Better Together: LLM and Neural Classification Transformers to Detect Sexism

Notebook for the EXIST Lab at CLEF 2024

Judith Tavarez-Rodríguez[1,*,†], Fernando Sánchez-Vega[1,2,†], Alejandro Rosales-Pérez[3,†] and Adrián Pastor López-Monroy[1,†]

[1]Mathematics Research Center (CIMAT), Jalisco S/N Valenciana, 36023, Guanajuato, Guanajuato, México
[2]Consejo Nacional de Ciencia y Tecnología (CONACYT), Av. Insurgentes Sur 1582, Col. Crédito Constructor, 03940, CDMX, México
[3]Mathematics Research Center (CIMAT), Monterrey, Av. Alianza Centro 502, Apodaca, 66628, Nuevo León, México.

Abstract

In this paper the approaches of the CIMAT-CS-NLP team for Task 1 (hard and soft settings) and Task 2 (hard setting) of the EXIST 2024 evaluation forum are presented. Task 1 consists in identifying instances of sexism in tweets (binary classification), while Task 2 is focused on determining the source intention in the sexist tweets (direct, reported and judgemental). The proposed methods for both tasks are based on unifying the knowledge of two different systems: zero-shot classification by using Large Language Models (LLMs) through a prompting refinement process, and supervised fine-tuning multilingual Transformers for classification. Results from both systems are combined by means of various techniques to determine the most effective approach. This methodology aims to leverage the strengths and robustness of different multilingual architectures to enhance classification results. The experimental results indicate that this approach is an effective method for sexism detection and categorization. Our best submitted system for sexism detection achieved third place in the hard-hard evaluation for all tweets, third place for tweets in Spanish and fourth place for tweets in English, with an F1 (positive class) of 0.7899, 0.8148 and 0.7576 respectively.

Keywords

LLMs, Transformers, Online Sexism, Prompt Refinement

1. Introduction

Sexism on social media has become a widespread problem, reflecting and perpetuating social prejudices within digital discourse [1]. The prevalence of gender-based discrimination on various online platforms highlights the urgency of having effective detection and mitigation strategies. The EXIST 2024 evaluation forum [2, 3] (http://nlp.uned.es/exist2024/) at CLEF is a campaign aimed at combating sexism and has been promoting research in its identification and categorization on social networks since 2021. The methods recently reported in EXIST for detecting online sexism primarily involve fine-tuning transformer-type models for classification, such as BERT [4], and integrating them with task-specific features [5]. However, it's worth noting that generative LLMs have recently emerged as powerful tools for language generation and understanding. Such procedures have not been extensively explored and evaluated in the EXIST test dataset until now. Particularly, prompt engineering is a process that involves designing, testing, and iteratively refining prompts to guide the model's responses more effectively. By carefully crafting prompts, the model's ability to understand context and generate relevant outputs can be significantly enhanced [6].

Even though works such as [7] have highlighted the benefits of prompting LLMs over the quantity of labeled data points for supervised fine-tuning, the core concept of our work is not just comparing, but rather leveraging the complementary knowledge encoded in supervised fine-tuned transformer models and in zero-shot settings with larger LLMs. By combining these diverse linguistic representations, the

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
judith.tavarez@cimat.mx (J. Tavarez-Rodríguez); fernando.sanchez@cimat.mx (F. Sánchez-Vega); alejandro.rosales@cimat.mx (A. Rosales-Pérez); pastor.lopez@cimat.mx (A. P. López-Monroy)

aim is to harness the collective intelligence encoded in the models to achieve more accurate sexism detection and categorization in social media text.
In this paper, traditional fine-tuning methods for transformer-based models were evaluated. Moreover, a prompt engineering process was conducted, during which sexism detection improved as the prompts were refined. We tested various unification strategies to combine the outputs from prompt engineering and classical fine-tuning. Our results show that a voting strategy, incorporating responses from seven different systems, was the most effective technique for sexism detection. Notably, four of these seven systems were developed through the prompt refining process. Our investigation suggests that integrating insights from multiple expert systems enhances the detection of sexism in social media.
Our main contributions are:

1. Different strategies for unification of experts have been evaluated and its been shown that some of them improved the results of classical supervised fine-tuning transformer models for classification.
2. It has been found that combining LLMs with Transformer-based classifiers can enhance the consensus in determining whether a tweet exhibits sexist content.
3. We have refined a prompt for zero-shot sexism detection which can be a starting point to devise other techniques such as few-shot, in-context learning, among others.

2. Related Work

Fine-tuning transformers for classification has been a strategy employed since the first edition of the EXIST evaluation forum [8]. It has been approached in several ways, including as monolingual and multilingual ensembles [9], ensembles of ensembles [10], and by finding optimal hyperparameters for over 30 pre-trained transformers available in HuggingFace [11]. This suggests that fine-tuning transformers for classification serves as a promising starting point for addressing sexism identification and categorization.
In the previous edition of the evaluation forum, EXIST 2023 [5], some teams [12, 13] used GPT-2 [14] and GPT-NeoX [15] models, with the latter achieving first place in the hard setting for tasks 1 and 2. Thus, it is promising to explore different and more recent LLMs and techniques for sexism identification and categorization tasks. For example, the authors in [16] conducted experiments on the EXIST 2021 and 2022 datasets, comparing techniques such as fine-tuning transformers for classification, zero-shot learning, and few-shot learning on T5 and Llama models. They evaluated the Spanish and English datasets separately, and their results demonstrated a favorable outcome for supervised fine-tuning on BERT models.
In other related domains, such as hate speech, authors in [17] compare zero-shot classification using T5 and Llama models with a fine-tuned multilingual BERT (mBERT) model on the HatEval dataset [18], which consists on detecting hate speech against immigrants and women in Spanish and English tweets. Their results were favorable for zero-shot classification in English tweets, but for Spanish tweets, supervised learning yielded the best results. Similar experiments and results are shown in [19].
In this work, more than comparing results from zero-shot classification and supervised fine-tuning, we aim at combining these two different approaches for leveraging the knowledge of several expert models for improving sexism detection in the EXIST 2024 dataset. Moreover, to the best of our knowledge, prompt refinement has never been done before for zero-shot sexism identification in the EXIST dataset, using the instructions provided in the annotations guideline [3].

3. Methodology

The systems that we developed for sexism identification have three general components:

- Zero-shot classification through LLMs and prompts
- Supervised fine-tuning of transformers for classification
- Unification of the knowledge from both methods.

Results from zero-shot classification using LLMs and various fine-tuning approaches applied to transformer models for classification have been obtained and analyzed. The aim is to discern the most effective method for unifying the insights from these diverse expertise sources. For this purpose, we seek to optimize the integration of knowledge extracted from the specialized capabilities of each model variant.

In the subsequent, systems for Task 1 hard setting are describe. For the systems of Task 1 with soft setting and Task 2 hard setting, see further details in section 3.3.

3.1. LLMs and Multilingual Transformers for Classification

LLMs. For the responses generated from LLMs, the Gemini API [20] was used through the *google-generativeai* python package. The model used was gemini-1.0-pro with safety settings indicating that no response should be blocked [20]. Subsequently, a prompt engineering processed was devised for performing classification under a zero-shot setting. The process consisted in defining a prompt that allows us to extract a response for the classification task with a regex expression. The prompt was refined until responses of three different prompts were obtained.

simple_prompt: *"Respond YES or NO. Is the following tweet sexist? Tweet: ".*

Then, this prompt was refined using instructions from the annotations guidelines provided in the EXIST 2024 files. As a result, a second prompt was obtained.

class_definitions_prompt: *"Given the following tweet, classify it as either YES or NO according to the instructions provided:\n\n**Instructions:**\n\n* **NO:** The tweet does not prejudice, underestimate, or discriminate against women.\n* **YES:** The tweet is sexist itself, describes a sexist situation, or criticizes sexist behavior.\n\n**Tweet:** ".*

The third prompt was obtained from the second one when asking to ChatGPT [21] to improve it. The resulting prompt was the following:

class_definitions_refined_prompt: *"**Instructions for Classification:**\n- **YES**: Classify the tweet as YES if it exhibits sexism directly, describes a sexist scenario, or criticizes sexist behavior.\n- **NO**: Classify the tweet as NO if it does not show prejudice against, undermine, or discriminate against women.\n\n**Tweet**: "*

Responses of a fourth prompt were obtained as well. This prompt is aimed to probably complement the responses generated from the three previous prompts, asking to the LLM to simulate the role of an expert in sexism. Incorporating a simulated profile of an expert in sexism into the language model is proposed as a method to potentially enhance the zero-shot classification [22].

profiled_simple_prompt: *"You are an expert in sexism and you know how to analyze texts from social media. Tell me if the tweet exhibits sexism directly, describes a sexist scenario, or criticizes sexist behavior. Just answer YES or NO.\n**Tweet:** "*

All tweets from *train*, *dev* and *test* partitions were classified with the LLM and the four different prompts. They were all asked in English although the tweets were in English and Spanish. Responses generated were cleaned for keeping only the YES or NO classification answer. Tweets that generated a blocked response in the Gemini API, were classified as sexist, due to the nature of the blocking (safety settings were modified to avoid such blockages, but policy of the API is to block the response for harmful content [23]).

Multilingual Transformers for Classification. According to evaluations in previous EXIST labs [5], fine-tuning multilingual transformer models for classification, such as XLM-RoBERTa [24], mBERT [4] and Twitter-XLM-Roberta [25], has lead to good performance in the sexism identification task. For this reason, fine-tuning of these three models was performed. The obtained classification results provided a baseline for comparison with the evaluations from last year's EXIST lab.

3.2. Experts Unification

Results from seven different types of evaluations (simple_prompt, class_definitions_prompt, class_definitions_refined_prompt, profiled_simple_prompt, variations of fine-tuned XLM-RoBERTa, variations of fine-tuned (FT) mBERT and variations of fine-tuned Twitter-XLM-R) were obtained. To integrate the knowledge extracted from each variation of the models, three strategies were taken into account:

- Creation of new input for fine-tuning
- Proportion of votes
- Best prompt response or best fine-tuned model

These strategies were the ones submitted for evaluations in the test set, for Task 1 Hard setting. They consisted in the following:

Creation of new input (Resp_aware_in) for fine-tuning. This strategy involved concatenating the tweet with the responses of the LLM generated using various prompts from the refinement process (e.g., Resp_aware_in = Tweet + "YES" + "NO" + "YES"). These new inputs were then passed through a fine-tuning process of a transformer model for classification (see Figure 1).

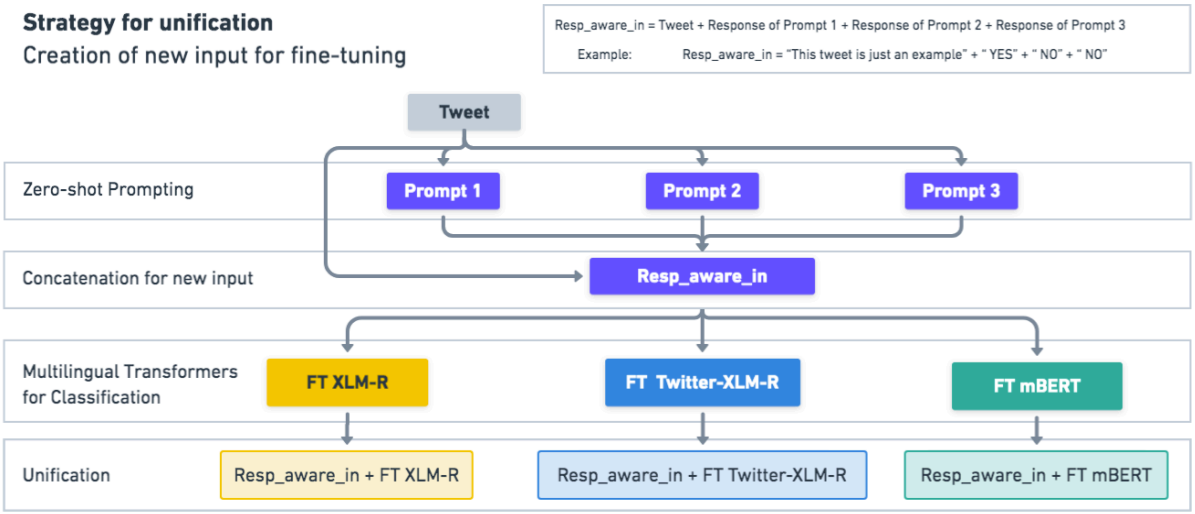

Figure 1: Unification strategy: Creation of new input for fine-tuning.

Proportion of votes. To emulate the "Learning with Disagreement"[26] paradigm present in the dataset annotations, this approach considers all the responses from the seven types of systems, taking into account the proportion of YES and NO answers. A threshold of 0.5 was used to decide whether a tweet was classified as sexist or not. Seven systems were considered to avoid ties (see Figure 2).

Best prompt response or best fine-tuned model. This approach involved creating an ensemble of the best response generated by the LLM and prompts, along with the best fine-tuned transformer for classification. The ensemble was based on a logical OR operation with the binary predictions for Task 1 (see Figure 3).

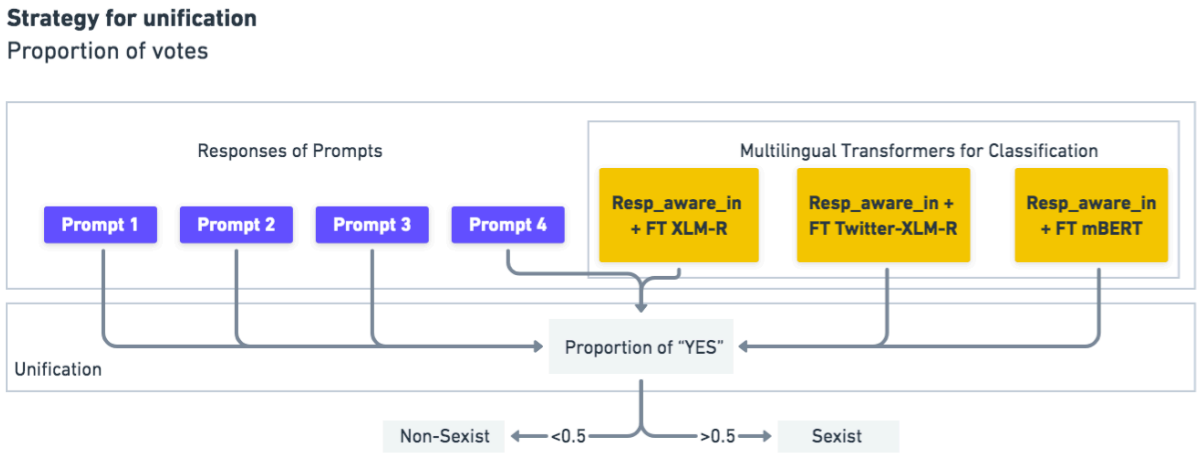

Figure 2: Unification strategy: Proportion of votes.

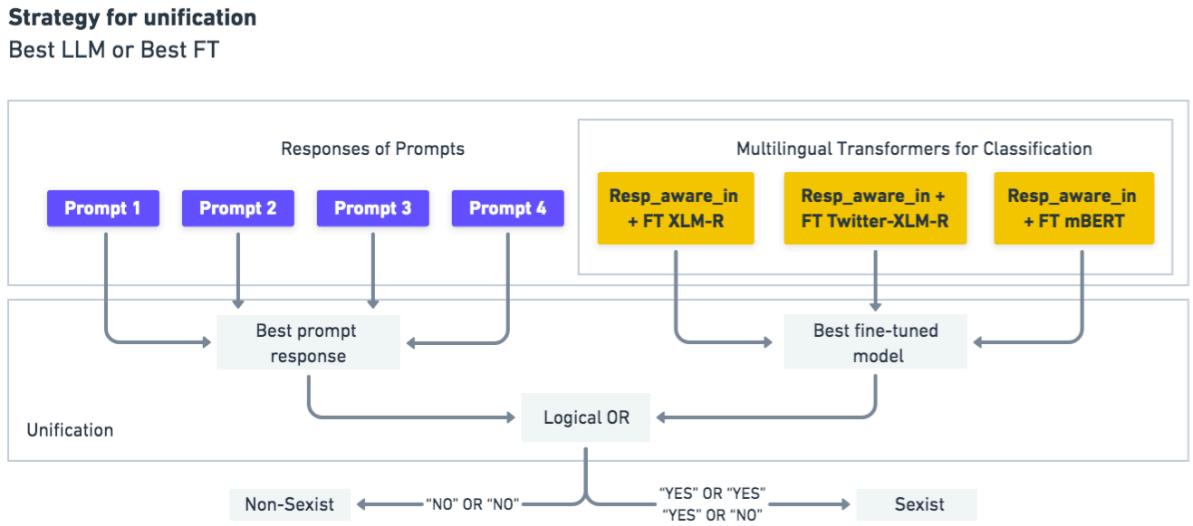

Figure 3: Unification strategy: Best prompt response or best fine-tuned model

3.3. Soft Setting & Source Intention

The previously explained strategies were primarily developed for the hard evaluation in Task 1. For the soft evaluation, a slight modification was made to the *proportion of votes* strategy. Instead of considering the results of all seven evaluations (simple_prompt, class_definitions_prompt, class_definitions_refined_prompt, profiled_simple_prompt, XLM-RoBERTa, mBERT, and Twitter-XLM-R), one of the systems (a different one for each submission) was excluded to only consider the responses of six "experts". This adjustment aimed to emulate the fact that the EXIST 2024 dataset has six different annotations for each tweet.

For Task 2, which involves classification at a finer level (the intention of the author), we employed the strategy of *creating new input for fine-tuning* with models in cascade. Initially, the response of a fine-tuned model for classification in Task 1 was used to identify sexist tweets. Subsequently, a second model was fine-tuned for classification into the three classes of the source intention task: DIRECT, REPORTED, and JUDGEMENTAL.

4. Data Pre-processing and Settings

The tweets from the train, dev, and test partitions of the EXIST 2024 dataset were pre-processed using the *pysentimiento* python library [27]. This involved replacing user mentions and URLs with special tokens, and handling emojis and hashtags. For the hard evaluations, instances where ties occurred in the annotations between sexist and non-sexist labels were identified. In these cases, no golden hard labels were available in the dataset. Therefore, these instances were removed from the training and evaluations for the development set.
The Gemini API was used with the default settings, except for the safety settings, which were modified to ensure that no harmful content would be blocked. If any blocking still occurs, the tweet is classified as sexist due to the nature of the API's blocking policies [23].
For all the fine-tuning classification experiments, the parameters set were: seed = 68, learning rate = $1e-5$, batch size = 8, number of epochs = 5, and maximum input length = 250 (to cover all the tweets and the concatenation of the prompt responses). The optimizer used was AdamW, and the loss function was Cross Entropy. The models used in the experiments were Twitter-XLM-RoBERTa-base, XLM-RoBERTa-large, and BERT-base-multilingual-uncased, all of which are available in the HuggingFace model repository [28, 29, 30]. All models were trained on a NVIDIA Titan RTX GPU, using PyTorch and the transformers library.

5. Experimental Results

In the following subsections, the results of the experiments of classification with the different systems described before are presented. The best evaluation metrics reported in the tables are shown in bold, while the second best are shown underlined. A hard-hard evaluation and a soft-soft evaluation were performed, depending on the experimental setup. The hard evaluation was used for experiments that provided hard category outputs, while the soft evaluation was applied to those that provided probabilistic outputs for each category.
The official metric for the hard-hard evaluation is the ICM (Information Contrast Measure) metric [31]. A normalized version of ICM (ICM Norm) is also reported, as well as F1 of positive class for Task 1 and Macro F1 for Task 2. For the soft-soft evaluation, a modification of the ICM metric, ICM-Soft [5], is used. A normalized version of ICM-Soft (ICM-Soft Norm) is also reported, as well as Cross Entropy (CE).

5.1. Results on Dev Partition

The results obtained by our systems for the dev partition are shown in this section. The dev set corresponds to the one provided in the EXIST 2024 dataset. The distribution in this set can be observed in Table 1. A *tie* is declared when half of the annotators classified the tweet as sexist and the other half as non-sexist.

Table 1
Distribution by classes of Task 1 in the hard setting and languages for the Dev partition.

Task 1 Hard label	English	Spanish	Total
YES	194	261	455
NO	250	229	479
tie	45	59	104

Zero-shot Prompts. For Task 1 Hard setting, the results of the zero-shot classification from LLM and prompts is shown in Table 2. These experiments were aimed to evaluate the performance of the LLM for sexism detection, according to the different provided prompts. The prompt that resulted of the refinement with ChatGPT is the one with the best performance in the table. It is interesting to notice that the prompt

refinement process (simple_prompt, class_definitions_prompt, class_definitions_refined_prompt) is reflected in the evaluation metrics, since they are improving as the simple_prompt is refined. In this way, class_definitions_refined_prompt is chosen as the best LLM response for the strategy mentioned in Section 3.2.

Table 2
Results of classifications with LLM and prompts in the dev set for Task 1, hard setting. The best evaluation metrics are shown in bold, while the second best are shown underlined

System	ICM	ICM-Norm	F1_YES
simple_prompt	-0.0224	0.4888	0.7015
class_definitions_prompt	<u>0.3621</u>	<u>0.6811</u>	0.7750
class_definitions_refined_prompt	**0.4635**	**0.7318**	**0.8033**
profiled_simple_prompt	0.3177	0.6589	<u>0.7820</u>

LLM Responses Aware Modeling. On the other hand, we compared fine-tuning a transformer for classification using the tweet as input, to fine-tuning with the tweet combined with responses from the prompts and the Gemini API. This was aimed to discern whether this form of unification of models could perform better than fine-tuning without modifications of inputs. Results are shown in Table 3. The experiments with Resp_aware_in (Tweet + prompt responses), performed better than its counterparts, except for the Twitter-XLM-R model, which is the best performing in the table. Nevertheless, in [32] is observed that the performance of variations of this model for hard labels in the EXIST test set could be improved. For this reason, we decided to choose the second best performing in ICM metric in Table 3, which is the Resp_aware_in + XLM-R model, as "the best" fine-tuned model.

Table 3
Results of classifications with fine-tuned models and fine-tuned models with new input in the dev set for Task 1, hard setting.

System	ICM	ICM-Norm	F1_YES
FT mBERT	0.4920	0.7461	0.8275
Resp_aware_in + FT mBERT	0.4925	0.7464	**0.8578**
FT XLM-R	0.5239	0.7621	0.8347
Resp_aware_in + FT XLM-R	<u>0.5832</u>	<u>0.7917</u>	<u>0.8509</u>
FT Twitter-XLM-R	**0.5878**	**0.7940**	0.8374
Resp_aware_in + FT Twitter-XLM-R	0.5382	0.7692	0.8146

Unification of Experts. For the rest of the unifying strategies, results are shown in Table 4. The unifying strategy that lead to the best result is the ensemble of *best LLM response or best fine-tuned model*. This strategy consisted in taking the ensemble of the response generated with class_definitions_refined_prompt and the response generated with the Resp_aware_in + FT XLM-R, with a logical OR operation in the binary predictions. This result was unexpected because this strategy only unifies two systems, while the *proportion of votes* strategy unifies knowledge of seven systems and was expected to be more robust. The proportion of votes considered the responses of the four prompts outlined in Table 2 and the three Resp_aware_in + FT models in Table 3. The purpose of these experiments is to explore unification strategies that do not require more computational resources and that can leverage the knowledge already generated in the previous experiments.

Soft task: Leave one expert out in the unification. For Task 1 Soft setting, the *proportion of votes* strategy was modified to consider only the proportion of YES and NO answers from six systems. This was aimed for trying to emulate the number of annotations present in the EXIST 2024 dataset. In this way, if the classification of the different systems were accurate, the distribution of sexism identification could be estimated more precisely. The predictions considered were the ones generated by the responses of prompts (Table 2) and the Resp_aware_in + FT models (Table 3). The best result in

Table 4
Results of classifications with combinations of responses from LLM and prompts and fine-tuned models with new input for Task 1, hard setting.

System	ICM	ICM-Norm	F1_YES
Proportion of votes	0.5778	0.7890	0.8530
class_definitions_refined_prompt OR Resp_aware_in + FT XLM-R	**0.6051**	**0.8027**	**0.8679**

ICM (Table 5) was achieved by the system that left out responses from simple_prompt, which is the prompt with less context or instructions.

Table 5
Results of classifications with combinations of responses from LLM and prompts and fine-tuned models with new input for Task 1, soft setting.

System	ICM-Soft	ICM-Soft-Norm	CE
out mBERT	0.7756	0.6253	0.9957
out Twitter-XLM-R	0.7529	0.6216	0.9978
out XLM-R	0.7494	0.6211	0.9809
out simple_prompt	**0.9356**	**0.6511**	1.0227
out class_definitions_prompt	0.8312	0.6343	**0.9466**
out class_definitions_refined_prompt	0.7666	0.6238	0.9990
out profiled_simple_prompt	0.8599	0.6389	0.9622

For **Task 2 Hard** setting, results of the experiments carried out are in Table 6. The tested models were variations of cascades of fine-tuned transformers for classification. The first model decided whether the tweet was sexist or not. Then, for the tweets classified as sexist, the second model decided the source intention between three classes (DIRECT, JUDGEMENTAL and REPORTED). For the first model, the Resp_aware_in + FT XLM-R was used for all the experiments because its metrics showed consistency in terms of ranking in Table 3. For the second model, different systems were considered and consisted in the ones listed in Table 6, which are fine-tuned transformers for classification with the tweets as input and with the modified Resp_aware_in set. The purpose of these experiments is to combine the techniques employed earlier and to evaluate their performance in a multiclass hierarchical classification setting. It can be noticed that, again, the Resp_aware_in models performed better than its counterparts in most of the metrics, which could lead to hypothesize that the fine-tuning process is, in fact, learning from the different responses of the prompts.

Table 6
Results of classifications with fine-tuned models and fine-tuned models with new input in the dev set for Task 2, hard setting.

System	ICM	ICM-Norm	Macro F1
FT mBERT	0.3559	0.6113	0.5888
Resp_aware_in + FT MBERT	0.3902	0.6220	0.5961
FT XLM-R	**0.4820**	**0.6507**	0.6261
Resp_aware_in + FT XLM-R	0.4640	0.6451	**0.6325**
FT Twitter-XLM-R	0.4146	0.6296	0.5952
Resp_aware_in + FT Twitter-XLM-R	0.4270	0.6335	0.6050

5.2. Results on Test

The test EXIST 2024 dataset consists of 2,076 tweets, divided into 1,098 from Spanish and 978 from English. Results of our submissions in the EXIST 2024 evaluation are shown in Tables 7, 8 and 9. The majority of our systems ranked in the top ten of all the evaluations. The best submitted system consists in the *proportion of votes* strategy, which ranked third for Task 1 Hard evaluation. It is worth to notice that the ranking order achieved in the dev set is not preserved in the test set. Even more, the submission with the best metrics in the dev set (ensemble of class_definitions_refined_prompt or Resp_aware_in + FT XLM-R) was the worst ranked system of all our submissions. This leads us to believe that more experiments with different seeds and parameters to ensure stability need to be performed.

Table 7
Results for submitted systems for Task 1, Hard-Hard evaluation, all instances. The proportion of votes strategy achieved the highest performance in terms of ranking.

System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
Proportion of votes	3	0.5926	0.7978	0.7899
Resp_aware_in + FT XLM-R	12	0.5486	0.7757	0.7746
class_definitions_refined_prompt OR Resp_aware_in + FT XLM-R	16	0.5357	0.7692	0.77

Table 8
Results for submitted systems for Task 1, Soft-Soft evaluation, all instances. Leaving out profiled_simple_prompt responses yielded better results than leaving out BERT results.

System	Rank	ICM-Soft	ICM-Soft-Norm	CE
out profiled_simple_prompt	5	0.9285	0.6489	1.2252
out mBERT	6	0.8468	0.6358	1.2538
out Twitter-XLM-R	8	0.8213	0.6317	1.2684

Table 9
Results for submitted systems for Task 2, Hard-Hard evaluation, all instances. For multiclass classification, our systems achieved competitive rankings.

System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
Resp_aware_in + FT Twitter-XLM-R	7	0.2643	0.5859	0.5171
Resp_aware_in + FT XLM-R	8	0.2346	0.5763	0.5195
Resp_aware_in + FT MBERT	13	0.1615	0.5525	0.4885

5.3. Best Ranks

All instances of the submitted systems were included in the evaluation, with separate assessments conducted for both Spanish and English. Our best systems achieved high rankings across evaluations involving all instances, as well as in the specific Spanish and English evaluations. Table 10 summarizes the best results for the hard settings of Tasks 1 and 2, while Table 11 summarizes the best soft evaluation results for Task 1.

6. Conclusions

In this work we observed that unifying classification strategies using different techniques, such as generating responses with LLMs, as well as fine-tuning transformers for classification, is a simple but

Table 10
Results for our best submitted systems, Hard-Hard evaluation. Our systems for binary classification performed better than our multiclass submissions.

Task	Instances	System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
1	all	Proportion of votes	3	0.5926	0.7978	0.7899
1	Spanish	Proportion of votes	3	0.6098	0.805	0.8148
1	English	Proportion of votes	4	0.5612	0.7864	0.7576
2	all	Resp_aware_in + FT Twitter-XLM-R	7	0.2643	0.5859	0.5171
2	Spanish	Resp_aware_in + FT Twitter-XLM-R	6	0.3203	0.6	0.5466
2	English	Resp_aware_in + FT Twitter-XLM-R	5	0.1764	0.5611	0.4714

Table 11
Results for the best submitted systems, Soft-Soft evaluation. The top-performing system was most effective with Spanish instances.

Task	Instances	System	Rank	ICM-Soft	ICM-Soft-Norm	CE
1	all	out profiled_simple_prompt	5	0.9285	0.6489	1.2252
1	Spanish	out profiled_simple_prompt	4	1.0223	0.664	1.1389
1	English	out profiled_simple_prompt	7	0.7691	0.6235	1.3221

effective approach that produces good and competitive results. These approaches proved to be effective for both hard and soft settings, for binary and fine-grained tasks in sexism detection (identification and categorization), and for results categorized by Spanish and English (see A). The previous findings suggest that multilingual approaches are competitive, and potentially more practical compared to use individual systems for each language. We think that more experiments need to be conducted to build robust systems that perform consistently across development and test partitions.
Additionally, there are numerous efforts to unify the responses of different models. This work represents a step towards that direction, and is expected to be extended with new and diverse techniques to optimize integration of insights generated by various models. As suggested from the prompt refinement process, the identification of sexist tweets improved as the method of requesting the response by a prompt from the model improved. This suggests an interesting direction to explore further, as LLMs could potentially better detect sexist situations if an optimal way to prompt for that identification is found. Interestingly, the zero-shot experiments in Gemini were able to obtain accurate insights about sexism. Therefore, exploring a few-shot setting could be a promising approach to investigate further.

Ethical Concerns

We acknowledge that this study is confined to social media texts, which may not represent all populations or cultures universally. Additionally, we recognize that LLMs can produce responses with various biases. Furthermore, the underrepresentation of specific groups in the training data can result in models that perform inadequately or inappropriately when addressing these groups. It is also crucial to mention that steps were taken to anonymize the tweets, ensuring individual privacy is protected.

Acknowledgments

Tavarez-Rodríguez acknowledges CONAHCYT and CIMAT for the support through the PhD scholarship (CVU 859147). The authors gratefully acknowledge *Centro de Investigación en Matemáticas* (CIMAT) and *Consejo Nacional de Humanidades, Ciencias y Tecnologías* (CONAHCYT) for the computing resources

provided by the CIMAT Bajio Supercomputing Laboratory (#300832) and the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies. Sanchez-Vega acknowledges CONAHCYT for its support through the program *"Investigadoras e Investigadores por México"* (Project ID.11989, No.1311). Rosales-Pérez acknowledges CONAHCYT for its support through the grant project *Búsqueda de arquitecturas neuronales eficientes y efectivas* (CBF2023-2024-2797).

References

[1] 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.

[2] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)., Springer, 2024.

[3] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024 – learning with disagreement for sexism identification and characterization in social networks and memes (extended overview), in: Working Notes of CLEF 2024- Conference and Labs of the Evaluation Forum. Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, Alba García Seco de Herrera Eds., 2024.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[5] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.

[6] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 255–269. URL: https://aclanthology.org/2021.eacl-main.20. doi:10.18653/v1/2021.eacl-main.20.

[7] T. Le Scao, A. Rush, How many data points is a prompt worth?, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2627–2636. URL: https://aclanthology.org/2021.naacl-main.208. doi:10.18653/v1/2021.naacl-main.208.

[8] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[9] A. F. M. de Paula, R. F. da Silva, I. B. Schlicht, Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models, arXiv preprint arXiv:2111.04551 (2021).

[10] E. Villa-Cueva, F. Sanchez-Vega, A. P. López-Monroy, Bi-ensembles of transformer for online bilingual sexism detection., IberLEF@ SEPLN (2022).

[11] R. Koonireddy, N. Adel, Roh_neil@ exist2023: detecting sexism in tweets using multilingual language models, Working Notes of CLEF (2023).

[12] A. Vetagiri, P. K. Adhikary, P. Pakray, A. Das, Leveraging gpt-2 for automated classification of online sexist content, Working Notes of CLEF (2023).
[13] L. Tian, N. Huang, X. Zhang, Efficient multilingual sexism detection via large language models cascades, Working Notes of CLEF (2023).
[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[15] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, S. Weinbach, GPT-NeoX-20B: An open-source autoregressive language model, in: A. Fan, S. Ilic, T. Wolf, M. Gallé (Eds.), Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models, Association for Computational Linguistics, virtual+Dublin, 2022, pp. 95–136. URL: https://aclanthology.org/2022.bigscience-1.9. doi:10.18653/v1/2022.bigscience-1.9.
[16] J. A. García-Díaz, R. Pan, R. Valencia-García, Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english, Mathematics 11 (2023) 5004.
[17] F. Plaza-del Arco, D. Nozza, D. Hovy, Leveraging label variation in large language models for zero-shot text classification. arxiv, arXiv preprint arXiv:2307.12973 (2023).
[18] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.
[19] F. M. Plaza-del Arco, D. Nozza, D. Hovy, et al., Respectful or toxic? using zero-shot learning with language models to detect hate speech, in: The 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, 2023.
[20] G. DeepMind, Gemini api, 2023. URL: https://ai.google.dev/gemini-api/docs?hl=es-419.
[21] OpenAI, Chatgpt: May 2024 version, 2024. URL: https://chat.openai.com/, large language model.
[22] X. Lu, X. Wang, Generative students: Using llm-simulated student profiles to support question item evaluation, arXiv preprint arXiv:2405.11591 (2024).
[23] G. DeepMind, Gemini api safety settings, 2023. URL: https://ai.google.dev/gemini-api/docs/safety-settings?hl=es-419.
[24] 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, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.
[25] 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.
[26] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, , M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 38 (2021) 1385–1470.
[27] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: A python toolkit for opinion mining and social nlp tasks, 2023. arXiv:2106.09462.
[28] C. NLP, twitter-xlm-roberta-base, 2022. URL: https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base.
[29] Facebook, xlm-roberta-large, 2019. URL: https://huggingface.co/FacebookAI/xlm-roberta-large.
[30] Google, Bert-base-multilingual-uncased, 2018. URL: https://huggingface.co/google-bert/bert-base-multilingual-uncased.
[31] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.
[32] G. Radler, B. Ersoy, S. Carpentieri, Classifiers at exist 2023: detecting sexism in spanish and english tweets with xlm-t, Working Notes of CLEF (2023).

A. Results on the test set, categorized by language

Tables 12, 13, 14, 15, 16 and 17 show the results for all of our nine systems submitted categorized by their performance on Spanish and English datasets. For Task 1 Hard evaluation, the best system ranked 3th and 4th for Spanish and English, respectively. For Task 1 Soft evaluation, rankings are 4th and 7th (Spanish and English). In Task 2 Hard evaluation, our systems ranked 6th for Spanish and 5th for English. All of our results on the test set were better for Spanish than for English for Task 1. Conversely, for Task 2, the opposite occurred. The previous results suggest that the multilingual approaches proposed in this work could be effective not only in the evaluation of multilingual datasets but also in monolingual settings.

Table 12
Results for systems submitted for Task 1, Hard-Hard evaluation, Spanish.

System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
Proportion of votes	3	0.6098	0.805	0.8148
Resp_aware_in + FT XLM-R	11	0.5599	0.78	0.7972
class_definitions_refined_prompt OR Resp_aware_in + FT XLM-R	13	0.5426	0.7714	0.7919

Table 13
Results for systems submitted for Task 1, Hard-Hard evaluation, English.

System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
Proportion of votes	4	0.5612	0.7864	0.7576
Resp_aware_in + FT XLM-R	18	0.5204	0.7656	0.7439
class_definitions_refined_prompt OR Resp_aware_in + FT XLM-R	20	0.5137	0.7622	0.7412

Table 14
Results for systems submitted for Task 1, Soft-Soft evaluation, Spanish.

System	Rank	ICM-Soft	ICM-Soft-Norm	CE
out profiled_simple_prompt	4	1.0223	0.664	1.1389
out mBERT	7	0.9468	0.6518	1.1495
out Twitter-XLM-R	8	0.9152	0.6468	1.1812

Table 15
Results for systems submitted for Task 1, Soft-Soft evaluation, English.

System	Rank	ICM-Soft	ICM-Soft-Norm	CE
out profiled_simple_prompt	7	0.7691	0.6235	1.3221
out mBERT	9	0.6753	0.6084	1.3709
out Twitter-XLM-R	11	0.6602	0.606	1.36632

Table 16
Results for systems submitted for Task 2, Hard-Hard evaluation, Spanish.

System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
Resp_aware_in + FT Twitter-XLM-R	6	0.3203	0.6	0.5466
Resp_aware_in + FT XLM-R	7	0.2883	0.5901	0.5501
Resp_aware_in + FT MBERT	12	0.203	0.5634	0.5148

Table 17
Results for systems submitted for Task 2, Hard-Hard evaluation, English.

System	Rank	ICM-Hard	ICM-Hard-Norm	F1_YES
Resp_aware_in + FT Twitter-XLM-R	5	0.1764	0.5611	0.4714
Resp_aware_in + FT XLM-R	11	0.148	0.5512	0.4771
Resp_aware_in + FT MBERT	14	0.0924	0.532	0.4496