi-hot vectors in a column indicating the presence of each sexism category.
12. **Sentiment Analysis:** For the English tweets, sentiment scores were computed using the VADER sentiment analyzer[4] [21]. Since VADER does not support Spanish, TextBlob[5] was used to derive polarity scores for the Spanish tweets. Based on these scores, each tweet was classified as Positive, Neutral, or Negative.
13. **Dataset Analysis (Optional):** Dataset-level statistics were analyzed, including text length distributions, language distributions, and label distributions. The visualization plots for sentiment and label distributions were generated and saved as PNG files.
14. **Data Merging (Optional):** If required, the training and development sets were combined into a large training set for training without evaluation.
15. **Saving Final Outputs:** The pre-processed sets were exported as CSV files to be used for training, evaluation and prediction.

Table 2
Number of training, development and test texts in both languages, per language, before and after data cleaning.

	Before Cleaning			**After Cleaning**		
	All	English	Spanish	All	English	Spanish
# Training Texts	6920	3260	3660	6050	2864	3186
# Development Texts	1038	489	549	931	444	487
# Test Texts	2976	978	1098	-	-	-

4. System Overview

4.1. Experimental Setup

The multilingual nature, as well as the source of the data, led to the decision to leverage open-source Transformer-based multilingual models from Hugging Face. These models are considered effective in capturing language-specific nuances and contexts, resulting in achieving more accurate and reliable results, as they have been pre-trained in various languages. More specifically, the large version of XLM-RoBERTa from Facebook[6] [22], the multilingual version of DeBERTa from Microsoft[7] [23], the multilingual XLM-R, which was re-trained on over 1 billion tweets from various languages until December 2022 from Cardiff NLP group at Cardiff University[8] [24] [25] as well as the multilingual version of BERT from Google[9] [26] were employed for baseline experiments. To achieve baseline scores, they were fine-tuned for multi-label text classification using *AutoModelForSequenceClassification*.

The initial round of baseline experiments was conducted using only the twitter-xlm-roberta-large-2022 model with various loss functions to find the best option for this multi-label task and to assess the need for positive weights (See section 4.5). This model was chosen because of its relevant training data and multilingual capabilities. The second round of baseline experiments included fine-tuning

[4] https://github.com/cjhutto/vaderSentiment
[5] https://github.com/sloria/TextBlob
[6] https://huggingface.co/FacebookAI/xlm-roberta-large
[7] https://huggingface.co/microsoft/mdeberta-v3-base
[8] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-large-2022
[9] https://huggingface.co/google-bert/bert-base-multilingual-uncased

the 4 aforementioned models in both languages and each language separately, with the best loss function revealed from the first round of baseline experiments (Distribution-Balanced Loss). All baseline experiments were carried out using the Transformers and Hugging Face libraries, in conjunction with 1 NVIDIA TITAN RTX GPU card with 24GB VRAM. Table 7 in Appendix A, demonstrates that the twitter-xlm-roberta-large-2022 attains the highest scores across all metrics in both languages and Spanish separately. Although the xlm-roberta-large achieved a higher ICM score in the English-only subset, Spanish remains the predominant language in the training and development data. Consequently, the twitter-xlm-roberta-large-2022 model was chosen as the best foundation for conducting further experiments in both languages.

Inspired by the multilingual multi-head model architecture for multi-label text classification developed by the Hierocles of Alexandria team during Touché at CLEF 2024 [27], a more advanced version of the model was created. This new architecture, which is presented in detail in the following section 4.2, integrates multi-tasking capabilities by incorporating sentiment analysis in addition to sexism categorization. This improvement aims not only to tackle the linguistic challenges inherent in each language, but also to enhance the model's ability to understand, detect and categorize sexism by adding the sentiment expressed in each text as additional information.

4.2. Multi-head Multi-task Model Architecture

The developed architecture is depicted in Figure 2 and was designed for multi-task and multilingual classification, targeting both multi-label sexism categorization and multi-class sentiment analysis (positive, neutral, negative) in English and Spanish tweets. At its foundation, the model leverages a pre-trained XLM-RoBERTa encoder to obtain deep multilingual representations. On top of the shared backbone, the architecture integrates language-specific classification heads for both tasks to address each language's linguistic challenges. More particulary, it includes 2 sexism classification heads (one for English and one for Spanish) and 2 sentiment classification heads (one for English and one for Spanish), allowing the model to learn language-adapted representations while benefiting from shared multilingual knowledge. This multi-head architecture allows for cross-lingual flexibility while maintaining specialization where needed.

Each classification head was implemented as a custom Transformer stack, which optionally consists of 1 to 4 Transformer layers depending upon the depth of the ablation experiments. These Transformer layers replicate components of the base architecture and are composed of:

- A self-attention mechanism for contextual token interactions.
- Layer normalization to stabilize and accelerate training.
- Feed-forward sub-layers introducing non-linearity and complexity.
- Residual connections for mitigating the vanishing gradient problem and allowing deeper networks.
- Dropout for preventing overfitting.

Following processing by the classification heads, sentence-level representations can be obtained using one of three available pooling methods. The option to select from these pooling methods was inspired by the author's previously proposed system for detecting signs of depression in social media texts, which was presented in Task 4 of LT-EDI@RANLP 2023 [18]. The pooling methods are the following:

- CLS pooling: returns the representation of the first token in the sequence (i.e., the [CLS] token).
- Mean pooling: returns the mean value of token embeddings, weighted by the attention mask.
- Max pooling: returns the maximum value across the token dimensions while using the attention mask.

For sexism classification, each language-specific head receives the pooled sequence representation from that language and the logits from the sentiment head corresponding to that language. This creates a task-aware input and enables support for auxiliary learning. The additional sentiment information is subsequently concatenated to the pooled representation before classification, providing contextual information related to the sentiment of the tweet, which may assist in identifying overly subtle sexist expressions. The proposed multi-task training process is as follows:

1. The input batch is passed through the shared XLM-RoBERTa encoder to generate contextualized token embeddings.
2. Based on language identifiers in the batch (en, es), the model dynamically routes each instance to the language-specific sentiment and sexism heads.
3. The sentiment heads for each language first generate sentiment logits, which are then used as auxiliary features for the corresponding sexism heads.
4. The sexism heads for each language produce sexism logits using the sentiment logits as auxiliary contextual input.
5. Both the sexism logits (multi-label) and sentiment logits (multi-class) are aggregated across the batch and passed as input to the loss function.
6. A single, joint loss function is used to optimize the model. The joint loss function enables the 2 tasks to learn jointly and leverage both shared representations and task-specific knowledge.

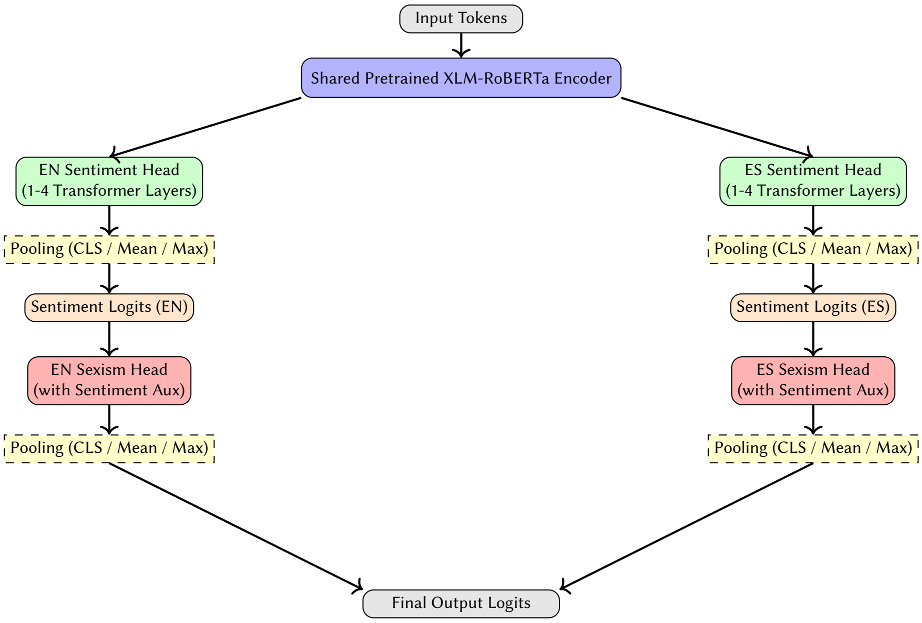
This architecture is particularly effective in the cross-lingual setting, as it allows the model to share knowledge via the base encoder while maintaining per-language specialization in the task heads. Furthermore, the integration of flexible pooling methods and adjustable head depth enables experimentation and comparison to find the optimal configuration/-s for this task. Notably, adding sentiment predictions as contextual auxiliary inputs to the sexism classification heads facilitates the model's ability to differentiate between positive, neutral and negative content, particularly in ambiguous cases where linguistic cues alone may be insufficient.

The best-performing model from the baseline experiments (See section 4.1), twitter-xlm-roberta-large-2022, was leveraged as the foundation for this advanced model architecture. 3 multi-head, multi-task models were developed, utilizing each of the 3 pooling methods, and starting with 1 Transformer layer in the classification head for each language and task. To assess performance improvements, additional Transformer layers were incorporated into the classification heads, resulting in a more complex architecture. Consequently, models comprising 2, 3, and 4 layers were created. In total, 12 models were trained using the provided training set and evaluated with the development set. Their results are illustrated in Table 3. The development and test predictions of these 12 models will be later leveraged for majority ensemble learning (See section 4.7), along with the predictions of (1) an LLM using prompt engineering, (2) an LLM using prompt engineering plus RAG (See section 4.3). Additionally, these 12 models will be trained on the entire provided dataset, training and development combined, with no validation during training. Their test predictions will be combined for another majority ensemble learning. All experiments with this model architecture were also conducted using the Transformers and Hugging Face libraries, alongside 1 NVIDIA TITAN RTX GPU card, which features 24GB of VRAM. The hyperparameters used for both the baselines and the multi-head, multi-task models can be found in Table 8 in Appendix A.

4.3. LLM Prompt Engineering & RAG

In a previously proposed system by the author for identifying hate speech, the targets and stance of hate speech within the *Climate Activism Stance and Hate Event Detection* Shared Task at CASE 2024, an LLM was fine-tuned using Parameter Efficient Fine-Tuning (PEFT), specifically employing Low-Rank Adaptation (LoRA) and prompt tuning. The results demonstrated that using prompting for classification yielded superior results compared to LoRA [19]. For this reason, the state-of-the-art, open-source Llama-3.2-3B-Instruct developed by Meta[10] [28], which supports English and Spanish, was leveraged in 2 Python scripts to classify tweets into hierarchical sexism categories and to generate predictions for both the development and test sets. This approach allows for a comparison of the performance of an LLM with that of multi-head, multi-task Transformer models.

[10] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct



Input Tokens
Shared Pretrained XLM-RoBERTa Encoder
EN Sentiment Head (1-4 Transformer Layers)
ES Sentiment Head (1-4 Transformer Layers)
Pooling (CLS / Mean / Max)
Pooling (CLS / Mean / Max)
Sentiment Logits (EN)
Sentiment Logits (ES)
EN Sexism Head (with Sentiment Aux)
ES Sexism Head (with Sentiment Aux)
Pooling (CLS / Mean / Max)
Pooling (CLS / Mean / Max)
Final Output Logits

Figure 2: Detailed architecture of the multi-task multi-head model. It performs multilingual classification for both multi-label sexism categorization and multi-class sentiment analysis. The shared backbone (XLM-RoBERTa encoder) feeds language-specific heads for English (EN) and Spanish (ES), each composed of optional stacked Transformer layers. Within each head, pooling strategies (CLS, Mean, Max) generate sentence-level representations before final classification. Sentiment logits are concatenated as auxiliary inputs to the sexism heads.

The first script used the LLM with prompt engineering. The prompt incorporated a pre-processed tweet in either English or Spanish, along with its sentiment derived during the pre-processing phase. It also included the sexism categories and their definitions as provided by the EXIST organizers. The instructions specified that the output should be a JSON containing all applicable labels; if no sexism categories applied, the script was to return "NO". Additionally, the sentiment was to inform the decision-making process, and no extra explanation or commentary was to be included in the output. The prompt is shown in Figure 3 in Appendix A. After experimentation with the prompt, the best results achieved an ICM = **-1.4709**, ICM-NORM = **0.1723**, and F_1 = **0.4147** on the development set in both languages.

The second script combined LLM prompt engineering and Retrieval Augmented Generation (RAG). RAG was employed as it provides external knowledge from the most relevant texts and minimizes LLM hallucinations. It was used to provide context to assist the LLM in performing classification more effectively [29]. The prompt incorporated a pre-processed tweet in English or Spanish, along with its sentiment derived during the pre-processing phase, the sexism categories and their definitions, a chain-of-thought reasoning process, as well as a context block retrieved from the 3 most topically similar labeled tweets, fetched through similarity search using ChromaDB[11] and multilingual Sentence-Transformers embeddings. These tweets were sourced from the training set when generating predictions for the development set and from the combined training and development set when generating predictions for the test set. The prompt also contained an annotator profile for the input tweet, synthesized from demographic metadata provided in the EXIST dataset formatted as: "female/male, {age} years old, {ethnicity}, {education}, living in {country}". The inclusion of annotator profiles was inspired by the ABCD team's approach from the previous year [14]. The instructions directed the model to adopt the perspective of the provided annotator profile and to think step-by-step, evaluate multiple sexism categories independently, incorporate sentiment and contextual examples into its decision-making, and return only the final JSON-formatted label list without any extra explanation or commentary. The prompt is shown in Figure 4 in Appendix A. The chain-of-thought reasoning was added in this prompt after several endeavors, and it was proved that it improved classification performance. After

[11] https://github.com/chroma-core/chroma

experimentation with the prompt, the best results achieved an ICM = **-0.9604**, ICM-NORM = **0.2860**, and F_1 = **0.4488** on the development set in both languages. These results demonstrate that prompt engineering, when combined with chain-of-thought reasoning and RAG, can significantly enhance LLM performance in specialized multi-label classification. This approach enables the LLM to act as an annotator based on a specific profile, grounding its predictions in sentiment and relevant human-labeled examples as contextual support, rather than relying solely on the isolated tweet.

By comparing the performance of the LLM (from both scripts) with the multi-head multi-task Transformer models across both languages, it became evident that the LLM scored significantly lower. Therefore, its predictions were included in the majority vote ensemble rather than being submitted individually. Notably, the inclusion of the LLM in the majority vote ensemble resulted in higher scores on the development set compared to using only the multi-head, multi-task Transformer models. The experiments with both scripts were conducted using the Hugging Face and bitsandbytes library for 4-bit quantization[12] and the LangChain library[13], alongside 2 NVIDIA TITAN RTX GPU cards, featuring each 24GB of VRAM.

4.4. Evaluation & Metrics

The hard-hard evaluation protocol is used for systems producing discrete, non-probabilistic outputs. The ground truth annotations produced by multiple human annotators are converted to hard labels using probabilistic thresholds defined for each sub-task. For sub-task 1.3 in this paper, a label is assigned to an instance only if more than one annotator has selected it. This thresholding approach captures only the labels with at least some level of inter-annotator agreement, thus providing a conservative yet reliable approximation of the ground truth.

The official metric of sub-task 1.3 used in the hard-hard evaluation framework is the Information Contrast Model (ICM) [30]. This metric is appropriately tailored towards hierarchical classification problems, as it penalizes classification errors based on the semantic distance between predicted and true labels, thereby taking into account the relationships among classes. In addition to ICM, the evaluation framework also reports the normalized version known as ICM-NORM, which enhances comparability across tasks, as well as the F-measure (F_1) implemented in the PyEvALL evaluation library[14]. Although the F_1 score (the harmonic mean of precision and recall) is reported for reference, it is not considered ideal for this context due to its limitation in capturing class relationships, as it treats errors between conceptually distant classes (e.g., predicting "NO" instead of a specific sexist category) equivalently to errors between similar positive classes (e.g., confusing two sexist categories), despite the former being far more severe.

4.5. Loss Functions & Class Weights

Several loss functions were evaluated during the baseline experiments, including Binary Cross-Entropy Loss with Logits[15], Focal Loss [31], Class-balanced Loss [32], Distribution-Balanced Loss [33], and Class-balanced Negative Tolerant Regularization Loss [34]. These were implemented by modifying the *Trainer* class from Hugging Face to identify the most effective loss function for the task. These loss functions were specifically chosen for their effectiveness in addressing data imbalance challenges in multi-label contexts. They have previously been employed for the *Human Value Detection* shared task by the PAI team at SemEval-2023 [35], as well as by the Hierocles of Alexandria team during Touché at CLEF 2024 [27].

Positive weights were calculated for each class and applied exclusively in experiments using the Binary Cross-Entropy Loss with Logits to enhance model performance in under-represented areas. Baseline experiments using the twitter-xlm-roberta-large-2022 model with the standard classification

[12] https://huggingface.co/docs/bitsandbytes/main/index
[13] https://github.com/langchain-ai/langchain
[14] https://github.com/UNEDLENAR/PyEvALL
[15] https://pytorch.org/docs/stable/generated/torch.nn.functional.binary_cross_entropy_with_logits.html

head (*AutoModelForSequenceClassification*) indicated that the Distribution-Balanced Loss yielded the most favorable results in terms of ICM, ICM-NORM, and F_1 scores, as detailed in Table 6 in Appendix A.

Distribution-Balanced Loss [33] addresses label imbalance and co-occurrence in multi-label classification. It employs instance-level reweighting to adjust the contribution of each label based on its inverse frequency in the dataset, while considering the number of positive labels in each instance. This approach helps mitigate overfitting to dominant classes and retains the learning signal for rare labels. Additionally, it applies a negative-tolerant regularization term to prevent overconfidence with negative labels in cases where most labels are negative, avoiding suppression of rare labels. Consequently, this loss function was selected for all submitted Transformer models without applying class positive weights.

4.6. Thresholds

All Transformer experiments, in both baselines and multi-head multi-task models, were carried out using various thresholds between 0.1 and 0.95. During fine-tuning, both the ICM and F1 scores were adapted to the *compute_metrics* function based on the evaluation package PyEvALL in Python, and were constantly monitored. A general ICM and F_1 score for all classes as well as ICM and F_1 scores per class were calculated in the *compute_metrics* function. For predictions post-training, the best thresholds for each class were saved in a JSON file in order to calculate metrics and create predictions for the development set. After implementing the sigmoid function, predictions were converted to 1 if they met or exceeded the threshold and to 0 if they fell below it. Consequently, 3 distinct prediction files were generated: one based on the default threshold of 0.5, another using the best general threshold across all classes, and a third employing the best threshold for each class. The development set predictions created using the optimal class-specific thresholds achieved the highest scores across all metrics. Applying the best threshold per class was also verified during last year's winning participation in sub-task 1 of Touché at CLEF 2024 [27]. Thus, all predictions submitted for the test set were generated based on the best threshold for each class.

4.7. Majority Ensemble Learning

A majority voting ensemble strategy of multiple multi-head multi-task Transformer-based models with different pooling methods and number of layers was employed to enhance the robustness and generalizability of the final development and test predictions. The idea was that the individual strengths and weaknesses of these models would complement one another when addressing the sexism categorization multi-label hierarchical task. The ensemble consisted of predictions from 12 runs of the twitter-xlm-roberta-large-2022 multi-head multi-task model, each defined by a specified pooling method (CLS, max, mean) and a classification head layer count ranging from 1 to 4. The ensemble also incorporated predictions from Llama-3.2-3B-Instruct in 2 variations: (i) prompt-only using category definitions, and (ii) prompt with RAG utilizing both category definitions and examples from the training set as contextual guidance.

The majority voting mechanism operated on both an instance and a label level. If a label was predicted in 7 out of 13 predictions, it was included in the final output. This threshold of 7 out of 13 was established as a simple majority, meaning that a label had to be predicted by more than half of the models to be included in the final results. This was a fair compromise in balancing accuracy and recall. The "NO" label was included if no sub-category reached this threshold; in such instances, "NO" was enforced to maintain the hierarchical nature of the task. Overall, the ensemble process effectively leveraged the complementary strengths of various Transformer model predictions and LLM-based predictions, resulting in more balanced and robust classification outputs.

Results produced from the provided development set, evaluated via the PyEvALL framework, indicated that the ensemble of the 12 transformer models along with the prompt-only Llama achieved the highest performance metrics: ICM = **0.5352**, ICM-NORM = **0.6192**, and F_1 = **0.6858**. When the prompt-only Llama-3.2-3B-Instruct was substituted with the combined prompt and RAG-based version, the performance metrics were slightly lower but still comparable: ICM = **0.5325**, ICM-NORM = **0.6186**, and

F_1 = **0.6852**. These findings highlight the effectiveness of the majority voting strategy in aggregating predictions from diverse models rather than relying on individual ones, while ensuring high-quality classifications from both Transformer-based models and the LLM-based approach. The majority voting initiative was inspired by the author's previous work on sexism detection at Task 10 of SemEval-2023 [17], where aggregating model predictions was shown to increase performance.

The 3 runs submitted for the test set predictions, hard-hard evaluation for sub-task 1.3 are as follows:

- **Run 1:** Majority vote ensemble of 12 multi-head multi-task models trained on the provided training and development sets combined, CLS, mean and max pooling, 1-4 Transformer layers per classification head, along with the prompt-only Llama-3.2-3B-Instruct.
- **Run 2:** Majority vote ensemble of 12 multi-head multi-task models trained on the provided training set and evaluated on the development set (See Table 3, CLS, mean and max pooling, 1-4 Transformer layers per classification head, along with the prompt-only Llama-3.2-3B-Instruct.
- **Run 3:** Majority vote ensemble of 12 multi-head multi-task models trained on the provided training and development sets combined, CLS, mean and max pooling, 1-4 Transformer layers per classification head, along with the combined prompt and RAG-based version of Llama-3.2-3B-Instruct with the training and development sets combined as RAG documents.

5. Results & Discussion

5.1. Development Set

From Table 3, it is evident that the twitter-xlm-roberta-large-2022 utilizing max pooling achieved the highest scores in the sentiment and sexism classification heads with 1 and 4 Transformer layers. Conversely, the model employing mean pooling attained its peak performance with 2 Transformer layers. The model applying CLS pooling outperformed others in the classification heads with 3 stacked Transformer layers. Notably, adding 2 and 4 layers in the heads resulted in a decline in performance, whereas 3 layers yielded lower yet significant scores compared to 1 layer per head. Despite the variations in scores, all of these models were incorporated in the majority ensemble.

According to Table 4, which illustrates the F_1 scores for each class of the multi-head multi-task cardiffnlp/twitter-xlm-roberta-large-2022 models based on various pooling methods (CLS, mean, max) and configurations with 1 to 4 Transformer layers per classification head, the "NO" class consistently achieves the highest F_1 scores, ranging from 0.86 to 0.87, across all model configurations. This reflects the model's ability to successfully detect non-sexist content, primarily due to the dominance of the "NO" label in the class distribution. Regarding the sexist categories, the "STEREOTYPING-DOMINANCE" and "IDEOLOGICAL-INEQUALITY" classes achieve moderately high F_1 scores between approximately 0.63 and 0.71, showing that the models are effective in capturing concepts of inequality between men and women and identifying mentions of stereotypical roles assigned to women, particularly when utilizing CLS and max pooling for "STEREOTYPING-DOMINANCE" and mean pooling for "IDEOLOGICAL-INEQUALITY", regardless of the number of layers per classification head. The category "OBJECTIFICATION" falls within the moderate range, achieving scores between 0.59 and 0.63. This suggests a balanced challenge arising from the need to navigate both explicit objectifying language and more subtle instances of objectification towards women. Nevertheless, the rest of the sexism categories mainly tend to achieve lower scores since they are the least represented in the dataset. The "SEXUAL-VIOLENCE" class attains a lower F_1 performance, ranging from 0.60 to 0.65, showing the difficulty of the models to capture sexual violence or sexual harassment indicators. Finally, the "MISOGYNY-NON-SEXUAL-VIOLENCE" class consistently produces the lowest F_1 scores across all model configurations, mostly between 0.52 and 0.56, suggesting that the models struggle to detect hatred or non-sexual violence expressions towards women.

Table 3

Performance of multi-head, multi-task models on the development set using `cardiffnlp/twitter-xlm-roberta-large-2022` as foundation trained with Distribution-Balanced Loss in both languages using the 3 pooling methods: CLS, Mean and Max pooling. They were trained on the provided training set and evaluated on the provided development set. Results are grouped by the number of Transformer layers per classification head. The results were computed using the PyEvALL evaluation library.

	All – 1 Transformer Layer Per Classification Head		
Model	**ICM**	**ICM-NORM**	**F$_1$**
CLS Pooling	0.4367	0.5973	0.6638
Mean Pooling	0.4395	0.5979	**0.6671**
Max Pooling	**0.4498**	**0.6002**	0.6664
	All – 2 Transformer Layers Per Classification Head		
Model	**ICM**	**ICM-NORM**	**F$_1$**
CLS Pooling	0.3852	0.5858	0.6540
Mean Pooling	**0.4031**	**0.5898**	**0.6607**
Max Pooling	0.3854	0.5858	0.6608
	All – 3 Transformer Layers Per Classification Head		
Model	**ICM**	**ICM-NORM**	**F$_1$**
CLS Pooling	**0.4352**	**0.5969**	**0.6690**
Mean Pooling	0.4252	0.5947	0.6660
Max Pooling	0.4229	0.5942	0.6679
	All – 4 Transformer Layers Per Classification Head		
Model	**ICM**	**ICM-NORM**	**F$_1$**
CLS Pooling	0.3208	0.5714	0.6556
Mean Pooling	0.3245	0.5723	0.6522
Max Pooling	**0.4154**	**0.5925**	**0.6576**

Table 4

Per-class F$_1$ scores of multi-head, multi-task models on the development set using `cardiffnlp/twitter-xlm-roberta-large-2022` as foundation trained with Distribution-Balanced Loss in both languages using the 3 pooling methods: CLS, Mean and Max pooling. They were trained on the provided training set and evaluated on the provided development set. Results are grouped by the number of Transformer layers per classification head. The results were computed using the PyEvALL evaluation library.

			All – 1 Transformer Layer Per Classification Head			
Model	**NO**	**IDEOLOGICAL-INEQUALITY**	**STEREOTYPING-DOMINANCE**	**OBJECTIFICATION**	**SEXUAL-VIOLENCE**	**MISOGYNY-NON-SEXUAL-VIOLENCE**
CLS Pooling	**0.8725**	0.6216	0.7067	0.6176	0.6091	0.5556
Mean Pooling	0.8649	**0.6437**	0.6743	**0.6303**	0.6195	**0.5697**
Max Pooling	0.8676	0.6235	**0.7072**	0.6236	**0.6375**	0.5389
			All – 2 Transformer Layers Per Classification Head			
CLS Pooling	0.8669	0.6408	**0.6901**	0.6112	0.6068	0.5079
Mean Pooling	0.8647	**0.6495**	0.6851	0.6055	0.6204	0.5393
Max Pooling	**0.8671**	0.6307	0.6802	**0.6140**	**0.6226**	**0.5503**
			All – 3 Transformer Layers Per Classification Head			
CLS Pooling	**0.8706**	0.6307	**0.6982**	**0.6319**	**0.6429**	0.5399
Mean Pooling	0.8560	**0.6614**	0.6932	0.6068	0.6364	**0.5424**
Max Pooling	0.8691	0.6527	0.6950	0.6034	0.6517	0.5351
			All – 4 Transformer Layers Per Classification Head			
CLS Pooling	0.8625	0.6345	0.6986	0.5927	**0.6203**	**0.5253**
Mean Pooling	0.8644	**0.6396**	0.6879	0.5888	0.6074	0.5252
Max Pooling	**0.8652**	0.6125	**0.7061**	**0.6313**	0.6066	0.5240

5.2. Test Set

From Table 5, it is illustrated that all submitted runs (See section 4.7 for more details) outperformed the two baselines in both languages, as well as individually for each language. The baselines include the *EXIST2025-test_majority-class*, which categorizes all instances as the majority class, and the *EXIST2025-test_minority-class*, which categorizes all instances as the minority class. It was shown that training the models with a larger dataset by combining both the training and development sets (as in Runs 1 and 3) resulted in improved performance. In contrast, Run 2, which relied solely on the training set for training, achieved the lowest rank among the three runs across both languages and within each language separately. In both languages (denoted as *ALL*), Run 3 achieved the 4th rank with ICM = **0.4842**, ICM-NORM = **0.6124**, and F$_1$ = **0.6335**, exceeding the submission scores of the ABCD team

Table 5
Hard-hard evaluation results of majority ensemble submitted runs on the unlabeled test set. The results were computed using the PyEvALL evaluation library.

		All		
Model	**Rank**	**ICM**	**ICM-NORM**	**F_1**
EXIST2025-test_gold	0	2.1533	1.0000	1.0000
Run 1	5	0.4814	0.6118	0.6324
Run 2	6	0.4515	0.6048	0.6272
Run 3	**4**	**0.4842**	**0.6124**	**0.6335**
EXIST2025-test_majority-class	108	-1.5984	0.1289	0.1069
EXIST2025-test_minority-class	128	-3.1295	0.0000	0.0288
		English		
Model	**Rank**	**ICM**	**ICM-NORM**	**F_1**
EXIST2025-test_gold	0	2.0402	1.0000	1.0000
Run 1	**5**	**0.3932**	**0.5964**	**0.6118**
Run 2	7	0.3847	0.5943	0.6159
Run 3	6	0.3928	0.5963	0.6108
EXIST2025-test_majority-class	105	-1.4563	0.1431	0.1111
EXIST2025-test_minority-class	127	-2.9279	0.0000	0.0301
		Spanish		
Model	**Rank**	**ICM**	**ICM-NORM**	**F_1**
EXIST2025-test_gold	0	2.2393	1.0000	1.0000
Run 1	5	0.5472	0.6222	0.6474
Run 2	6	0.4946	0.6104	0.6357
Run 3	**4**	**0.5536**	**0.6236**	**0.6501**
EXIST2025-test_majority-class	108	-1.7269	0.1144	0.1030
EXIST2025-test_minority-class	125	-3.3196	0.0000	0.0276

from last year's EXIST competition, which secured the first rank [14]. This performance highlights that the inclusion of chain-of-thought reasoning, along with the annotators' profiles in the LLM prompt and the use of RAG to provide relevant contextual information from tweets, contributes to improved results by reducing hallucinations. Run 3 maintained the same rank when evaluated solely in Spanish, possibly due to the over-representation of this language in the dataset. However, it ranked slightly lower than Run 1, which secured the 6^{th} rank in English alone, with ICM = **0.3928**, ICM-NORM = **0.5963**, and F_1 = **0.6108**. Run 1 achieved the 5^{th} rank across both languages as well as in each language separately and recorded the highest score among the submitted runs for English only, with ICM = **0.3932**, ICM-NORM = **0.5964**, and F_1 = **0.6118**. This indicates that incorporating category definitions and considering the sentiment of the input tweets can significantly enhance performance, particularly when the tweets are exclusively in English.

6. Conclusion & Future Work

The system developed for the EXIST Lab at CLEF 2025 participated in sub-task 1.3, focusing on tackling sexism categorization in the tweet partition in hard evaluation through hierarchical multi-label text classification. The submitted system involved fine-tuning the multilingual twitter-xlm-roberta-large-2022 Transformer model 12 times, utilizing a multi-head and multi-task model architecture that facilitates both sentiment analysis and sexism categorization. Variations in these 12 runs included CLS, mean and max pooling, as well as adjustments to the number of layers in the classification heads. Additionally, the system leveraged the multilingual Llama-3.2-3B-Instruct model and explored 2 methods: one implementing classification via prompt engineering and the other combining classification with prompt engineering and RAG to incorporate contextual information for improved performance. The predictions from the 12 multi-head multi-task Transformer models and the LLM were aggregated using majority ensemble learning, resulting in 3 submissions for hard-hard evaluation on the unlabeled test set (See section 4.7). The experimental strategy employed a multi-step pre-processing pipeline, appropriate loss functions for multi-label classification with imbalanced classes, positive class weights, and varying thresholds per class, all aimed at alleviating class imbalance and enhancing the models' ability to comprehend and classify texts more effectively.

The submissions achieved commendable rankings, securing the 4^{th}, 5^{th}, and 6^{th} positions out of a total of 132 submissions in both English and Spanish. All submitted runs surpassed the baseline scores across both languages and each language separately. Run 3 and Run 1, which combined the training and development sets for training, outperformed Run 2, which solely utilized the training set for training. It was revealed that integrating contextual information through RAG and employing chain-of-thought reasoning contributed significantly to successful LLM classification. Notably, Run 3, as the majority ensemble approach with the 12 Transformer models along with RAG, chain-of-thought and annotators' profiles included in the prompt, yielded the highest score and ranked 4^{th} in both languages. This method also achieved a 4^{th} rank in Spanish, consistent with the other two runs, which maintained the same rank for both languages. This pattern may be attributed to the minor over-representation of Spanish within the EXIST dataset. When focusing solely on English, Run 1 attained a slightly higher score than Run 3, ranking 5^{th}, while Run 3 and Run 2 ranked 6^{th} and 7^{th}, respectively.

To optimize model performance for multi-label sexism categorization, future research should explore the use of larger Transformer language models within the multi-task multi-head architecture, such as XLM-RoBERTa-xl[16] or XLM-RoBERTa-xxl[17]. Additionally, experimenting with larger LLMs for classification and refining the prompt with chain-of-thought reasoning could yield significant improvements. Incorporating data augmentation techniques or synthetic data generation may further enhance overall performance.

7. Limitations

The data analysis and baseline experiments revealed a significant issue of class imbalance among the labels. Despite the efforts to utilize loss functions specifically designed to address class imbalance and to explore various thresholds and class-positive weights for each label, detecting instances of sexism while simultaneously identifying one or more categories of sexism remains a challenging task. This difficulty primarily arises from the overwhelming prevalence of "NO" sexism annotations in the training and development datasets, along with an inadequate distribution of sexism categories, particularly concerning labels such as "SEXUAL-VIOLENCE" and "MISOGYNY-NON-SEXUAL-VIOLENCE". Additionally, the disproportionate number of Spanish tweets compared to English tweets poses additional challenges for models due to the smaller token count available for Spanish compared to English. Compounding these issues, limitations in GPU VRAM have hindered experimenting with larger Transformer models, such as facebook/xlm-roberta-xl or LLMs with more parameters like Gemma 3.

Acknowledgments

The research that contributed to these results was funded by the European Union's Horizon Europe research and innovation programme, in the context of the TITAN project, grant agreement No. 101070658, and the AI4TRUST project, grant agreement No. 101070190. This research has also been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. The views expressed in this paper are exclusively those of the author, and the European Commission is not responsible for any use that may be made of the information it contains.

Declaration on Generative AI

During the preparation of this work, the author used Grammarly for grammar and spelling checks. After using this tool/service, the author reviewed and edited the content as needed and takes full

[16] https://huggingface.co/facebook/xlm-roberta-xl
[17] https://huggingface.co/facebook/xlm-roberta-xxl

responsibility for the publication's content.

References

[1] G. Masequesmay, Sexism, 2022. URL: https://www.britannica.com/topic/sexism.

[2] M. F. Wright, B. D. Harper, S. Wachs, The associations between cyberbullying and callous-unemotional traits among adolescents: The moderating effect of online disinhibition, Personality and Individual Differences 140 (2019) 41–45. URL: https://www.sciencedirect.com/science/article/pii/S0191886918301910. doi:https://doi.org/10.1016/j.paid.2018.04.001, personality pathologies in the world: beyond dichotomies.

[3] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, Computers in Human Behavior 52 (2015) 436–442. URL: https://www.sciencedirect.com/science/article/pii/S0747563215004641. doi:https://doi.org/10.1016/j.chb.2015.06.02.

[4] L. d. Meco, A. MacKay, Social media, violence and gender norms: The need for a new digital social contract, 2022. URL: https://www.alignplatform.org/resources/blog/social-media-violence-and-gender-norms-need-new-digital-social-contract.

[5] J. Valenti, Toxic twitter - women's experiences of violence and abuse on twitter, 2022. URL: https://www.amnesty.org/en/latest/news/2018/03/online-violence-against-women-chapter-3-2/.

[6] G. K. Pitsilis, H. Ramampiaro, H. Langseth, Effective hate-speech detection in twitter data using recurrent neural networks, Applied Intelligence 48 (2018) 4730–4742. URL: https://doi.org/10.1007/s10489-018-1242-y. doi:10.1007/s10489-018-1242-y.

[7] A. Arsht, D. Etcovitch, The human cost of online content moderation, 2018. URL: https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation.

[8] 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.

[9] 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.

[10] F. J. Rodríguez-Sanchez, J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Proces. del Leng. Natural 67 (2021) 195–207. URL: https://api.semanticscholar.org/CorpusID:250527210.

[11] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74/. doi:10.18653/v1/2022.semeval-1.74.

[12] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

[13] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz Garcia, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 — Learning with Disagreement for Sex-

ism Identification and Characterization in Tweets and Memes, 2024, pp. 93–117. doi:10.1007/978-3-031-71908-0_5.

[14] L. M. Quan, D. V. Thin, Sexism identification in social networks with generation-based language models, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, volume 3740 of *CEUR Workshop Proceedings*, 2024. URL: http://ceur-ws.org/Vol-3740/paper-109.pdf.

[15] Y.-Z. Fang, L.-H. Lee, J.-D. Huang, Nycu-nlpat exist 2024: Leveraging transformers with diverse annotations for sexism identification in social networks, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, volume 3740 of *CEUR Workshop Proceedings*, 2024. URL: http://ceur-ws.org/Vol-3740/paper-93.pdf.

[16] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based mixture of transformers for exist2024, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, volume 3740 of *CEUR Workshop Proceedings*, 2024. URL: http://ceur-ws.org/Vol-3740/paper-108.pdf.

[17] C. Christodoulou, NLP_CHRISTINE at SemEval-2023 task 10: Utilizing transformer contextual representations and ensemble learning for sexism detection on social media texts, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 595–602. URL: https://aclanthology.org/2023.semeval-1.81/. doi:10.18653/v1/2023.semeval-1.81.

[18] C. Christodoulou, NLP_CHRISTINE@LT-EDI-2023: RoBERTa & DeBERTa fine-tuning for detecting signs of depression from social media text, in: B. R. Chakravarthi, B. Bharathi, J. Griffith, K. Bali, P. Buitelaar (Eds.), Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 109–116. URL: https://aclanthology.org/2023.ltedi-1.16/.

[19] C. Christodoulou, NLPDame at ClimateActivism 2024: Mistral sequence classification with PEFT for hate speech, targets and stance event detection, in: A. Hürriyetoğlu, H. Tanev, S. Thapa, G. Uludoğan (Eds.), Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 96–104. URL: https://aclanthology.org/2024.case-1.13/.

[20] C. Baziotis, N. Pelekis, C. Doulkeridis, Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754.

[21] C. J. Hutto, E. E. Gilbert, VADER: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, 2014.

[22] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.

[23] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[24] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[25] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260. URL: https://aclanthology.org/2022.acl-demo.25. doi:10.18653/v1/2022.acl-demo.25.

[26] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers

for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[27] S. Legkas, C. Christodoulou, M. Zidianakis, D. Koutrintzes, M. Dagioglou, G. Petasis, Hierocles of alexandria at touché: Multi-task & multi-head custom architecture with transformer-based models for human value detection, in: G. Faggioli, N. Ferro, P. Galuscakova, A. G. S. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 3419–3432. URL: http://ceur-ws.org/Vol-3740/paper-330.pdf.

[28] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, Z. Zhao, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[29] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. URL: https://arxiv.org/abs/2312.10997. arXiv:2312.10997.

[30] E. Amigó, 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), Association for Computational Linguistics, Dublin, Ireland, 2022, p. 5809–5819.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. URL: https://arxiv.org/abs/1708.02002. arXiv:1708.02002.

[32] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, 2019. URL: https://arxiv.org/abs/1901.05555. arXiv:1901.05555.

[33] T. Wu, Q. Huang, Z. Liu, Y. Wang, D. Lin, Distribution-balanced loss for multi-label classification in long-tailed datasets, 2021. URL: https://arxiv.org/abs/2007.09654. arXiv:2007.09654.

[34] Y. Huang, B. Giledereli, A. Köksal, A. Özgür, E. Ozkirimli, Balancing methods for multi-label text classification with long-tailed class distribution, 2021. URL: https://arxiv.org/abs/2109.04712. arXiv:2109.04712.

[35] L. Ma, Z. Sun, J. Jiang, X. Li, PAI at SemEval-2023 task 4: A general multi-label classification system with class-balanced loss function and ensemble module, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 256–261. URL: https://aclanthology.org/2023.semeval-1.34. doi:10.18653/v1/2023.semeval-1.34.

A. Appendix

Table 6
Performance of cardiffnlp/twitter-xlm-roberta-large-2022 on the development set using different loss functions. Baseline experiments were conducted using hard labels and both languages (English and Spanish). The results were computed using the PyEvALL evaluation library.

Loss Function	**ICM**	**ICM-NORM**	**F$_1$**
BCE	0.3870	0.5862	0.6630
BCE + Pos Weights	0.4218	0.5939	**0.6660**
Focal Loss	0.4095	0.5912	0.6608
Class-Balanced Loss	0.3866	0.5861	0.6625
Class-Balanced Negative Tolerant Regularization Loss	0.4144	0.5923	0.6635
Distribution-Balanced Loss	**0.4254**	**0.5948**	0.6654

Table 7
Performance of different Transformer models using the best-performing Distribution-Balanced Loss (DBloss) on the development set. Baseline experiments were conducted using hard labels in both languages and separately per language. The results were computed using the PyEvALL evaluation library.

	All		
Model	**ICM**	**ICM-NORM**	**F$_1$**
FacebookAI/xlm-roberta-large	0.3385	0.5754	0.6505
microsoft/mdeberta-v3-base	0.3277	0.5730	0.6449
cardiffnlp/twitter-xlm-roberta-large-2022	**0.4254**	**0.5948**	**0.6654**
google-bert/bert-base-multilingual-uncased	-0.1538	0.4657	0.5873
	English		
Model	**ICM**	**ICM-NORM**	**F$_1$**
FacebookAI/xlm-roberta-large	**-1.0051**	0.2761	**0.4336**
microsoft/mdeberta-v3-base	-1.1401	0.2461	0.4188
cardiffnlp/twitter-xlm-roberta-large-2022	-1.0058	**0.2760**	0.4301
google-bert/bert-base-multilingual-uncased	-1.0794	0.2596	0.4195
	Spanish		
Model	**ICM**	**ICM-NORM**	**F$_1$**
FacebookAI/xlm-roberta-large	-0.8544	0.3097	0.4556
microsoft/mdeberta-v3-base	-0.9346	0.2918	0.4562
cardiffnlp/twitter-xlm-roberta-large-2022	**-0.8278**	**0.3156**	**0.4840**
google-bert/bert-base-multilingual-uncased	-1.2728	0.2165	0.4028

Table 8
Transformer Models' Hyperparameters used for baseline experiments and all multi-head, multi-task models.

Hyperparameter	**Value**
Seed	2025
Number of Epochs	30
Early Stopping Patience	10
Training Max Sequence Length	145
Development Max Sequence Length	131
Test Max Sequence Length	161
Train Batch Size	32
Validation / Test Batch Size	32
Learning Rate	1e-5
Max grad Norm	1.0
Optimizer	AdamW
AdamW Epsilon	1e-8
LR Scheduler	Linear
Mixed Precision	fp16
Special Tokens	<user>, <email>, <date>, <number>, <phone>

```
[INST]
You are a sexism detection assistant analyzing tweets in English or Spanish. Each tweet is accompanied by its sentiment.

Your task is:
- To detect all applicable **sexism categories** from a predefined list.
- Use the sentiment to inform your judgment (e.g., aggressive tone may signal abuse).
- Consider **each category independently**; multiple categories can apply.
- If no sexism is present, return ONLY ["NO"].

Tweet:
"{tweet}"
Sentiment: {sentiment_prediction}

### Sexism Categories (with definitions):
1. **IDEOLOGICAL-INEQUALITY**: The text discredits the feminist movement, rejects inequality between men and women, or presents men
↪  as victims of gender-based oppression.
2. **STEREOTYPING-DOMINANCE**: The text expresses false ideas about women that suggest they are more suitable to fulfill certain
↪  roles (mother, wife, caregiver, submissive, etc.), or inappropriate for certain tasks (e.g., driving, hard work), or claims that
↪  men are superior.
3. **OBJECTIFICATION**: The text presents women as objects, disregarding their dignity and personality, or assumes physical traits
↪  women must have to fulfill traditional gender roles (beauty standards, hypersexualization, women's bodies at men's disposal,
↪  etc.).
4. **SEXUAL-VIOLENCE**: The text includes or describes sexual suggestions, requests for sexual favors, or harassment of a sexual
↪  nature, including rape or sexual assault.
5. **MISOGYNY-NON-SEXUAL-VIOLENCE**: The text expresses hatred or non-sexual violence toward women (e.g., insults, aggression, or
↪  psychological abuse without sexual undertone).
6. **NO**: Use this only if none of the above categories are present.

### Output Format:
- Return only the following JSON.
- Do not explain your answer.
- Select **all** categories that apply (or only ["NO"] if none apply).

{
  "labels": ["<CATEGORY1>", "<CATEGORY2>", ...]
}

Answer:
[/INST]
```

Figure 3: Prompt template for the first script, the prompt-only version.

```
[INST]
You are a sexism detection assistant analyzing tweets in English or Spanish. You have the perspective of the following person:

{annotator_profile}

Each tweet is accompanied by its sentiment.
You are also given examples of previously labeled tweets to help guide your classification.

Before making a decision, think step-by-step:
- Does the tweet express sexist stereotypes, gender roles, objectification, or violence?
- Does it refer to power, bodies, or inequality?
- Consider sentiment and examples carefully.

Tweet:
"{tweet}"

Sentiment: {sentiment_prediction}

Context examples from labeled tweets:
{retrieved_context}

### Sexism Categories (with definitions):
1. **IDEOLOGICAL-INEQUALITY**: The text discredits the feminist movement, rejects inequality between men and women, or presents men
↪  as victims of gender-based oppression.
2. **STEREOTYPING-DOMINANCE**: The text expresses false ideas about women that suggest they are more suitable to fulfill certain
↪  roles (mother, wife, caregiver, submissive, etc.), or inappropriate for certain tasks (e.g., driving, hard work), or claims that
↪  men are superior.
3. **OBJECTIFICATION**: The text presents women as objects, disregarding their dignity and personality, or assumes physical traits
↪  women must have to fulfill traditional gender roles (beauty standards, hypersexualization, women's bodies at men's disposal,
↪  etc.).
4. **SEXUAL-VIOLENCE**: The text includes or describes sexual suggestions, requests for sexual favors, or harassment of a sexual
↪  nature, including rape or sexual assault.
5. **MISOGYNY-NON-SEXUAL-VIOLENCE**: The text expresses hatred or non-sexual violence toward women (e.g., insults, aggression, or
↪  psychological abuse without sexual undertone).
6. **NO**: Use this only if none of the above categories are present.

### Output Format:
- Return only valid JSON exactly in the format shown below.
- Do not include explanations or extra text.

{
  "labels": ["<CATEGORY1>", "<CATEGORY2>", ...]
}

Answer:
[/INST]
```

Figure 4: Prompt template for the second script, combining prompt engineering with RAG-based context.