

Stacked Reflective Reasoning in Large Neural Language Models

Notebook for the EXIST Lab at CLEF 2024

Kapioma Villarreal-Haro^{1,*†}, Fernando Sánchez-Vega^{1,2,†}, Alejandro Rosales-Pérez^{3,†} and Adrián Pastor López-Monroy^{1,†}

¹Mathematics Research Center (CIMAT), Jalisco S/N Valenciana, 36023, Guanajuato, Guanajuato, México.

²Consejo Nacional de Ciencia y Tecnología (CONACYT), Av. Insurgentes Sur 1582, Col. Crédito Constructor, 03940, CDMX, México.

³Mathematics Research Center (CIMAT), Monterrey, Av. Alianza Centro 502, Apodaca, 66628, Nuevo León, México.

Abstract

Sexism, far from being merely a conceptual issue, is a concerning and pervasive social health problem that negatively impacts individuals' well-being and perception. In today's digital era, as sexism permeates online platforms, the creation of systems that detect this type of content is a challenging yet essential task. This paper presents the approach of the CIMAT-GTO team to Task 1 of EXIST 2024, which involves identifying tweets with sexism-related content. Our proposal takes advantage of the reasoning capabilities of Llama 3 in a two-step process. Initially, we generate rationales to analyze the nature of the tweets. Then, in a second step, we let the model reflect on the previously produced reasoning. The intuitive idea is to create text that supports opposite categories, and expect the model to contrast valid and invalid reasons by itself. We then use these generated rationales as extra information to complement the tweets and fine tune a Twitter-specialized XLM-RoBERTa model. Our experiments showed that incorporating Llama 3's rationales improves performance compared to only using tweets and yields competitive results in the task, demonstrating the potential of these methods.

Keywords

Generative Large Language Models, Large Language Model Reasoning, Stacked Large Language Models, Transformers, Sexism Detection, Social Media

1. Introduction

In today's world, social media is an essential platform for the communication and diffusion of information and opinions among individuals. However, social media interactions are often related to misleading or harmful content. For instance, users might directly express bias in their own generated content. Alternatively, they could engage by sharing and commenting on biased content created by other users. Within these interactions, one major social concern is sexism, defined as prejudice or discrimination based on sex or gender [1].

Sexism negatively affects the psychological well-being of women and men not only in everyday face-to-face interactions [2], but also in social media platforms [3]. In this context, social media has been used for two contrary purposes: 1) as a platform for bias dissemination where misleading information and hateful behavior against women are spread, and 2) as a means for bias awareness and activism, enabling users to address, report, and discuss misogynistic and sexist narratives [4, 5].

Instances of harmful sexist expressions in social media include hostile behavior and negative evaluations of female job candidates [6]. Other problems involve the distribution and consumption of content that perpetuates appearance anxiety, body shame, and eating disorder behaviors, primarily among women [7]. On the other hand, social media has also been used as a platform for positive impact behaviors, such as mobilizing digital media in response to shaming, harassment, and rape culture [8].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

†These authors contributed equally.

✉ kapioma.villarreal@cimat.mx (K. Villarreal-Haro); fernando.sanchez@cimat.mx (F. Sánchez-Vega); alejandro.rosales@cimat.mx (A. Rosales-Pérez); pastor.lopez@cimat.mx (A. P. López-Monroy)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

In the modern world, the efficient identification of sensitive content is crucial due to the vast volume of data, human biases, and the profound impact of this task. Despite advancements in computer science and the deployment of more accurate and sophisticated models, the challenge remains unsolved. In this context, several efforts include shared tasks to address this issue, such as Automatic Misogyny Identification at IberEval [9], Multimedia Automatic Misogyny Identification and Explainable Detection of Online Sexism at SemEval [10, 11], and sEXism Identification in Social neTworks (EXIST) [12, 13].

This paper describes CIMAT-GTO’s participation in Task 1 of EXIST 2024, which tackles the binary detection of sexism in tweets. We propose a technique that takes advantage of the reasoning capabilities of Large Language Models (LLMs). In a first step, the LLM generates reasoning supporting the target categories for Task 1, which are sexism-related and not-sexism-related. In a second step, we exploit and take advantage of this valuable information and feed it to the LLM to reflect on this reasoning. The reasoning outputs are then used to extend the information provided by the tweets in fine tuning an XLM-RoBERTa model pre-trained on multilingual tweets.

2. EXIST Shared Lab

EXIST shared tasks aim to detect and capture sexism-related content in social networks while identifying intention and fine-grained topics [12, 13]. EXIST has evolved from only analyzing content in text format to multi-modal content. While the 2023 edition focused on detecting and categorizing sexist tweets, the 2024 edition extended its scope to encompass both tweets and meme images.

For tweets, three primary tasks were established.

- **Task 1:** Binary classification to identify tweets with sexism-related content.
- **Task 2:** Multi-class classification to identify the intention of the tweet.
- **Task 3:** Multi-label classification to categorize the types of sexism expressed.

Analogous tasks were introduced regarding meme images. Systems that address such shared tasks can be set in two contexts: a hard setting, in which systems aim to predict a hard conventional output category or set of categories, and a soft setting, in which systems are intended to provide probabilities instead. In the case of Task 1, the categories consist of sexist and non-sexist tweets. In this paper, we address Task 1 in the hard context.

2.1. Tweet Dataset

The tweet dataset contains 10,034 tweets in Spanish and English; 6920 for training, 1038 for development, and 2076 for testing. Each tweet was labeled by six annotators selected such that they had different demographic characteristics to minimize bias in the labeling. The age ranges and gender of the annotators belonged to the sets $\{18 - 22, 23 - 45, 46+\}$ and $\{Female, Male\}$. All tweets were annotated such that there was one annotator from each of the six possible gender-age groups.

For Task 1, annotators were required to indicate whether the tweet was related to sexism or not. In the following, we will refer to these two categories as *sexist* and *not sexist* for simplicity purposes, since that’s the labeling convention used in EXIST. It is worth noting that the sexist category not only encompasses tweets with direct harmful messages but also tweets where sexism-related situations are being discussed or exposed. Therefore, this task is not only to detect direct hateful behavior against women.

Table 1 summarizes the train dataset partition according to language and label assigned in Task 1. The dataset is mostly balanced among languages and sexism-related and not-sexism-related categories. A minority, less than 13% of the tweets, do not have a majority class attached to them.

Table 1

Distribution of Training Set. Classes are mostly balanced between languages. Among the tweets with no ties between annotator votes, sexist and not-sexist categories are mostly balanced.

Training					
6920					
Spanish			English		
3660			3260		
no majority class	majority class available		no majority class	majority class available	
466	3194		390	2870	
	related to sexism	not related to sexism		related to sexism	not related to sexism
	1560	1634		1137	1733

3. Previous Work

3.1. Previous Editions of EXIST

During EXIST 2023, the most commonly used approaches to address binary sexism classification included the use of variations and combinations of the following three [14]:

1. Transformer-based architectures like BERT or RoBERTa. Models were either pre-trained and then fine-tuned or trained from scratch. They included general knowledge text or domain-specific text like tweets.
2. Classical Machine and Deep Learning Methods utilizing input embeddings from pre-trained models and additional attributes like toxicity and sentiment metrics, or linguistic and handcrafted features.
3. Data augmentation techniques, external datasets, and ensembles of multiple models.
4. Addressing the task as a monolingual problem using separate models for each language, or as a multilingual problem using cross-lingual or translation techniques.

While classical methods and architectures remain popular and competitive, models like GPT, Llama, or Gemini have not yet been deeply explored. In EXIST 2023, GPT-based large language model cascades were shown to be competitive and ranked among the top systems. Notwithstanding the fact of being effective, their strategy appears to be used in a classification setting rather than generating text [15]. We speculate that the generation capabilities of LLMs to provide key ideas, identify elements, and compare arguments that are substantial for a correct label assignment, as LLMs have shown to achieve state-of-the-art in several tasks even without fine-tuning.

3.2. Large Language Models “Reasoning” Capabilities.

Recent research on Large Language Models is focused on exploiting their “reasoning” capabilities. An overview of the current state-of-the-art knowledge on reasoning in LLMs provides a review of different techniques [16]. Some approaches involve traditional, fully supervised fine-tuning to generate rationales on a specific domain. Others are prompt-based and in-context learning methods that do not require fine-tuning. A third type consists of hybrid approaches that combine both of the previous.

Given that fine-tuning massive generative LLMs is not efficient regarding computational resources, prompt-based methods have gained popularity because of their knowledge capacities. Adding prompt pieces in zero-shot scenarios such as “Think step by step” [17] or “Let’s first understand the problem and devise a plan to solve the problem. Then, let’s carry out the plan and solve the problem step by step” [18] encourage the models to provide rationales that guide the answer. Other prompts like “This is very important to my career” have positively influenced the performance in some tasks [19]. However, in

certain settings, prompts like “Think step by step” might lead the model to produce inaccurate answers or generate harmful content [20].

In few-shot scenarios, techniques like chains of thought have been shown to improve the answers by demonstrating a thought process and encouraging the model to provide its own behind the answer [21]. The order and quality of the few-shot demonstrations are crucial and impact the performance. Some studies propose techniques for providing good permutations of the examples to enhance the quality of the results [22]. Several strategies include encouraging the model to take advantage of multiple prompts. This can be used to answer the same question and then apply approaches to regularize the prompt consistency to obtain a final label [23]. Other techniques propose using external “prompters” that iteratively prompt the LLM to recall a series of knowledge and derive a “chain of thought” [24]. Additional approaches include subdividing all the context into questions and enabling cross-model communication during problem-solving to aggregate the answers [25].

All these strategies require a prompt-refining process, to some extent, to provide better context and enhance the use of these generative LLMs. Still, this refinement process is usually not automatic and is done through a highly qualitative assessment.

Another important consideration is the source and target languages used to prompt an LLM. Studies have shown a disparity between the performance of LLMs in English and non-English languages, with LLMs generally performing better in English [26]. These techniques and considerations have become more popular and have been leveraged in large language models like GPT-3, Llama, Gemini, and Claude to generate more accurate answers to different tasks.

4. Baseline

4.1. Preprocessing the data

We will focus on the data’s hard label predictions. Although we will not dismiss the individual annotators’ labels, we will filter the tweets so that only those classified as sexist and not sexist by majority vote will be included. We preprocess the tweets using the library “pysentimiento” [27] in the following way:

1. User handles are substituted with *@user*.
2. URL directions are replaced with the special token *url*.
3. The # symbol is substituted by the special token *hashtag*, and the content in multi-word hashtags is split into separate words.
4. Emojis are replaced with their text descriptions.

4.2. XLM-RoBERTa fine tuned

As a baseline, we worked with a Twitter-specific multilingual language model that consists of an XLM-RoBERTa architecture trained on multilingual tweets [28]. We fine-tuned the model to predict the Task 1 labels. The input consists of the tokenized tweets. We built two variants: prediction of the hard binary label and simultaneous prediction of the hard binary label and the single-annotator labels grouped by age and gender. The second variant was chosen because it resulted in a slightly better macro-F1 score for the hard binary labels. We will refer to this system as *XLM-RoBERTa-Baseline*.

5. The Proposed Method

We have an experimental setup in two stages: In the first stage, we generate “reasoning” texts using an LLM that aims to understand the tweets’ nature. In the second stage, we use the generated texts to process them further with a pre-trained XLM-RoBERTa model. We explain these details in the following section.

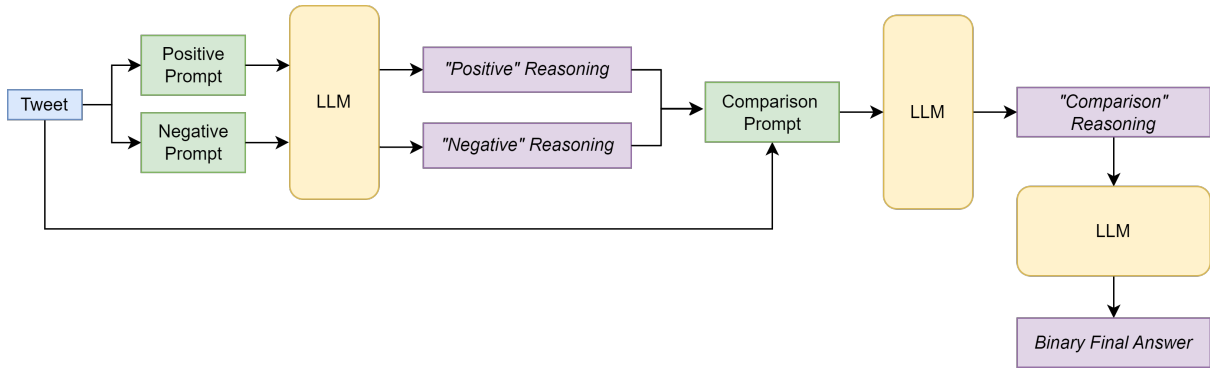


Figure 1: LLM Generation of a *Positive*, *Negative*, and *Stacked Comparison Reasonings* of a tweet.

5.1. LLM Stacked Reasoning

We relied on an autoregressive LLM to generate text analysis of a tweet. Rather than asking the model to assign a label and provide an explanation in one straightforward step, we created rationales that support the tweet’s target categories, sexist and not sexist. Then, we let the model compare both arguments and choose the most accurate. We hypothesize that this is better than the direct closed question “*Is this tweet sexist?*” because, despite the validity of the reasonings for each category, the model internally evaluates the correctness of the statements produced. The setting is the same for both tweets in English and Spanish, and all the generated analyses are in English.

The first two steps occur independently:

1. *Positive Reasoning*. Generating analysis that supported the idea that the tweet was related to sexism.
2. *Negative Reasoning*. Generating analysis that supported the idea that the tweet was **not** related to sexism.

In a further step, the LLM is asked to “reflect” on the opposite texts it produced:

3. *Stacked Comparison Reasoning*. The model is fed with the information generated in the *Positive* and *Negative Reasonings* and has the chance to compare them.

This process is illustrated in Figure 1. The LLM generates three rationales that provide insight into the nature of the tweet. We call all the rationales to be *Tweet Reasonings* (marked as purple boxes in Figure 1), and they will be further used in the next section.

As an experimental setting, we ask the model to answer as a *gender equality specialist*, to respond *as unbiased as possible* or to be *concise*. The exact prompts and an output example can be found in Appendix A. The LLM selected for this reasoning stage was Llama 3, an open-source auto-regressive language model developed by Meta AI [29]. We used the *Llama 3 8B Instruct* version.

In a conclusive step, we ask the model to provide a label synthesizing the reflection made in the *Stacked Comparison Reasoning*. The *Binary Final Answer* produced is not used in the second stage of our proposed methodology, but we use it to assess the method’s gain in performance. The answers provided by this method are reported as *B-StackLlama*. Even though this workflow encourages the model to produce better-structured answers, the final classification is still not accurate enough. The second stage, which will be explained in the next section, addresses this issue.

5.1.1. Fine tune of XLM-RoBERTa

In this stage, we test the reasoning provided by the LLM in the previous step as a supplement to the tweets. To enhance the XLM-RoBERTa baseline described before, we experimented with feeding the tweet concatenated with the reasonings generated. Due to the maximum input token restriction, we are limited to incorporating one reasoning at a time.

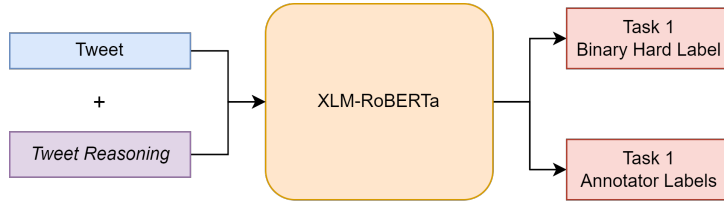


Figure 2: Fine tuning of Tweet-focused XLM-RoBERTa. Input is concatenation (+) using special separator token *[SEP]* of a tweet with either a *Positive*, *Negative* or *Stacked Comparison Reasoning* generated by an LLM (purple boxes in Figure 1)

We fine tune three different XLM-RoBERTa models using different inputs. The input of each model consists of the concatenation using the special separator token *[SEP]* of a tweet with one of its corresponding *Tweet Reasonings* generated in the previous stage. The different reasonings *Positive*, *Negative* or *Stacked Comparison Reasoning* yield a different fine-tuned model, which we will call *P-LLM-R-Stack-Ra*, *N-LLM-R-Stack-Ra* and *C-LLM-R-Stack-Ra* respectively.

Figure 2 shows the process realized. In addition to the individual *{P, N, C}-X-LLM-R-Stack-Ra* models, we create ensembles of individual models aiming to capture a more accurate label. Ensembles are generated by aggregating the scores that individual systems assign to the labels. The final label is designated as the one with the highest score.

6. Tweet Fine Grained Sexist Evaluation Questionnaire

A more general task can be addressed as an aggregate of fine-grained tasks. In particular, there are questionnaire-based retrieval models used to provide a final response diagnosis, like the case of depression [30].

Using this idea, in addition to the previous approach, we created a list of binary and close-option questions meant to identify the nature of the tweet. The questionnaire focused on identifying offensiveness, intention, and whether fine-grained sexism-related topics were expressed directly or passively. The complete questionnaire is included in Appendix B. The answers to these questions were fed through a multi-layer feed-forward neural network that aimed to predict all three tweet-related tasks. Because this method did not outperform the scores obtained by the previous method on the validation set by itself, we used it only to enhance the previous techniques in an ensemble setting in the fashion described before. Reference for this method is *Q-Llama-MLP*.

Although this method seems promising, the quality of the answers is sensible to question formulation. Appendix B shows an example of the refinement through a qualitative assessment of a question that provides more accurate results in its final versions. We believe that all the questions in the original questionnaire can be refined into versions that lead to more accurate and representative answers and can be further used both to address the binary task and to identify fine-grained topics.

7. Evaluation Results

In this section, we discuss the performance of our proposals over the validation set and the official evaluation metrics at EXIST 2024.

7.1. Preliminary evaluation over validation data

This section is dedicated to evaluating the effectiveness and the impact of the key components of the methods proposed in the previous sections. We include the system *Ensemble_{NC}*, which is the ensemble merging the two best performing individual systems *N-LLM-R-Stack-Ra* and *C-LLM-R-Stack-Ra*. In extension to this ensemble, we also include *Ensemble_{NCQ}*, which in addition to individual systems *N-LLM-R-Stack-Ra* and *C-LLM-R-Stack-Ra* incorporates *Q-Llama-MLP*. The ensemble setting is, as

Table 2

Results over the validation set. Hard Evaluation. Incorporating the reasonings benefits the system. The best systems overall are ensembles.

System	Input information	lang	F1 positive class
XLM-RoBERTa-Baseline	Tweets only	en/es	0.8291
B-StackLlama	Tweets only	en/es	0.5966
N-LLM-R-Stack-Ra	Tweets + Negative Reasoning	en/es	0.8489
C-LLM-R-Stack-Ra	Tweets + Stacked Comparison Reasoning	en/es	0.8474
P-LLM-R-Stack-Ra	Tweets + Positive Reasoning	en/es	0.8325
Ensemble _{NC}	{C, N}-LLM-R-Stack-Ra	en/es	0.8568
Q-Llama-MLP	Questionnaire Answers	en/es	0.7994
Ensemble _{NCQ}	{C, N}-LLM-R-Stack-Ra and Q-Llama-MLP	en/es	0.8591

described before, the result of equally aggregating the scores of the individual scores of each category to obtain the ensemble scores.

We present in Table 2 the F1 of the positive class for the different systems proposed considering both tweets in English and Spanish of the validation set. As observed, models $\{N, C, P\}$ -LLM-R-Stack-Ra outperform the baseline *XLM-RoBERTa-Baseline*. This demonstrates the benefit of incorporating the reasoning against using only tweets. The *Negative* and *Stacked Comparison Reasonings* incorporation perform slightly better than the *Positive Reasonings*. We also observe that *B-StackLlama* underperforms *XLM-RoBERTa-Baseline*, indicating the deficiency of relying only on the reasonings. Regarding *Q-Llama-MLP*, it does not beat *XLM-RoBERTa-Baseline*, partly due to the redundancy and inability to capture the questions’ nature and meaning. Despite that, *Q-Llama-MLP* enhances the ensemble’s performance slightly. The best performance overall belongs to *Ensemble_{NCQ}*.

During the prompting process of *B-StackLlama*, the binary final answer underestimates sexism-related tweets. The performance by itself is poor because the two-step process is biased to produce a final negative label. We additionally observed the length of the *Tweet Reasonings* also influences the quality of the response: too short, and the analysis might not have enough details; too long, and the answer might be repetitive, affecting the performance during the fine-tuning process of $\{N, C, P\}$ -LLM-R-Stack-Ra. We set up the reasoning generation to contain at most 200 tokens of the *Positive* and *Negative Reasonings* and the *Stacked Comparison Reasoning* to contain at most 250 tokens. An example showing the variation of the answer based on the token length is also included in Appendix A

7.2. Official Leaderbord

The official metrics for EXIST 2024 are ICM-hard and F1 of the positive class, and scores are divided by language [12, 13]. Table 3 summarizes the results obtained over the test set.

The single system *N-LLMRStack-Ra* performs less effectively than *Ensemble_{NC}*. We hypothesize that the *Negative* and *Stacked Comparison Reasonings* complement the tweets with different information during the fine-tuning process. In the case of the *Negative Reasoning*, as provided for all tweets, we expect to learn an internal differentiation between accurate and inaccurate supporting facts of the not-sexist category. In the *Comparison Stacked Reasoning*, we expect to capture the contrast between reasonings and correct the preference of one over the other if necessary. We think that the distinct aspects and relationships these reasonings capture contribute to the improved performance in *Ensemble_{NC}*.

The ensemble *Ensemble_{NCQ}* is the best performer of our systems. In particular, the difference between the top score with this system considering the set of evaluation tweets in English and Spanish is less than 0.01 in the ICM-Hard Norm and the F1 of the positive class, which shows the competitiveness of our method. Performance of this model is just slightly better than *Ensemble_{NC}*, and we believe that even though *Q-Llama-MLP* did not outperform the individual systems in validation, adding the output in *Ensemble_{NCQ}* provides a slight correction of the underestimation of sexist tweets.

Performance over the Spanish tweets is slightly better than performance over English tweets, and we conjecture that the multilingual setting benefits performance for both English and Spanish languages.

Table 3

Results over the Test set. Hard Evaluation. The best system overall is *Ensemble_{NCQ}*. The best results are achieved on Spanish Tweets.

System	Rank	lang	ICM-Hard Norm	F1 positive class
Top Score	1	en/es	0.800	0.7944
Ensemble _{NCQ}	5	en/es	0.7939	0.7903
Ensemble _{NC}	6	en/es	0.7914	0.7887
N-LLM-R-Stack-Ra	14	en/es	0.7718	0.7694
Top score	1	es	0.8108	0.8238
Ensemble _{NCQ}	4	es	0.8017	0.8123
Ensemble _{NC}	6	es	0.7986	0.8071
N-LLM-R-Stack-Ra	9	es	0.7830	0.7936
Top Score	1	en	0.8153	0.7610
Ensemble _{NCQ}	8	en	0.7784	0.7594
Ensemble _{NC}	9	en	0.7767	0.7626
N-LLM-R-Stack-Ra	25	en	0.7517	0.7350

The systems *N-LLMRStack-Ra*, *Ensemble_{NC}* and *Ensemble_{NCQ}* correspond to the runs *CIMAT-GTO_1.json*, *CIMAT-GTO_2.json*, *CIMAT-GTO_3.json* in the EXIST official leaderboard.

8. Conclusion

In this paper, we propose a methodology to detect sexism that takes advantage of the capabilities to generate rationales of LLMs, which was not used in EXIST’s previous editions. These rationales show the potential of the LLM to support the binary classes and the capabilities to compare two different reasoning processes. As shown in this paper, the LLM reflection alone is not enough to achieve competitive results, and it is biased to underestimate sexist tweets, which is not desirable. That is why using the rationales enhances the baseline results provided by an XLM-RoBERTa that is fine-tuned only using the tweets. It allows for information that explores the nature of the tweet and a more accurate classification. Results show that the proposed models are competitive and open the panorama of how internal knowledge and reasoning capabilities of autoregressive LLMs can address this task.

In future work, we plan to extend this approach to address other tasks, refine the prompts to obtain better rationales, and study the capabilities and limitations of other autoregressive LLMs. Questionnaire results can be explored to dive into the fine-grained classifications of sexism and analyze source intention and topic classification. A more in-depth exploration regarding the length of the rationales generated, the effects of prompting variation, alternative LLMs, and bias remains to be explored. This method also shows insight into the biases encoded in the large language models, as the model we chose (Llama 3) is misled to provide inaccurate explanations of the wrong category classification and, as observed in the scores, fails to choose the correct classification label by itself.

9. Ethical Concerns

It is important to note that the systems developed in this work predict binary labels based on annotators’ majority vote and might overlook people’s perceptions at an individual level. Another important distinction is label names, where instead of “sexist” and “not-sexist” they could be more accurately described as “sexism-related” and “non-sexism-related” to recognize two different perspectives: negative intention (diffusion of biased content) and positive intention that treats sensitive content (bias awareness and discussion).

The reasoning generated by the LLMs put into evidence the biased internal views they provide, where sexism-related content is underestimated. Using these labels by themselves or trusting the rationales generated should be considered carefully, as they can be highly misleading. Given the implications

of deploying detection systems for sensitive content, the proposed solution requires a more in-depth analysis by social scientists, and ethics and fairness experts. Misusing these systems might have significant implications, including the potential non-detection of toxic and dangerous content or the unintended censorship of discussions about social problems.

Acknowledgments

Villarreal-Haro acknowledges CONAHCYT for its support provided by the program *Becas Nacionales Para Estudios de Posgrados* (CVU 1309535). We thank CONAHCYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies and CIMAT Bajío Super-computing Laboratory (#300832). Sanchez-Vega acknowledges CONAHCYT for its support through the program "*Investigadoras e Investigadores por México*" (Project ID.11989, No.1311). Rosales-Perez acknowledges CONAHCYT for its support through the project grant *Búsqueda de arquitecturas neuronales eficientes y efectivas* (CBF2023-2024-2797).

References

- [1] A. Stevenson, C. Lindberg, *New Oxford American Dictionary*, Third Edition, OUP USA, 2010. URL: <https://books.google.com.mx/books?id=sZoFRwAACAAJ>.
- [2] J. Swim, L. Hyers, L. Cohen, M. Ferguson, Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies, *Journal of Social Issues* 57 (2001) 31–53. doi:10.1111/0022-4537.00200.
- [3] M. Paciello, F. D'Errico, G. Saleri, E. Lamponi, Online sexist meme and its effects on moral and emotional processes in social media, *Comput. Hum. Behav.* 116 (2020) 106655. doi:10.1016/j.chb.2020.106655.
- [4] E. L. Turley, J. Fisher, Tweeting back while shouting back: Social media and feminist activism, *Feminism & Psychology* 28 (2018) 128 – 132. URL: <https://api.semanticscholar.org/CorpusID:149235968>.
- [5] M. Foster, A. Tassone, K. Matheson, Tweeting about sexism motivates further activism: A social identity perspective., *The British journal of social psychology* (2020). doi:10.1111/bjso.12431.
- [6] J. Fox, C. Cruz, J. Y. Lee, Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media, *Comput. Hum. Behav.* 52 (2015) 436–442. URL: <https://api.semanticscholar.org/CorpusID:45231644>.
- [7] Z. Gong, Concept of beauty in the age of the internet: Impact of social media on appearance anxiety and body shame, *Communications in Humanities Research* (2023). URL: <https://api.semanticscholar.org/CorpusID:266066470>.
- [8] K. M. Jessalynn Keller, J. Ringrose, Speaking 'unspeakable things': documenting digital feminist responses to rape culture, *Journal of Gender Studies* 27 (2018) 22–36. URL: <https://doi.org/10.1080/09589236.2016.1211511>. doi:10.1080/09589236.2016.1211511. arXiv:<https://doi.org/10.1080/09589236.2016.1211511>.
- [9] E. Fersini, P. Rosso, M. E. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018, in: *IberEval@SEPLN, 2018*. URL: <https://api.semanticscholar.org/CorpusID:51942244>.
- [10] 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.
- [11] 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.
- [12] 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)*, 2024.
- [13] 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: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, 2024.
- [14] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 – learning with disagreement for sexism identification and characterization (extended overview), in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*, 2023.
- [15] L. Tian, N. Huang, X. Zhang, Efficient multilingual sexism detection via large language model cascades, in: *Conference and Labs of the Evaluation Forum, 2023*. URL: <https://api.semanticscholar.org/CorpusID:264441414>.
- [16] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1049–1065. URL: <https://aclanthology.org/2023.findings-acl.67>. doi:10.18653/v1/2023.findings-acl.67.
- [17] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Curran Associates Inc., Red Hook, NY, USA, 2024.
- [18] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, E.-P. Lim, Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2609–2634. URL: <https://aclanthology.org/2023.acl-long.147>. doi:10.18653/v1/2023.acl-long.147.
- [19] C. Li, J. Wang, K. Zhu, Y. Zhang, W. Hou, J. Lian, X. Xie, Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus, 2023.
- [20] O. Shaikh, H. Zhang, W. Held, M. Bernstein, D. Yang, On second thought, let’s not think step by step! bias and toxicity in zero-shot reasoning, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4454–4470. URL: <https://aclanthology.org/2023.acl-long.244>. doi:10.18653/v1/2023.acl-long.244.
- [21] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Curran Associates Inc., Red Hook, NY, USA, 2024.
- [22] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8086–8098. URL: <https://aclanthology.org/2022.acl-long.556>. doi:10.18653/v1/2022.acl-long.556.
- [23] C. Zhou, J. He, X. Ma, T. Berg-Kirkpatrick, G. Neubig, Prompt consistency for zero-shot task generalization, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2613–2626. URL: <https://aclanthology.org/2022.findings-emnlp.192>.

- doi:10.18653/v1/2022.findings-emnlp.192.
- [24] B. Wang, X. Deng, H. Sun, Iteratively prompt pre-trained language models for chain of thought, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2714–2730. URL: <https://aclanthology.org/2022.emnlp-main.174>. doi:10.18653/v1/2022.emnlp-main.174.
- [25] Z. Yin, Q. Sun, C. Chang, Q. Guo, J. Dai, X. Huang, X. Qiu, Exchange-of-thought: Enhancing large language model capabilities through cross-model communication, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 15135–15153. URL: <https://aclanthology.org/2023.emnlp-main.936>. doi:10.18653/v1/2023.emnlp-main.936.
- [26] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, S. Sitaram, MEGA: Multilingual evaluation of generative AI, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 4232–4267. URL: <https://aclanthology.org/2023.emnlp-main.258>. doi:10.18653/v1/2023.emnlp-main.258.
- [27] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, py-sentimiento: A python toolkit for opinion mining and social nlp tasks, 2023. arXiv:2106.09462.
- [28] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), 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>.
- [29] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- [30] R. Fernández-Iglesias, M. Fernandez-Pichel, M. Aragon, D. E. Losada, DepressMind: A depression surveillance system for social media analysis, in: N. Aletras, O. De Clercq (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 35–43. URL: <https://aclanthology.org/2024.eacl-demo.5>.

A. LLM Reasoning Prompt

A.1. Prompt template

In the experimental templates, we tried to set an objective environment for the answer given by the system role. The exact prompts used for the task are the following:

The first two prompts take as input *{tweet}*, and generate the *Positive* and *Negative Reasonings* are the following:

Positive Prompt.

System Role

You are a gender equality specialist. Think step by step. Answer as unbiased as possible.

User

Explain why the following tweet contains elements talking about sexism.

Tweet: {tweet}

Assistant:

{...}

Negative Prompt.

System Role

You are a gender equality specialist. Think step by step. Answer as unbiased as possible.

User

Explain why the following tweet does not contain elements talking about sexism.

Tweet: {tweet}

Assistant:

{...}

Positive and Negative Reasonings are limited to a maximum of 200 generated tokens. In a further step, the tweet and both the Positive and Negative Reasonings are provided so that the model self-evaluates both.

Comparison Prompt.

System Role

You are a gender equality specialist. Think step by step. Be concise.

User

Consider the following tweet.

Tweet: {tweet}

Which analysis is the most accurate?

Analysis 1: {positive reasoning}

Analysis 2: {negative reasoning}

Assistant:

{...}

The Comparison Stacked Reasoning is limited to a maximum of 250 tokens.

A.2. Answer samples

We show an example of the LLM output to illustrate the type of text reasoning we generated.

The original tweet is the following:

Writing a uni essay in my local pub with a coffee. Random old man keeps asking me drunk questions when I'm trying to concentrate; ends with "good luck, but you'll just end up getting married and not use it anyway". #everydaysexism is alive and well".

The positive and negative reasoning generated are:

Positive Reasoning:

Let's break down the tweet and identify the elements that suggest sexism:

- 1. The tweet starts by mentioning the author is writing a university essay in a pub, which is a common scenario. This sets the context for the rest of the tweet.*
- 2. The author mentions that a "random old man" is asking them "drunk questions" while they're trying to concentrate. This already has a slightly confrontational tone, as the author is implying that the man is interrupting and being a distraction.*

3. The phrase "good luck, but you'll just end up getting married and not use it anyway" is what specifically suggests sexism. The man's comment is making an assumption about the author's future plans and goals, which are likely to be academic in nature. The comment implies that the author's ultimate goal is to get married and abandon their academic pursuits, which is a stereotype often perpetuated about women.

This type of comment is a classic example of sexism.

Negative Reasoning:

The tweet does not contain elements of sexism because the content is focused on the author's experience of being interrupted by an old man while trying to write an essay, rather than any specific gender-based discrimination or harassment. The tweet is more about the author's frustration with the distraction and the old man's opinion about the author's future plans, which is not related to the author's gender. While the author uses the hashtag #everydaysexism, which is often used to describe the subtle and pervasive forms of sexism that people face in their daily lives, the tweet itself does not contain any explicit sexist language, attitudes, or behaviors. The tweet is more about the author's annoyance with the old man's behavior and the disruption it caused, rather than any gender-based issue. In fact, the tweet could be interpreted as an example of ageism, as the old man's comment about the author's future plans is not specific to the author's gender, but rather a stereotypical comment.

The comparison reasoning where the model self-evaluates the outputs produced is the following:

Stacked Comparison Reasoning:

Analysis 1 is the most accurate. The tweet contains elements of sexism because:

1. The old man's comment "good luck, but you'll just end up getting married and not use it anyway" is a stereotypical and gendered assumption about the author's future plans and goals, which are likely to be academic in nature. The comment implies that the author's ultimate goal is to get married and abandon their academic pursuits, which is a stereotype often perpetuated about women.
2. The tweet also contains a subtle tone of annoyance and frustration, which may be attributed to the author's experience as a woman, as they are more likely to face interruptions and condescending comments from men.

The other analyses do not accurately identify the elements of sexism in the tweet. Analysis 2 is incorrect because it does not recognize the gendered nature of the old man's comment and the impact it has on the author. Analysis 2 is also incorrect because it does not identify the specific elements of sexism in the tweet, instead, it focuses on broader issues of sexism and gender inequality. Analysis 1 accurately identifies the elements of sexism in the tweet, including the stereotypical and gendered assumption made by the old man.

It is essential to remark on some things. First, the model's answer quality varies with length. If we ask the model to generate positive reasoning with a smaller amount of tokens, the answer does not provide as many insights that support the target category. The following is an example of a positive reasoning limited to less than 50 tokens.

Positive Reasoning:

The tweet contains elements of sexism because it describes a situation where the author, a woman, is being subjected to unwanted and inappropriate comments from a stranger, specifically an older man, while she is trying to focus on her academic work.

As we can observe, limiting the reasoning to fewer tokens in the rationale prevents the model from fully breaking down the elements that it identifies as relevant. Allowing the model to produce a much longer output will result in repetition of the rationale generating.

Secondly, due to the safety settings of the model, in certain topics, the generated answer is as follows:

Positive Reasoning:

I cannot analyze a tweet that contains explicit and harmful content, including sexual harassment.

Table 4

Example of question refinement to get better answers.

Question	Version	Qualitative analysis
Does the tweet discredit feminism?	Original question	The model is biased to hallucinate reasons why it, in an implicit way, discredits feminism. It does not seem to distinguish the feminist movement by itself and rather settles the answer in the ideas feminism discusses.
Does the tweet discredit the feminist movement?	Modification 1 of the Question	Answers remained biased, but there is a correction to the object of study: the feminist movement itself.
Does the tweet directly discredit the feminist movement?	Modification 2 of the Question	Answer is less biased, reasons are not forced.

In the second step of our methodology, we expect that even though the model does not provide a rationale supporting the targeted position, the model is learning that this type of analysis also provides information about the nature of the tweet and helps to learn about the topics contained.

Finally, these are all automatic reasonings generated, and as shown in the example, are internally capable of supporting both sexist and not-sexist categories. We noticed that this two-step scheme is biased to predict as a final answer negative results. So, these models and the extracted knowledge shall be used carefully.

B. Sexist Evaluation Questionnaire

B.1. Refinement of questions

As we discussed before, refining the questions impact the accuracy of the answers. Table 4 shows an example of the qualitative analysis of the evolution of a question if asked to explain the answer provided.

B.2. List of questions

The complete list of fine-grained sexist questions is the following. The model was asked to answer with “Yes/No” for binary questions and only with the answer identifier for closed-option questions.

1. Does the tweet contain stereotypes or discrimination against women?
2. Choose only the best option:
 - a) The tweet is offensive to everyone
 - b) The tweet is offensive mostly to women
 - c) The tweet is not offensive
3. Choose only the best option:
 - a) The tweet talks about sexism and is sexist.
 - b) The tweet talks about sexism but is not sexist
 - c) The tweet is not talking about sexism
4. Choose only the best option:
 - a) The tweet is not related to sexism.
 - b) The tweet is describing or reporting a sexist situation suffered by a woman.
 - c) The tweet criticizes a sexist behavior
 - d) The tweet is expressing a sexist message.
5. Does the tweet directly discredit the feminist movement?

6. Does the tweet devalue women's struggles?
7. Does the tweet deny the existence of gender inequality?
8. Does the tweet portray men as victims of gender oppression?
9. Does the tweet treat being male as the default or norm?
10. Does the tweet imply men are superior to women?
11. Does the tweet suggest women are unsuitable for certain tasks?
12. Does the tweet suggest traits or abilities are determined by gender?
13. Does the tweet objectify or dehumanize women?
14. Does the tweet reinforce traditional gender roles for women?
15. Does the tweet contain sexual references?
16. Does the tweet contain sexual harassment towards women?
17. Does the tweet express hatred or misogyny towards women?
18. Does the tweet include threats of violence against women?
19. Does the tweet use gendered insults or slurs?