RoJiNG-CL at EXIST 2024: Leveraging Large Language Models for Multimodal Sexism Detection in Memes

Notebook for the EXIST Lab at CLEF 2024

Jing Ma[1,†], Rong Li[1,*,†]

[1]University of Zurich, Zurich, Switzerland

Abstract

This paper addressed Task 4 of the challenge posed by Sexism Identification in Social Networks (EXIST) at Conference and Labs of the Evaluation Forum (CLEF) 2024, which involves binary classification to determine the presence of sexism in memes. The task dataset contains memes in both English and Spanish. We explored the application of Large Language Models (LLMs), specifically GPT-4, for extracting textual descriptions from memes. Our methodology integrated these descriptions with associated texts to fine-tune various models, both monolingual and multilingual, to enhance the classifiers' ability to identify sexist content in memes using hard labels. By experimenting with diverse models and hyperparameters, we tailored our approach to optimize performance. Our submissions achieved the top three positions on the hard-hard evaluation leaderboard, which includes both English and Spanish instances.

Keywords

Memes sexism identification, Classification, Large Language Models (LLMs), Prompt engineering

1. Introduction

Various social networking platforms provide a virtual space where internet users can freely express themselves. However, this freedom is tainted by the presence of sexist or misogynistic content, potentially leading to physical and psychological harm to women [1]. Thus, developing effective mechanisms to detect and identify such content is crucial. The rapid expansion of Natural Language Processing (NLP) in the social sciences has prompted researchers to explore its capabilities for identifying sexist content in textual data. Previous studies have applied Long-Short-Term Memory networks (LSTMs) [2] and Convolutional Neural Networks (CNNs) [3] to classify such content [4]. Additionally, more advanced language models such as Electra [5], BERT [6], RoBERTa [7], and GPT-2 [8] have demonstrated significant efficacy in the classification of sexist text.

The challenge extends beyond textual analysis as sexism in online content often includes visual elements, particularly in memes. Memes, often considered jokes, gain attention through their rapid digital dissemination within online communities [9]. While frequently humorous, memes can also subtly propagate hate messages, including sexism and misogyny, causing harm at both individual and societal levels. The Sexism Identification in EXIST [10], part of CLEF 2024 [11], reflects this complexity by addressing sexism in both tweets and memes. Our research specifically focuses on Task 4: identifying sexism within memes, aiming to effectively classify these multimodal expressions. The integration of text and image in detecting sexism necessitates sophisticated vision-language models. While models like Residual Network (ResNet) [12] and Vision Transformer (ViT) [13] are essential for processing images, recent advancements have introduced more integrated models capable of handling the complexities of memes. For instance, CLIP [14] and multimodal models like mPLUG-Owl [15] and OpenFlamingo [16] have shown substantial proficiency in image classification tasks. Despite their effectiveness, these architectures require significant computational resources and extensive processing time.

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

[†] These authors contributed equally.

* Corresponding author.

jing.ma2@uzh.ch (J. Ma); rong.li@uzh.ch (R. Li)

To address these challenges and enhance cost-efficiency, our work incorporates zero-shot prompting techniques with GPT-4 [17] to extract textual descriptions from memes, considering both text and image information. This approach allows for a nuanced understanding of the meme's context, crucial for detecting underlying sexist themes. We refine this approach by fine-tuning models on a dataset comprising both provided texts and GPT-4 generated descriptions. Given the multilingual nature of social media content, we process memes in both English and Spanish, employing language-specific models to process memes in the respective language, and using multilingual models for the entire dataset.

2. Related Work

2.1. Text-Based Sexism Detection

Research on text-based sexism detection has mainly centered around analyzing social media texts. The pioneering work by Waseem and Hovy [18] on detecting hate speech on Twitter, including sexist content, highlighted the importance of linguistic and extra-linguistic features and expert annotations in training classifiers. The research indicated the potential of character n-grams to outperform other textual features like word n-grams and user demographic metadata. However, they also noted the challenge of scalability due to the labor-intensive nature of manual tagging. More recent researchers have explored automated feature extraction using transformer-based models like BERT and its variants [6][7], which excel in contextual understanding and have shown remarkable improvements in detecting complex and subtle sexist expressions. This advancement is evident in [19], which created the first Spanish corpus for sexism on Twitter. They involved a combination of traditional classifiers like Logistic Regression and Random Forest, neural network approaches including Bi-LSTM networks, and BERT. Their results showed that BERT outperformed other methods. EXIST 2023 [10] tackled the challenge of detecting sexism in tweets, focusing on identifying sexism, determining the source's intention, and categorizing types of sexism. The approaches for these tasks primarily involved fine-tuning models such as mBERT, XLM-Roberta, GPT-NeoX, BERTIN-GPT-J-6B, and Bernice [20] [21] [22], employing techniques like ensembling and contrastive learning. These methods demonstrated excellent performance on the tasks.

2.2. Multimodal Approaches to Sexism Identification

Given the complex nature of internet expressions, particularly within the realm of social media, sexism detection has necessarily expanded beyond textual data to include visual content, where images often carry implicit messages not evident in text alone. Fusion of multimodal information has become a popular method in this domain. For instance,[23] employed a methodology for detecting sexism in advertisements by combining outputs from visual and textual classifiers. The visual classifier analyzed features like Local Binary Pattern (LBP) and deep learning features extracted using a pre-trained CNN AlexNet[24], while the textual classifier utilized n-grams, syntactic tags, metadata about word usage, and word embeddings. A notable study by [25] pioneered the challenge of identifying sexist content in memes, proposing a framework that uses both unimodal and multimodal classifiers. This research developed unimodal classifiers that analyzed either textual or visual meme features independently using models like Support Vector Machines (SVM), Naive Bayes (NB), Decision Tree (DT), and 1-Nearest Neighbors (1NN). The study explored both early and late fusion techniques for integrating these modalities. Early fusion combined features at the input level before classification, whereas late fusion aggregated outputs from the unimodal classifiers post-analysis to determine the presence of sexism. Their findings indicated that textual classifiers typically outperformed visual classifiers, suggesting that textual cues are stronger indicators of sexism in memes. Additionally, late fusion was found to be more effective than early fusion, demonstrating that preserving the integrity of modality-specific features by combining classifier outputs after individual analyses can enhance overall accuracy. The study concluded that while unimodal approaches hold value, particularly in textual analysis, their integration

with multimodal strategies significantly improves the effectiveness and reliability of sexism detection in memes.

More recent works on detecting sexist or misogynous memes used state-of-the-art (SOTA) pre-trained models [26] [27]. Visual features are extracted using CLIP [14], and multimodal models such as mPLUG-Owl [15] and OpenFlamingo [16] have also been employed. These pipelines achieve high performance but are computationally intensive, necessitating substantial computational power and memory. Considering these limitations, our work uses GPT-4, which can simultaneously process image and text inputs, generating detailed descriptions of image content and integrating textual information for comprehensive analysis. This capability is crucial for understanding the humor, context, and cultural symbols in memes, addressing challenges highlighted by [25], which emphasize the difficulties of relying solely on visual features, as they can be ambiguous and less directly indicative of sexism compared to text.

3. Method

Our work focused on Task 4: Sexism Identification in Memes, which is a binary classification task aimed at determining whether a given meme is sexist or not. Our approach pipeline is outlined in Figure 1: First, inputs consisting of English and Spanish prompts, along with memes in the corresponding languages, were processed using zero-shot Chain-of-Thought (CoT) prompting by the GPT-4 model. This resulted in a one-sentence description of the meme and a hard-label output. Subsequently, this descriptive text was concatenated with the text extracted from the memes. This combined text, which included both visual and textual elements of the memes, was employed to fine-tune various language models. Depending on the language-specific requirements, we either used the entire dataset or the Spanish and English dataset respectively. Ultimately, the model made binary decisions, producing a definitive 'YES' or 'NO' hard-label.

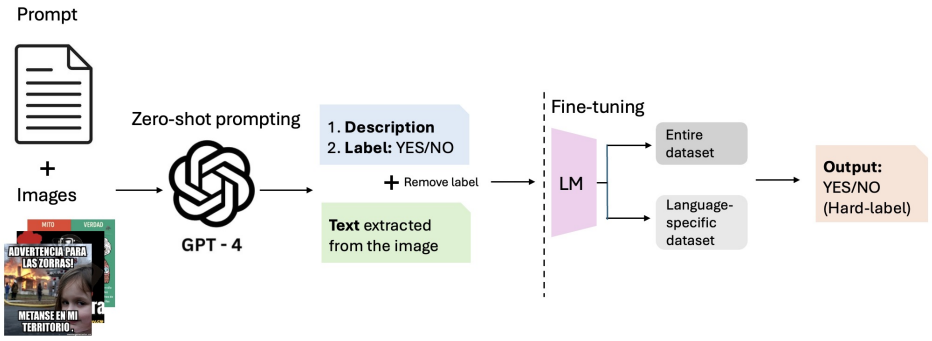

Figure 1: Pipeline of the EXIST Task4

3.1. Dataset

The dataset used in this study is provided by EXIST 2024 and encompasses both training and test datasets. The training dataset comprises a total of 4,044 memes, distributed between 2,034 Spanish memes and 2,010 English memes. The test dataset consists of 1,053 memes, with 540 classified as Spanish and 513 as English. The associated metadata for these memes is archived in a JSON file, which includes extracted texts from the memes and task-specific annotations: six labels annotated by six different annotators. Additionally, it captures detailed demographic information about each annotator, including their gender, age, ethnicity, educational attainment, and country of residence. The gold labels are provided in a separate JSON file; for hard-labels, when the human-annotated labels for a meme are evenly split between "YES" and "NO," such memes are excluded from the training dataset.

To facilitate a more effective evaluation of the results, we subdivide the training sets into separate training and validation datasets, using a random allocation of 20% for validation. The validation dataset is designed to monitor model performance on unseen data and plays a crucial role in adjusting hyperparameters to mitigate overfitting. For the multilingual mDeBERTa-v3-base model[28], we split the entire dataset into training and validation subsets as a whole. In contrast, when fine-tuning monolingual models such as BETO-uncased[29] for Spanish and bert-base-uncased[30] for English, we performed splits for the Spanish and English memes first, followed by targeted fine-tuning for each language group.

3.2. Image Information Extraction

Our task focuses on memes that typically contain both visual and textual elements. We employed a two-step process to analyze these memes. First, we extracted the text from the memes provided in the JSON dataset. Following this, we analyzed the visual content using the GPT-4 Turbo model API, chosen for its robust multimodal understanding capabilities.

Prompting input and output example for Spanish memes	
	
Prompt	Eres un experto en identificar la presencia de sexismo en los memes. Tu tarea es evaluar si el contenido de un meme retrata o critica el sexismo a través de la marginación, prejuicios contra las mujeres basados en estereotipos de género, creencias en la superioridad masculina o vistas misóginas. Describa y analice brevemente el contenido del meme en una frase, luego etiquételo como 'YES' si el meme encarna o critica el sexismo, o 'NO' si no lo hace. Por favor, utilice el siguiente formato: 1. Description: [Tu descripción]. 2. Label: [YES/NO].
GPT-4-Turbo	Description: "El meme muestra dos paneles. En el primero, bajo el título \"MITO\", Albert Einstein aparece diciendo \"2+2=5\", con un texto que menciona un mito sobre su bajo rendimiento en la escuela. En el segundo panel, titulado \"VERDAD\", Einstein dice \"2+2=4\" y está acompañado de un texto que critica al feminismo moderno, diciendo que defiende la estupidez humana y no los derechos de las mujeres, como ellas afirman.", Lable: "YES"

Figure 2: Prompting Input and Output for Spanish Memes

Prompting input and output example for English memes	
	
Prompt	You are an expert in identifying the presence of sexism within memes. Your task is to assess whether the content of a meme portrays or criticizes sexism through marginalization, prejudice against women based on gender stereotypes, beliefs in male superiority, or misogynistic views. Briefly describe and analyze the content of the meme in one sentence, then label it as 'YES' if the meme embodies or criticizes sexism, or 'NO' if it does not. Please use the following format: 1. Description: [Your description]. 2. Label: [YES/NO].
GPT-4-Turbo	Description:"The meme features the character Kermit the Frog drinking tea, with text suggesting that societal issues with gender identity wouldn't exist without rigid gender roles, using the phrase \"But that's none of my business\" to imply a dismissive attitude.", Lable:"YES"

Figure 3: Prompting Input and Output for English Memes

Prompt Engineering. Prompt engineering is crucial to optimize the performance of the model. We aim to keep the input and output not only informative, but also with a relatively short length. Recent studies have highlighted the potential of LLMs in role-playing scenarios. Assigning specific roles to an LLM can enhance the naturalness and interactivity of its responses [31, 32], and improve its performance in complex tasks [33]. We therefore configured the system's role as an expert in sexism

detection in memes. We experimented with prompts directing the model to analyze solely the image or the combination of image and text. Results indicated that the latter approach yielded more informative insights. This is due to the complementary nature of the image and text in memes, which, when analyzed together, provide a fuller and more accurate understanding of the meme's thematic message. In contrast, the analysis based on images alone often leads to neutral descriptions that are less relevant to the memes' themes. This aligns with observations by [25], which highlight the interpretative challenges posed by memes: first, identical images can be perceived as sexist or not based on the accompanying text, which can alter the conveyed message. Second, sexism may be manifested through the image alone, the text alone, or a combination of both.

We conducted prompt engineering primarily in English, and subsequently translated it into Spanish to accommodate the Spanish memes in the dataset using GPT-4. We also involved three bilingual (Spanish-native) speakers to evaluate the quality of the translations, ensuring the translations maintain the efficacy and accuracy of the original prompts. The final prompts, along with the input meme and the model output, are shown in Figure 2 (Spanish) and Figure 3 (English).

Model Configuration. The model was configured with specific settings to enhance performance and efficiency:

- **Model Setting:** `gpt-4-turbo`
- **Temperature:** 0.75 (to modulate the randomness of the outputs, ensuring their coherence and relevance)
- **Seed:** 1234 (to promote consistency in model responses across various runs, though absolute consistency cannot be guaranteed)
- **Detail:** Low (to process images in a resource-efficient manner, as fine details are not critical for our task)

3.3. Model Fine-tuning

At this stage, we experimented with various models, including multilingual models as well as monolingual models for English and Spanish. For the Spanish dataset, we fine-tuned the BETO-uncased model[29], and for the English dataset, the BERT-uncased model[30]. Additionally, we employed multilingual models including mBERT [30] (both cased and uncased), mDeBERTa[28], XLM-R[34], and XLM-Twitter[35]. The hyperparameter optimization was facilitated using Optuna[36], a framework that automates the search for optimal hyperparameters through systematic exploration, considering factors like learning rate, number of training epochs, batch size, warmup steps, and weight decay. The learning rate was varied between 1×10^{-5} and 5×10^{-5}, with the number of training epochs ranging from 3 to 5. Batch sizes are set at 8 and 16, warmup steps ranged from 0 to 500, and weight decay from 0.0 to 0.3 to add a regularization term to the loss function to minimize overfitting. Early stopping mechanisms were also incorporated to curtail training upon stabilization of validation losses. The objective function for optimization was defined based on accuracy, with a total of 15 trials conducted to strike a balance between obtaining the best hyperparameters and managing computational resources.

Table 1 presents the optimal hyperparameter settings for the three NLP models: mDeBERTa-v3-base, BETO-uncased, and bert-base-uncased. The table includes values for learning rate, epoch, train batch size, warmup steps, weight decay, and dropout rate. Each model's settings are specifically configured to enhance its training efficacy and overall performance in tasks, reflecting a strategic approach to fine-tuning.

Following the identification of the best hyperparameters by Optuna, the models were fine-tuned again on the full training dataset. The performance of the fine-tuned models was then evaluated using the validation dataset, with particular focus on accuracy and the F1 score for the positive class (pos_label=1).

Table 1
Best Hyperparameters Setting for Different Models

Parameter / Model	mDeBERTa-v3-base	BETO-uncased	bert-base-uncased
Learning rate	1.58	3.50	1.58
Epoch	4	4	5
Train batch size	8	8	8
Warmup steps	5	58	171
Weight decay	0.00067	0.01185	0.25118
Dropout rate	0.00925	0.23191	0.14459

4. Results

Figure 4 depicts the training loss for four models: bert-base-uncased, BETO-uncased, mDeBERTa-v3-base, and a baseline model over five epochs. To have deeper insights into our approach, we introduced a baseline model to assess whether adding descriptions to the input enhances model prediction performance. This baseline model is a fine-tuned version of mDeBERTa-v3-base, which used only the meme texts provided in the dataset as features, excluding any descriptions generated by GPT-4.

The bert-base-uncased model demonstrates a significant and steady reduction in training loss from just below 0.6 to approximately 0.2. The BETO-uncased model exhibits a similar initial loss but a less steep decline, stabilizing just below 0.4. The mDeBERTa-v3-base starts with the highest initial loss at around 0.65 but shows a rapid decrease, converging close to 0.3, similar to the BETO-uncased in the final epochs. The baseline model, however, indicates a fluctuating decrease in training loss, starting at around 0.6 and ending slightly above 0.4. This effectively illustrates the varying efficiency and speed of learning across the models, with the bert-base-uncased model achieving the most pronounced improvement in training loss.

Alongside a baseline model, Figure 5 shows the change of model accuracy throughout four and five training epochs for three distinct models: BETO-uncased, mDeBERTa-v3-base, and bert-base-uncased. Interestingly, the accuracy of the bert-base-uncased model shows a steady rising trend, peaking at epoch 3 before settling, suggesting strong learning potential. The accuracy of the BETO-uncased model, on the other hand, has a more gradual growth, with significant improvement especially during the second epoch. While it still exhibits consistent improvement, the mDeBERTa-v3-base model displays minor fluctuations in the later epochs. The comparison illustrates the variations in learning dynamics and stability among the models in this comparison.

Table 2
Comparative Performance of NLP Models

Metric	Baseline	bert-base-uncased	mDeBERTa-v3-base	BETO-uncased
Evaluation Loss	0.648	0.722	0.614	**0.540**
Accuracy (%)	68.13	**78.07** (+9.94)	76.50 (+8.37)	76.38 (+8.25)
F1 Score Positive (%)	74.94	79.78 (+4.84)	80.29 (+5.35)	**81.38** (+6.44)
Runtime (s)	3.197	1.703	6.538	3.571
Samples per Second	213.98	200.825	104.769	96.059
Steps per Second	13.452	25.25	13.153	1.68
Epoch	4	5	4	4

With an emphasis on evaluation loss, accuracy, and F1 Score Positive, Table 2 compares the performance of several NLP models on a number of metrics. The mDeBERTa-v3-base model comes in second at 0.614, while the bert-base-uncased model has a greater evaluation loss of 0.722. The BETO-uncased model has the lowest evaluation loss at 0.540, suggesting a better capacity to decrease mistakes during the evaluation phase. The bert-base-uncased model has the highest accuracy at 78.07%, indicating a noteworthy improvement of almost 9.94% above the baseline's observed 68.13%. The BETO-uncased

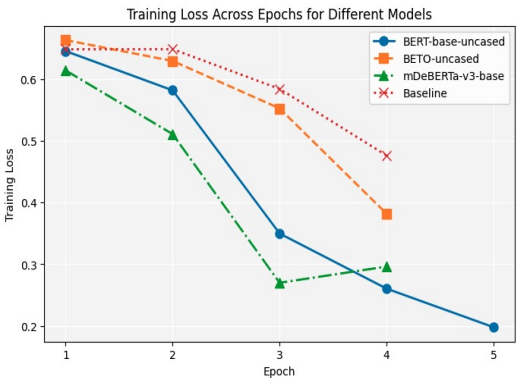

Figure 4: Model Training Loss Across Epochs

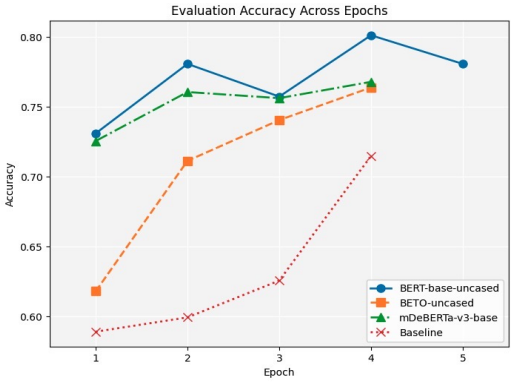

Figure 5: Model Evaluation Accuracy Across Epochs

Table 3
Official Results for Task 4 (Hard-hard evaluation for ALL Instances)

Run	ICM-Hard	ICM-Hard Norm	F1_YES	Ranking
RoJiNG-CL_3.json	0.3182	0.6618	0.7642	1
RoJiNG-CL_2.json	0.2272	0.6155	0.7437	2
RoJiNG-CL_1.json	0.1863	0.5947	0.7274	3
EXIST2024-majority-class.json	-0.4038	0.2947	0.6821	39
EXIST2024-minority-class.json	-0.6468	0.1711	0.0000	46

Table 4
Official Results for Task 4 (Hard-hard evaluation for EN and ES Instances)

Run	Language	ICM-Hard	ICM-Hard Norm	F1_YES
RoJiNG-CL_3.json	EN	0.3422	0.6737	0.7760
RoJiNG-CL_3.json	ES	0.2941	0.6498	0.7534
RoJiNG-CL_2.json	EN	0.2698	0.6370	0.7486
RoJiNG-CL_2.json	ES	0.1837	0.5936	0.7395
RoJiNG-CL_1.json	EN	0.2086	0.6059	0.7259
RoJiNG-CL_1.json	ES	0.1629	0.5830	0.7288

model has the greatest F1 Score Positive 81.38%, which is 6.44% higher than the baseline. This indicates that BETO-uncased, which effectively balances recall and precision, is especially good at properly recognizing the positive class even with a reduced evaluation loss. Overall, these findings highlight the distinct advantages of each model, with bert-base-uncased achieving the highest accuracy and BETO-uncased performs well at evaluation loss minimization and F1 score optimization.

Table 3 presents the official rankings for Task 4 on the Leaderboard. Our first run combined BERT predictions fine-tuned on English data and BETO fine-tuned on Spanish data. Our second run employed mDeBERTa, while the third run was the GPT-4 output results. We achieved top rankings out of more than 50 results, although surprisingly, the GPT-4 based predictions emerged as the most effective, delivering top results in a zero-shot setting, showcasing its exceptional capacity to comprehend and analyze complex sexist memes.

Further analysis, as shown in Table 4, indicates a consistent trend where all models achieved higher scores on the English dataset compared to the Spanish one. Despite the size similarities between BERT and BETO, BETO's lower performance relative to BERT highlights the challenge of achieving effectiveness gap between English and Spanish.

5. Ablation Study

This ablation study was initiated to address the performance discrepancies observed between the English and Spanish datasets. To explore whether translating the English datasets into Spanish could serve as a method of data augmentation, we translated the entire English dataset, including both the original meme texts and descriptions generated by GPT-4, into Spanish using the DeepL API. This translated data was then combined with the existing Spanish training dataset. Following the integration, the combined dataset was divided into training (80%) and validation sets (20%). For model evaluation purposes, the validation portion of the original Spanish dataset was repurposed as our test set due to the unavailability of gold labels for the original Spanish test dataset.

Table 5
Comparative Performance of NLP Models

Metric	Baseline	BETO-uncased	BETO-uncased (Ablation)
Accuracy (%)	68.13	76.38 (+8.25)	72.59 (+4.46)
F1 Score Positive (%)	74.94	**81.38** (+6.44)	78.44 (+3.50)

As illustrated in Table 5, the ablation study revealed no performance improvement with the augmented Spanish dataset. Several factors might have influenced this result. Primarily, the translations provided by the DeepL API could have introduced semantic inaccuracies or noise, complicating the training process. Although these translation tools offer a quick method for converting large datasets from one language to another, they may not always capture the nuanced cultural contexts and idiomatic expressions necessary for accurate sentiment and thematic analysis. These translation errors likely reduced the model's ability to learn effectively, resulting in worse performance. Additionally, the BETO-uncased model may not have been optimally fine-tuned for the nuances of the Spanish language, potentially limiting its processing and comprehension abilities on the Spanish dataset. To mitigate these issues, future research should concentrate on enhancing the quality of translations, employing advanced data augmentation strategies, and ensuring thorough fine-tuning of the models for specific language contexts.

6. Conclusion

In this working notes, we have demonstrated that LLMs, particularly GPT-4, can serve as competitive tools for extracting textual information from memes. Our methodology, with the strategic use of prompt engineering, has sidestepped the complexities typically associated with multimodal approaches and focused on generating descriptive texts directly from meme content. This approach not only simplifies the computing resources needed but also enhances our system's ability to detect subtlety that cannot be fully understood by mere texts. Our results are promising, showing that the application of LLMs, when finely tuned with tailored prompts, can effectively interpret and describe meme content. This is crucial for tasks requiring not just textual extraction but also an understanding of underlying societal and cultural contexts conveyed through humor and satire in memes.

The scope and generalizability of this study are constrained by several factors. First, the outputs of LLMs may exhibit intrinsic biases originating from their training data. These biases, particularly gender biases, could potentially lead to descriptions that are not accurate or appropriate, thereby misrepresenting the intent or sentiment of the memes. Another notable limitation is our reliance on binary ('YES' or 'NO') hard labels for sexism classification. However, in reality, sexism often exists on a continuous spectrum rather than a simple classification question. This complexity is particularly pronounced in the context of memes, which are inherently open to interpretation. Variability in perceptions among different individuals is common, as reflected in our dataset, where annotators frequently disagree. In future work, we aim to explore this diversity of human perspectives more thoroughly by incorporating soft labels that better capture the spectrum of responses.

Acknowledgments

We would like to express our gratitude to Simon Clematide and Andrianos Michail for their invaluable support and insightful suggestions. We also extend our thanks to the Department of Computational Linguistics at the University of Zurich for their financial support in making this project possible.

References

[1] B. Karthikeyan, S. Sundarraj, C. Sampathkumar, K. Mouthami, N. Yuvaraj, Sexism classification in social media using machine learning algorithms, in: A. Abraham, T. Hanne, N. Gandhi, P. Manghirmalani Mishra, A. Bajaj, P. Siarry (Eds.), Proceedings of the 14th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2022), Springer Nature Switzerland, Cham, 2023, pp. 14–23.

[2] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780.

[3] K. O'Shea, R. Nash, An introduction to convolutional neural networks, ArXiv abs/1511.08458 (2015). URL: https://api.semanticscholar.org/CorpusID:9398408.

[4] A. Kalra, A. Zubiaga, Sexism identification in tweets and gabs using deep neural networks, ArXiv abs/2111.03612 (2021). URL: https://api.semanticscholar.org/CorpusID:243832598.

[5] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, ArXiv abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[9] C. Iloh, Do it for the culture: The case for memes in qualitative research, International Journal of Qualitative Methods 20 (2021) 16094069211025896. URL: https://doi.org/10.1177/16094069211025896. doi:10.1177/16094069211025896. arXiv:https://doi.org/10.1177/16094069211025896.

[10] 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.

[11] 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. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 770–778. URL: https://api.semanticscholar.org/CorpusID:206594692.

[13] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017. URL: https://api.semanticscholar.org/CorpusID:13756489.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[15] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, F. Huang, mplug-owl: Modularization empowers large language models

with multimodality, ArXiv abs/2304.14178 (2023). URL: https://api.semanticscholar.org/CorpusID:258352455.

[16] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Y. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, L. Schmidt, Openflamingo: An open-source framework for training large autoregressive vision-language models, ArXiv abs/2308.01390 (2023). URL: https://api.semanticscholar.org/CorpusID:261043320.

[17] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[18] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: J. Andreas, E. Choi, A. Lazaridou (Eds.), Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: https://aclanthology.org/N16-2013. doi:10.18653/v1/N16-2013.

[19] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, Automatic classification of sexism in social networks: An empirical study on twitter data, IEEE Access 8 (2020) 219563–219576. doi:10.1109/ACCESS.2020.3042604.

[20] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, Ai-upv at exist 2023 - sexism characterization using large language models under the learning with disagreement regime, ArXiv abs/2307.03385 (2023). URL: https://api.semanticscholar.org/CorpusID:259376983.

[21] L. Tian, N. Huang, X. Zhang, Efficient multilingual sexism detection via large language model cascades., 2023.

[22] J. Angel, S. T. Aroyehun, A. F. Gelbukh, Multilingual sexism identification using contrastive learning., in: CLEF (Working Notes), 2023, pp. 855–861.

[23] F. Gasparini, I. Erba, E. Fersini, S. Corchs, Multimodal classification of sexist advertisements, in: International Conference on E-Business and Telecommunication Networks, 2018. URL: https://api.semanticscholar.org/CorpusID:52121016.

[24] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. Burges, L. Bottou, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 25, Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[25] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist meme on the web: A study on textual and visual cues, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 226–231. doi:10.1109/ACIIW.2019.8925199.

[26] H. B. Zia, I. Castro, G. Tyson, Racist or sexist meme? classifying memes beyond hateful, in: A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, Online, 2021, pp. 215–219. URL: https://aclanthology.org/2021.woah-1.23. doi:10.18653/v1/2021.woah-1.23.

[27] S. Chen, U. Naseem, I. Razzak, F. Salim, Unveiling misogyny memes: A multimodal analysis of modality effects on identification, in: Companion Proceedings of the ACM on Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1864–1871. URL: https://doi.org/10.1145/3589335.3651974. doi:10.1145/3589335.3651974.

[28] M. Laurer, W. v. Atteveldt, A. S. Casas, K. Welbers, Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI, Preprint (2022). URL: https://osf.io/74b8k, publisher: Open Science Framework.

[29] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[30] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[31] Y. Tao, A. Agrawal, J. Dombi, T. Sydorenko, J. I. Lee, Chatgpt role-play dataset: Analysis of user motives and model naturalness, in: International Conference on Language Resources and Evaluation, 2024. URL: https://api.semanticscholar.org/CorpusID:268723733.

[32] Z. Wang, Z. Peng, H. Que, J. Liu, W. Zhou, Y. Wu, H. Guo, R. Gan, Z. Ni, M. Zhang, Z. Zhang, W. Ouyang, K. Xu, W. Chen, J. Fu, J. Peng, Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models, ArXiv abs/2310.00746 (2023). URL: https://api.semanticscholar.org/CorpusID:263334495.

[33] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, Camel: Communicative agents for" mind" exploration of large scale language model society (2023).

[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

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

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.