RMIT-IR at EXIST Lab at CLEF 2024

Notebook for the EXIST Lab at CLEF 2024

Tony Kim Smith[1,†], H Ruda Nie[2,†], Johanne R. Trippas[1] and Damiano Spina[1,*]

[1]RMIT University, Melbourne, Australia
[2]Tay Nguyen University, Buon Ma Thuot, Vietnam

Abstract

This paper describes RMIT-IR team's participation in the EXIST Lab at CLEF 2024. The proposed approaches aim to address sexism characterization on microblog posts (Tasks 1, 2, and 3) and sexism identification on memes (Task 4). For Tasks 1–3, we studied the effectiveness of zero-shot In-Context Learning (ICL) [1] with off-the-shelf pre-trained Large Language Models (LLMs) to mimic the scenario of minimal intervention of a practitioner aiming to build sexism characterization systems. Our approaches for meme classification (Task 4) utilize CLIP (Contrastive Language-Image Pre-training) [2] to experiment with multi-modal embeddings and zero-shot sexism identification models. We report the performance of our approaches under the learning with disagreements regime (Soft evaluation) and also for label predictions (Hard evaluation). The code of our submission is available at https://github.com/rmit-ir/exist2024/.

Warning: Some of the examples included in this paper may contain offensive language and explicit descriptions of sexist behavior, which may be disturbing to the reader.

Keywords

sexism characterization, large language models, in-context learning, multi-modal contrastive learning

1. Introduction

Social media has had a considerable impact on human societies. Applications such as Facebook, YouTube, Instagram, and TikTok have helped move the zeitgeist while creating large communities in the millions. However, many social media platforms have issues with people creating and posting harmful information. The challenge of detecting and managing such harmful content remains a concern for social media companies, contributing to consequences ranging from misinformation to adverse effects on mental health [3]. In addition, the rise of social media has empowered influencers who often unwittingly or deliberately propagate harmful stereotypes and negative gender norms. This type of content attracts an audience and drives advertising revenue, perpetuating a cycle of negativity [4]. As a result, it often fosters negative behaviour towards women and minority groups, impacting many people negatively [5].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France
*Corresponding author.
†These authors contributed equally.
s3972733@student.rmit.edu.au (T. K. Smith); rudadhtn89@gmail.com (H. R. Nie); j.trippas@rmit.edu.au (J. R. Trippas); damiano.spina@rmit.edu.au (D. Spina)
https://www.johannetrippas.com (J. R. Trippas); https://www.damianospina.com (D. Spina)
0009-0009-4752-6679 (T. K. Smith); 0000-0002-1194-2496 (H. R. Nie); 0000-0002-7801-0239 (J. R. Trippas); 0000-0001-9913-433X (D. Spina)

Sexism is the belief that the members of one sex or gender are less than the members of the other sex, especially that women are less able than men [6]. This can be categorized into hostile sexism and benevolent sexism.

Sexism can limit the opportunities and roles people of different sexes and genders are expected to take. It can be conveyed through any form of expression, like images, cartoons, memes, objects, gestures, and symbols, and can be spread offline or online. This oppression can take different forms, such as economic exploitation and social domination [7].

Sexist attitudes and behaviours can perpetuate stereotypes of social and gender roles based on one's biological sex. Usually, people are socialized with sexist concepts that teach traditional gender roles for males and females [8]. Hostile sexism represents a form of sexist ideology, marked by explicit hostility towards women and the perception of them as inferior and submissive[9]. This deeply ingrained perception often results in the mistreatment of women at both individual and institutional levels [8]. Benevolent sexism is a nuanced manifestation that ingrains in men the belief that they should be responsible for providing for women in intimate relationships [9]. This belief system dictates specific roles and behaviours for women, such as expecting them to demonstrate motherly instincts, subtly reinforcing traditional gender roles. A society that has high rates of hostile and benevolent sexism often has high rates of violence against women, such as domestic violence, rape, and the commodification of women and their bodies [10, 11].

There has been a recent increase in research on identifying different forms of hate speech, corresponding with advancements in generative pre-trained transformers and, in general, large language models (LLMs) [12]. Researchers are asking how LLMs can be trained to identify subtle and overt sexist content [13, 14, 15]. However, many questions on how state-of-the-art LLMs can be used for sexism detection are still open. What criteria should be used to evaluate what constitutes sexism in varied cultural contexts? If a dataset with binary classifications[1] is employed, can a machine learning model accurately capture the nuances within the text? And how do we address the evolution of language with new slang and phrases continually emerging? These questions highlight the complexity of sexism detection. The cost and technical skills required to create a system that incorporates LLMs that can identify sexism make it unattainable for most individuals. We aim to simplify the process using pre-trained LLMs and prompts to address the EXIST lab tasks of classifying and labelling tweets.

In addition to the text classification in Tasks 1–3, we address the problem of identifying sexism in multi-modal formats for Task 4. Memes — ideas, images, or videos that are spread very quickly on the internet [16] — exist not only in text form but also include any accompanying images. Therefore, combining text and the attached image (i.e., making the input multi-modal) can be more conducive to identifying whether a meme is sexist. Multi-modal models are usually proposed to deal with multi-modal datasets for classification tasks. Among existing multi-modal systems, Contrastive Language-Image Pre-Training (CLIP) [2] is a powerful vision-and-language (VL) pre-trained model that can directly learn raw text about images. In addition, CLIP has the ability to map data of different modalities, text and images into a shared embedding space. Hence, CLIP has been shown to be a powerful tool for zero-shot image and text classification [2]. Furthermore, CLIP can be beneficial for image-text feature fusion, which

[1]We acknowledge that the classification of sex and gender into two categories is a simplification of people's identities.

can boost model performance on natural language processing (NLP) downstream tasks such as text classification [17] and multi-modal sarcasm detection [18]. Motivated by the success of CLIP on various VL downstream tasks, this study aims to investigate the following research questions for Task 4:

- How effective is CLIP for zero-shot sexism identification?
- How can the naturally inherited multi-modal knowledge from pre-trained CLIP be extracted to identify sexism effectively?

Addressing the first research question, we proposed *Prompt-CLIP* for zero-shot sexism identification. For the latter question, we employed CLIP to perform supervised sexism classification. Inspired by the impressive performance of multi-view CLIP for sarcasm detection in a previous study [18], we adopted multi-view CLIP for supervised sexism classification, namely, text-image multi-view CLIP (TIMV-CLIP) and proposed text-image multi-modal models via CLIP-Guided Learning (TI-CLIP) as a baseline.

The paper is organized as follows. Details about the tasks participated in are described in Section 2. Section 3 provides details about the proposed approaches. In Section 4, we provide and discuss the results. Finally, we conclude in Section 5.

2. Tasks Addressed

The sEXism Identification in Social neTworks (EXIST) [19] lab at the Conference and Labs of the Evaluation Forum (CLEF) 2024 [20] aims to identify and characterize sexism using the learning with disagreements paradigm [21, 22, 23]. This edition of the EXIST lab consists of sexism characterization on microblog posts (tweets) and memes.

2.1. Tasks 1–3: Sexism Characterization of Microblog Posts

- **Task 1:** Addresses sexism identification in tweets as a binary classification, requiring the system to classify whether a tweet is sexist (YES) or not (NO).
- **Task 2:** Focuses on determining the source intention in tweets as a multi-class classification, requiring the system to classify the tweet's intention as *Direct*, *Reported*, or *Judgemental*.
- **Task 3:** Involves sexism categorization in tweets as a multi-label classification, requiring the system to classify tweets into categories such as *Ideological Inequality*, *Stereotyping Dominance*, *Objectification*, *Sexual Violence*, and *Misogyny-Non-Sexual Violence*.

2.2. Task 4: Sexism Identification of Memes

While the above tasks address sexism identification in text, Task 4 deals with multi-modal input. Task 4 aims to address sexism identification as a binary classification, requiring the systems to classify whether a given meme is sexist or not.

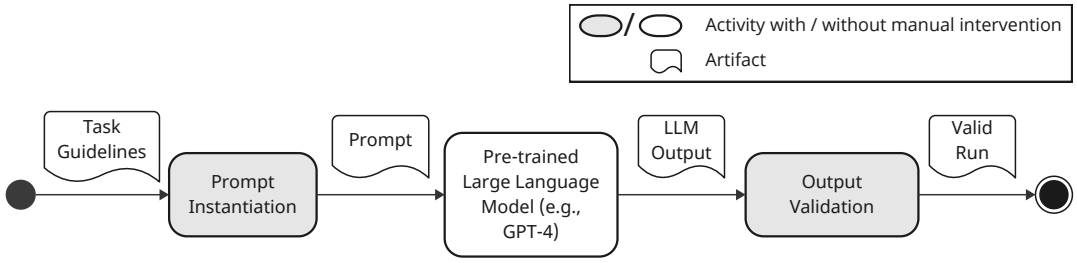

Figure 1: Overview of the workflow to create our unsupervised in-context learning runs for sexism characterization on microblog posts (Tasks 1–3).

2.3. Evaluation approaches

- **Soft-Soft Evaluation:** For systems that produce probabilities for each category, soft-soft evaluation is provided to compare the probabilities assigned by the systems with those assigned by the set of human annotators. The official evaluation metric is ICM-soft [24, 23]. Additionally, Cross Entropy is also reported.

- **Hard-Hard Evaluation:** Hard labels are derived from the different annotators' labels through a probabilistic threshold computed for each task. Hard-hard evaluation is provided to evaluate systems that return Hard labels as output by comparing against a ground truth that combines multiple annotations into one. The original ICM [25] and $F1$ score are used as evaluation metrics.

3. Proposed Approaches

3.1. Unsupervised In-Context Learning for Sexism Characterization in Microblog Posts

Our goal was to examine the procedure of developing a functional solution with readily available LLMs while minimizing the manual effort required from the practitioner. As shown in Figure 1, the basic architecture involves giving the researcher a set of labeling or classification tasks and asking the LLM to generate an accurate output. To ensure that the responses followed the predefined criteria for each task, the outputs were systematically stored in a JSON format and manually inspected for errors in the "value" field. Responses such as "YES", "YES", or variations with additional text or punctuation like "Yes, the ... is sexist" required manual corrections to conform to the expected format. Incidences of token limit rates that resulted in "HTTP" errors were addressed by re-running the task for the affected tweet using its unique ID. These occurrences were uncommon, making manual correction a more efficient solution than an automated task given the time constraint.

The prompts used for runs submitted to Tasks 1, 2, and 3 were designed with multiple parts:

- **Definition of the underlying concept being addressed in the task (e.g., *sexism*):** Sexism, prejudice or discrimination based on sex or gender, especially against women and

girls. Although its origin is unclear, the term sexism emerged from the "second-wave" feminism of the 1960s through '80s and was most likely modeled on the civil rights movement's term racism (prejudice or discrimination based on race). Sexism can be a belief that one sex is superior to or more valuable than another sex. It imposes limits on what men and boys can and should do and what women and girls can and should do. The concept of sexism was originally formulated to raise consciousness about the oppression of girls and women, although by the early 21st century it had sometimes been expanded to include the oppression of any sex, including men and boys, intersex people, and transgender people.

- **Instruction to Address Task and to Obtain Consistent Outputs:** You are a robot who detects sexism from text given in the prompt.

- **Perspectivism:**
  - **Level of Education:** For each response, consider the perspective of individuals representing the following study levels: [study_levels_annotators]

  - **Level of Education and Gender:** For each response, consider the perspective of individuals representing the following study levels: *[study_levels_annotators]* and gender: *[gender_annotators]*.

- **Output Format:**
  - **Task 1 (Soft):** Give me 6 answers with NO or YES. Format: [NO], [YES]

  - **Task 1 (Hard)**: Give me 1 answer with [NO] or [YES]

  - **Task 2 (Soft)**: Give me 6 answers with NO, DIRECT, REPORTED or JUDGEMENTAL using commas for each answer. Example: [NO], [DIRECT], [REPORTED], [JUDGEMENTAL], [JUDGEMENTAL], [NO]

  - **Task 2 (Hard):** Give me 1 answer with NO, DIRECT, REPORTED or JUDGEMENTAL using commas for each answer. Example: [NO], [DIRECT], [REPORTED], [JUDGEMENTAL], [JUDGEMENTAL], [NO]

  - **Task 3 (Soft):** Give me 6 answers with NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, or MISOGYNY-NON-SEXUAL-VIOLENCE using commas for each answer. Example: [NO], [IDEOLOGICAL-INEQUALITY], [STEREOTYPING-DOMINANCE], [OBJECTIFICATION], [SEXUAL-VIOLENCE], [MISOGYNY-NON-SEXUAL-VIOLENCE]

  - **Task 3 (Hard):** Give me 1 answers with NO, IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, or MISOGYNY-NON-SEXUAL-VIOLENCE using commas for each answer. Example: [NO], [IDEOLOGICAL-INEQUALITY], [STEREOTYPING-DOMINANCE], [OBJECTIFICATION], [SEXUAL-VIOLENCE], [MISOGYNY-NON-SEXUAL-VIOLENCE]

- **Instance to Classify:** #### *[tweet]* ####

An example of a prompt submitted and output obtained from gpt-4-turbo to classify instance with id_EXIST 600090 using RMIT-IR_3 for Task 2 (Soft):

- **Input:** Sexism, prejudice or discrimination based on sex or gender, especially against women and girls. Although its origin is unclear, the term sexism emerged from the "second-wave" feminism of the 1960s through '80s and was most likely modeled on the civil rights movement's term racism (prejudice or discrimination based on race). Sexism can be a belief that one sex is superior to or more valuable than another sex. It imposes limits on what men and boys can and should do and what women and girls can and should do. The concept of sexism was originally formulated to raise consciousness about the oppression of girls and women, although by the early 21st century it had sometimes been expanded to include the oppression of any sex, including men and boys, intersex people, and transgender people. You are a robot who detects sexism from text given in the prompt. For each response, consider the perspective of individuals representing the following study levels: ["High school degree or equivalent", "Bachelor's degree", "Bachelor's degree", "Bachelor's degree", "Bachelor's degree", "High school degree or equivalent"]. Give me 6 answers with NO, DIRECT, REPORTED or JUDGEMENTAL using commas for each answer. Example: [NO], [DIRECT], [REPORTED], [JUDGEMENTAL], [JUDGEMENTAL], [NO].
  #### Girls, don't let anyone ever tell you, you're not as good as a man #gender #girlpower #equity ####

- **Output:** [NO],[NO],[NO],[NO],[NO],[NO]

To find the distribution of responses, we initially tried to use GPT to figure out the likelihood percentage. Unfortunately, GPT only gave absolute values (either 100 or 0) or a consistent split of 70/30 most of the time. We directed the model to generate six responses for each tweet, which matched the number of annotators per tweet. For example, a set of responses like "YES", "NO", "YES", "NO", "YES", and "NO" would result in a calculated distribution of 50%. Additionally, in our final submissions, experimental runs two and three included prompts that provided additional context, such as the annotators' gender or educational backgrounds. This was done to see if providing relevant background information would improve the LLM's ability to predict annotator responses. The formats for each task can be seen above along with an example prompt used in RMIT-IR_3 for Task 2 (Soft).

3.1.1. Runs Submitted to Tasks 1–3

We used OpenAI's API to submit prompts to the pre-trained model `gpt-4-turbo-2024-04-09` [26]. For each tweet in the test set, we instantiated the prompt from above by appending the textual content of the instance. We used the syntax `#### [tweet] ####` to provide explicit delimiters to the model. For Soft tasks, we asked for six instances and then created a distribution based on the frequency of the predicted labels.

We experimented with multiple versions of prompt templates using the development set supplied by the EXIST organizers (we did not use the training set). We found that the following

Table 1
Summary of the runs submitted to Tasks 1–3. The Output Format was according to the type of task (Soft and Hard) as detailed previously. The instance to classify was appended at the end of the prompt.

Run	Definition	Instruction	Perspectivism	Output Format
RMIT-IR_1	✓	✓	–	✓
RMIT-IR_2	✓	✓	Level of Education	✓
RMIT-IR_3	✓	✓	Level of Education + Gender	✓

elements were especially effective in directing the model to concentrate on the specific task and to ensure the responses were properly formatted (i.e., single-word answers and capitalized):

- Employing a role-playing technique of framing the task with the prompt *"You are a robot who detects sexism from text given in the prompt."*
- Giving explicit formatting instructions such as *"Give me 6 answers with NO or YES. Format: [NO], [YES]"*.

3.2. Multi-modal Contrastive Learning for Sexism Identification on Memes

Inspired by the successful applications of CLIP [2] for NLP [17] and computer vision tasks [27], [28], we adopted CLIP for the sexism identification task (Task 4). Unlike conventional methods that rely heavily on labelled image-text pairs, CLIP is a cross-modality model pre-trained with 400M noisy image-text pairs collected from the internet to learn high-level semantic features. CLIP consists of two encoders that embed texts and images into a uniform mathematical space. Then, for the matched image-text pair, CLIP is encouraged to maximize the cosine similarity between the embedding of the two modalities. Otherwise, the similarity is minimized for the model to find the most suitable paired images and texts. Our motivation for using CLIP-based learning for sexism identification is to capture cross-modal ambiguity by explicitly measuring the correlation between texts and images of targeted memes and to guide the feature-fusing and decision-making stages.

We propose two supervised contrastive learning models based on CLIP: Text-Image multi-modal model via CLIP-guided learning (TI-CLIP) and Text-Image Multi-View multi-modal model via CLIP-guided learning (TIMV-CLIP). The architecture of TIMV-CLIP is shown in Figure 2. We also propose *Prompt-CLIP* to address zero-shot sexism classification and CLIP-based models for supervised sexism classification.

- **TI-CLIP:** The overall architecture of TI-CLIP consists of two feature encoding models used to encode texts and images. These embeddings are then combined into a multi-modal embedding before passing into a feedforward network for sexism classification.
- **TIMV-CLIP:** We adopted a novel multi-view CLIP framework (MV-CLIP) [18] for sexism identification, namely TIMV-CLIP (Figure 2). In addition to encoding image and text as TI-CLIP, TIMV-CLIP further considers modelling relationships across text and image

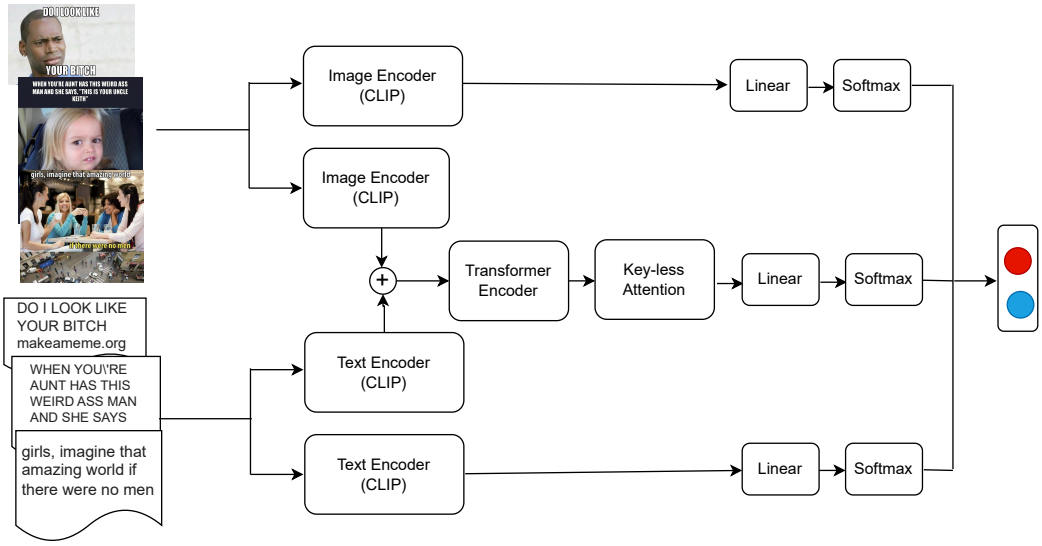

Figure 2: The architecture of our proposed multi-modal model TIMV-CLIP for supervised sexism identification on memes (Task 4).

modality using a transformer encoder, which aims to capture the interaction across different modalities. Unlike MV-CLIP in a previous study [18], TIMV-CLIP employs BERT Base Multilingual (mBERT) [29] to encode texts.

- **Prompt-CLIP:** Prompt-CLIP performs zero-shot sexism identification. Prompt-CLIP uses a pre-trained CLIP model to create a custom classifier without training and considers images as inputs. It further encodes pre-defined classes (sexism and not sexism) with more description, known as prompts, into a learned latent space, and compares their similarity to the image latent space. In this study, we used *"an image contains no information about sexism"* and *"an image contains information about sexism and against women"* as prompts for Prompt-CLIP. The pre-trained text encoder transforms the class names (e.g., prompts) into a text embedding vector, while the pre-trained Image Encoder embeds the image.

- **Model Training:** We first randomly split the training subset into training (80%) and validation (20%) for cross-validation purposes. We implemented TI-CLIP and TIMV-CLIP based on the Hugging Face library [30] and adopted `clip-vit-base-patch32` as the backbone. Both TI-CLIP and TIMV-CLIP were trained directly with Soft labels. We use Adam as an optimizer to optimize the parameters in both TI-CLIP and TIMV-CLIP models. After several trials with other hyperparameters, we selected the parameters that performed best on the validation set. Specifically, the batch size is 32. The learning rate for CLIP is 1e-6 and for the other parts is 5e-4. Finally, we use the dropout percentage of 0.3 and train the models for 10 epochs.

Table 2
Runs submitted to Task 4: sexism identification on memes.

Run	Model
RMIT-IR_1	TI-CLIP (feedforward network)
RMIT-IR_2	TIMV-CLIP (Transformer encoder)
RMIT-IR_3	Prompt-CLIP (zero-shot)

3.2.1. Runs Submitted to Task 4

The proposed multi-modal sexism identification models mainly focus on Soft label predictions. For the Hard submissions, hard labels are directly assigned by applying the max function, i.e., based on the highest probability score.

Table 2 presents the submitted runs for Task 4, which can be summarized as follows:

- **RMIT-IR_1:** For the first submission, the trained TI-CLIP model was used to predict whether given memes are sexist or not sexist.
- **RMIT-IR_2:** We used the trained TIMV-CLIP to generate the second submission.
- **RMIT-IR_3:** Prompt-CLIP was used to predict Soft and Hard labels for the third submission.

4. Results and Discussion

4.1. Tasks 1–3

The performance of our proposed approaches for Tasks 1, 2, and 3 are presented in Tables 3, 4, and 5, respectively. As shown in Table 3, simplifying the architecture required for creating a LLM that can identify sexism shows promise, as evidenced by the classification of English and Spanish tweets. Although the current model achieved a 49% ICM-Soft score, which is 17% lower than the best-performing run, this result indicates the potential to use prompting for classification tasks.

The cost of this process, which included testing with a development dataset and submitting with a gold dataset, was close to $150 AUD. In particular, it did not require knowledge of cloud computing, expensive hardware, or much energy. The average time taken to produce an output was about 90 minutes.

Table 3
Results of the proposed approaches for Task 1 (Soft).

	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
All Test Instances (English + Spanish)				
EXIST2024_gold	0	3.1182	1.0000	0.5472
EXIST2024_majority_class	36	−2.3585	0.1218	4.6115
EXIST2024_minority_class	40	−3.0717	0.0075	5.3572
Task-specific Prompt (RMIT-IR_1)	23	−0.0011	0.4998	2.7892
Task-specific + Perspectivism (Education) (RMIT-IR_2)	31	−0.3941	0.4368	2.9956
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	26	−0.3016	0.4516	2.8235
English Test Instances				
EXIST2024_gold	0	3.1141	1.0000	0.5770
EXIST2024_majority_class	36	−2.1991	0.1469	4.2166
EXIST2024_minority_class	40	−3.8158	0.0000	5.7521
Task-specific Prompt (RMIT-IR_1)	26	−0.2873	0.4539	2.8722
Task-specific + Perspectivism (Education) (RMIT-IR_2)	30	−0.5951	0.4044	2.9824
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	28	−0.4949	0.4205	2.8222
Spanish Test Instances				
EXIST2024_gold	0	3.1177	1.0000	0.5208
EXIST2024_majority_class	36	−2.5421	0.0923	4.9631
EXIST2024_minority_class	37	−2.5742	0.0872	5.0055
Task-specific Prompt (RMIT-IR_1)	22	0.1840	0.5295	2.7153
Task-specific + Perspectivism (Education) (RMIT-IR_2)	28	−0.2800	0.4551	3.0074
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	27	−0.1908	0.4694	2.8247

Looking at Table 3, Task 1, the use of the Task-specific Prompt yielded an ICM-Soft Norm score of 49%, securing the 23rd position overall. It is interesting to note that the inclusion of additional clues such as the annotator's education level and gender did not bolster performance; instead, it diminished the score. In particular, when looking at the Spanish test instances, the model scored higher across all Spanish test instances compared to all English test instances, despite being trained on an English-based GPT. This underscores the robust cross-lingual applicability of the model, showcasing its proficient handling of Spanish data despite its primary training on English.

Table 4
Results of the proposed approaches for Task 2 (Soft).

	Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy
All Test Instances (English + Spanish)				
EXIST2024_gold	0	6.2057	1.0000	0.9128
EXIST2024_majority_class	27	−5.4460	0.0612	4.6233
EXIST2024_minority_class	35	−32.9552	0.0000	8.8517
Task-specific Prompt (RMIT-IR_1)	23	−4.5481	0.1336	3.5776
Task-specific + Perspectivism (Education) (RMIT-IR_2)	33	−6.1535	0.0042	4.0930
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	29	−5.7632	0.0357	3.9903
English Test Instances				
EXIST2024_gold	0	6.1178	1.0000	0.9354
EXIST2024_majority_class	25	−5.2028	0.0748	4.2291
EXIST2024_minority_class	35	−39.4948	0.0000	8.9579
Task-specific Prompt (RMIT-IR_1)	22	−4.2180	0.1553	3.3660
Task-specific + Perspectivism (Education) (RMIT-IR_2)	30	−6.7055	0.0000	3.9259
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	29	−6.1443	0.0000	3.8353
Spanish Test Instances				
EXIST2024_gold	0	6.2431	1.0000	0.8926
EXIST2024_majority_class	30	−5.6674	0.0461	4.9745
EXIST2024_minority_class	35	−28.7093	0.0000	8.7570
Task-specific Prompt (RMIT-IR_1)	26	−4.8962	0.1079	3.7660
Task-specific + Perspectivism (Education) (RMIT-IR_2)	33	−6.0168	0.0181	4.2418
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	32	−5.7527	0.0393	4.1283

In Task 2, our analysis revealed challenges in multi-class classification as shown in Table 4. The approach yielded a 13% ICM-Soft Norm score, indicating considerable difficulty in discerning the intention of the tweets. The introduction of additional clues, such as the annotator's education level, generally led to a decline in performance. However, adding gender information resulted in a slight improvement, elevating the score from almost 0% to 3%. The results indicated that the approach performed more effectively in English without additional clues; however, its performance diminished once clues were introduced. Conversely, our analysis demonstrated that the GPT model exhibited greater efficacy with clues in Spanish, suggesting potential advantages in providing contextual information in non-English scenarios.

Table 5
Results of the proposed approaches for Task 3 (Soft).

	Rank	ICM-Soft	ICM-Soft Norm
All Test Instances (English + Spanish)			
EXIST2024_gold	0	9.4686	1.0000
EXIST2024_majority_class	28	−8.7089	0.0401
EXIST2024_minority_class	33	−46.1080	0.0000
Task-specific Prompt (RMIT-IR_1)	19	−7.2098	0.1193
Task-specific + Perspectivism (Education) (RMIT-IR_2)	21	−7.8944	0.0831
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	27	−8.5680	0.0476
English Test Instances			
EXIST2024_gold	0	9.1255	1.0000
EXIST2024_majority_class	25	−8.2105	0.0501
EXIST2024_minority_class	33	−46.9473	0.0000
Task-specific Prompt (RMIT-IR_1)	20	−7.8798	0.0683
Task-specific + Perspectivism (Education) (RMIT-IR_2)	28	−9.3039	0.0000
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	29	−10.4428	0.0000
Spanish Test Instances			
EXIST2024_gold	0	9.6071	1.0000
EXIST2024_majority_class	28	−9.0314	0.0300
EXIST2024_minority_class	33	−45.4260	0.0000
Task-specific Prompt (RMIT-IR_1)	19	−6.7226	0.1501
Task-specific + Perspectivism (Education) (RMIT-IR_2)	20	−6.8696	0.1425
Task-specific + Perspectivism (Edu + Gender) (RMIT-IR_3)	21	−7.1653	0.1271

Task 3 was also challenging with multi-label classification. The initial ICM-Soft Norm score, as shown in Table 5, stood at 11%. Following a similar trend to Task 2, the integration of clues such as the level of education of the annotator, resulted in a reduction in performance to 8%. Subsequently, when both education and gender clues were included, performance decreased further to 4%. The English scores are quite similar to Task 2. Notably from the initial performance on the Spanish dataset exceeded that of Task 2, but declined with the addition of educational clues and further declined with the incorporation of both education and gender clues. This observation underscores a consistent pattern of diminishing returns with the incorporation of more specific annotator information.

Our proposed approaches for the second and third runs involve implementing few-shot and in-context learning. The experiments for these runs were conducted using gpt-4-turbo, and future tests should include gpt-4o along with other pre-trained LLMs to determine their efficacy in this context. Our experiment results show that involving few-shot and in-context learning does not improve model performance on sexism identification in tweets (as shown in Tables 3–5). Although prompting requires less coding and understanding of LLMs, producing the exact desired response 100% of the time was challenging. The prompts had to be carefully designed to ensure that the GPT provided a single and consistent answer, especially when dealing with distribution. Although pre-processing text for an LLM is more complex than pre-processing answers from GPTs, ensuring the response is in the correct format is simpler.

Table 6
Results of the proposed approaches for Task 4 (Soft and Hard).

	Rank (Soft)	ICM-Soft	ICM-Soft Norm	Cross Entropy	ICM-Hard	ICM-Hard Norm	$F1_{YES}$
All Test Instances (English + Spanish)							
EXIST2024_gold	0	3.1107	1.0000	0.5852	0.9832	1.0000	1.0000
EXIST2024_majority_class	36	−2.3568	0.1212	4.4015	−0.4038	0.2947	0.6821
EXIST2024_minority_class	38	−3.5089	0.0000	5.5672	−0.6468	0.1711	0.0000
TI-CLIP (RMIT-IR_1)	29	−1.2819	0.2940	1.0128	−0.6468	0.1711	0.0000
TIMV-CLIP (RMIT-IR_2)	8	−0.3780	0.4392	0.9852	−0.0123	0.4938	0.6726
Prompt-CLIP (RMIT-IR_3)	24	−1.0894	0.3249	1.1206	−0.2601	0.3677	0.6040
English Test Instances							
EXIST2024_gold	0	3.0794	1.0000	0.5528	0.9848	1.0000	1.0000
EXIST2024_majority_class	34	−2.2236	0.1390	4.4798	−0.4076	0.2931	0.6880
EXIST2024_minority_class	36	−3.1235	0.0000	5.4888	−0.6381	0.1761	0.0000
TI-CLIP (RMIT-IR_1)	33	−1.2889	0.2907	1.0115	−0.6381	0.1761	0.0000
TIMV-CLIP (RMIT-IR_2)	1	−0.0011	0.4998	0.9243	0.1536	0.5780	0.7250
Prompt-CLIP (RMIT-IR_3)	25	−1.0106	0.3359	1.1316	−0.2089	0.3940	0.5641
Spanish Test Instances							
EXIST2024_gold	0	3.1360	1.0000	0.6160	0.9815	1.0000	1.0000
EXIST2024_majority_class	36	−2.4997	0.1014	4.3270	−0.4001	0.2962	0.6765
EXIST2024_minority_class	38	−3.9408	0.0000	5.6416	−0.6557	0.1660	0.0000
TI-CLIP (RMIT-IR_1)	29	−1.2730	0.2970	1.0141	−0.6557	0.1660	0.0000
TIMV-CLIP (RMIT-IR_2)	17	−0.7851	0.3748	1.0431	−0.1762	0.4103	0.6192
Prompt-CLIP (RMIT-IR_3)	27	−1.1903	0.3102	1.1101	−0.3140	0.3400	0.6332

Another unexpected aspect of this architecture was the ability to assist the GPT with hints. We tested how adding biases by including the annotator's education level and gender affected its ability to classify or label tweets. This gave insights into how such biases can influence model performance and classification accuracy.

4.2. Task 4

Table 6 present the results of the proposed approaches for Task 4. Among the proposed approaches, TIMV-CLIP performs best in all cases (English+Spanish, English, or Spanish test instances) considering both Soft and Hard evaluation scenarios. This indicates the importance of effectively utilizing deep interactions between texts and images of memes with CLIP. Furthermore, TIMV-CLIP achieved the best-performance model (RMIT-IR_2) on English test instances with an ICM-Soft Norm score of 0.4998, ranked first in the leaderboard considering the Soft evaluation (English test instances). This observation confirms the advantages of CLIP for text-image pair classification tasks. However, the performance of TIMV-CLIP has dropped in Spanish test instances, which leads to lower performance in all test instances (Spanish test instances). We believe using a translation component for Spanish text in memes could lead to better overall performance.

5. Conclusions

This paper proposed unsupervised in-context learning with off-the-shelf pre-trained LLMs to address sexism characterization on microblog posts (Tasks 1, 2, and 3). Dealing with multi-modal inputs, we proposed multi-modal contrastive learning, including Prompt-CLIP, TI-CLIP, and TIMV-CLIP for sexism identification in memes (Task 4).

The results of our experiment demonstrated the effectiveness of TIMV-CLIP under the Learning with Disagreements regime, indicating the need to consider capturing sexism cues from different perspectives, including image, text, and image-text interactions.

Future work includes further experimentation with unsupervised In-Context Learning in other tasks or meta-tasks such as MonsterCLEF [31], and the inclusion of machine translation for multi-modal contrastive learning.

Acknowledgments

This research has been carried out in the unceded lands of the Woi Wurrung and Boon Wurrung peoples of the eastern Kulin Nation. We pay our respects to their Ancestors and Elders, past, present, and emerging. This research is partially supported by the Australian Research Council (ARC, project nr. DE200100064 and CE200100005).

References

[1] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, Z. Sui, A Survey on In-context Learning, 2023. arXiv:2301.00234.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[3] L. Braghieri, R. Levy, A. Makarin, Social Media and Mental Health, American Economic Review 112 (2022) 3660–3693. doi:10.1257/aer.20211218.

[4] R. Young, V. Kananovich, B. G. Johnson, Young Adults' Folk Theories of How Social Media Harms Its Users, Mass Communication and Society 26 (2023) 23–46. doi:10.1080/15205436.2021.1970186.

[5] S. R. Stephanie Wescott, X. Zhao, The Problem of Anti-feminist 'Manfluencer' Andrew Tate in Australian Schools: Women Teachers' Experiences of Resurgent Male Supremacy, Gender and Education 36 (2024) 167–182. doi:10.1080/09540253.2023.2292622.

[6] Cambridge Dictionary, Definition of sexism, https://dictionary.cambridge.org/dictionary/english/sexism, 2024. Accessed: 2024-07-04.

[7] M. McIntosh, The State and the Oppression of Women, in: Feminism and Materialism (RLE Feminist Theory), Routledge, 2013, pp. 254–289.

[8] G. Masequesmay, Sexism | Definition, Types, Examples, & Facts | Britannica, 2024. URL: https://www.britannica.com/topic/sexism.

[9] P. Glick, S. T. Fiske, Ambivalent sexism, in: Advances in Experimental Social Psychology, volume 33, Academic Press, 2001, pp. 115–188. doi:10.1016/S0065-2601(01)80005-8.

[10] K. R. Blake, S. M. O'Dean, J. Lian, T. F. Denson, Misogynistic Tweets Correlate with Violence against Women, Psychological science 32 (2021) 315–325. doi:10.1177/0956797620968529.

[11] K. Barker, O. Jurasz, Online Misogyny, Journal of International Affairs 72 (2019) 95–114.
[12] E. Hashmi, S. Y. Yayilgan, Multi-class Hate Speech Detection in the Norwegian Language Using FAST-RNN and Multilingual Fine-Tuned Transformers, Complex & Intelligent Systems 10 (2024) 4535–4556. doi:10.1007/s40747-024-01392-5.
[13] J. A. García-Díaz, R. Pan, R. Valencia-García, UMUTeam at EXIST 2023: Sexism Identification and Categorisation Fine-tuning Multilingual Large Language Models, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, 2023, pp. 985–999. URL: https://ceur-ws.org/Vol-3497/paper-080.pdf.
[14] H. Abburi, P. Parikh, N. Chhaya, V. Varma, Multi-task Learning Neural Framework for Categorizing Sexism, Comput. Speech Lang. 83 (2024). doi:10.1016/j.csl.2023.101535.
[15] L. Plaza, J. Carrillo-de Albornoz, R. Morante, J. Gonzalo, E. Amigó, D. Spina, P. Rosso, Overview of EXIST 2023: sEXism Identification in Social neTworks, in: Proceedings of ECIR'23, 2023, pp. 593–599. doi:10.1007/978-3-031-28241-6_68.
[16] Cambridge Dictionary, Definition of meme, https://dictionary.cambridge.org/dictionary/english/meme, 2024. Accessed: 2024-07-04.
[17] L. Qin, W. Wang, Q. Chen, W. Che, CLIPText: A New Paradigm for Zero-shot Text Classification, 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. 1077–1088. doi:10.18653/v1/2023.findings-acl.69.
[18] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, R. Xu, MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System, 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. 10834–10845. doi:10.18653/v1/2023.findings-acl.689.
[19] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, EXIST: sEXism Identification in Social neTworks, http://nlp.uned.es/exist2024/, 2024. Accessed: 2024-07-04.
[20] Conference and Labs of the Evaluation Forum (CLEF), https://www.clef-initiative.eu/, 2024. Accessed: 2024-07-04.
[21] L. Plaza, J. Carrillo-de Albornoz, E. Amigó, J. Gonzalo, R. Morante, P. Rosso, D. Spina, B. Chulvi, A. Maeso, V. Ruiz, EXIST 2024: sEXism Identification in Social neTworks and Memes, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, p. 498–504. doi:10.1007/978-3-031-56069-9_68.
[22] 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.
[23] 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.
[24] 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 (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, pp. 813–854. URL: https://ceur-ws.org/Vol-3497/paper-070.pdf.
[25] E. Amigó, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, 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. 5809–5819. doi:10.18653/v1/2022.acl-long.399.
[26] OpenAI, https://help.openai.com/en/articles/8555510-gpt-4-turbo-in-the-openai-api, 2024. Accessed: 2024-07-04.
[27] J. Fu, S. Xu, H. Liu, Y. Liu, N. Xie, C.-C. Wang, J. Liu, Y. Sun, B. Wang, CMA-CLIP: Cross-Modality Attention Clip for Text-Image Classification, in: 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 2846–2850. doi:10.1109/ICIP46576.2022.9897323.
[28] M. V. Conde, K. Turgutlu, CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 3951–3955. doi:10.1109/CVPRW53098.2021.00444.
[29] 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. doi:10.18653/v1/N19-1423.
[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.
[31] N. Ferro, J. Gonzalo, J. Karlgren, H. Müller, MonsterCLEF: One Lab to Rule Them All, 2024. URL: https://monsterclef.dei.unipd.it/, Accessed: 2024-07-04.