LightGBM for Sexism Identification in Memes*

Notebook for the EXIST Lab at CLEF 2024

Arnau Garcia i Cucó[1,†], Miquel Obrador Reina[2,†]

[1]Universitat Politècnica de València (UPV), València, Spain
[2]Universitat Politècnica de València (UPV), València, Spain

Abstract

In this paper, we present our participation in the EXIST2024 competition, focusing on the detection and categorization of sexism in memes through binary and multiclass classification tasks. For Task 4, which involves binary classification to identify whether a meme is sexist, we utilized a two-stage approach by fine-tuning the Contrastive Language-Image Pre-training (CLIP) model followed by training a Light Gradient Boosting Machine (LightGBM) classifier on the obtained embeddings. For Task 6, which categorizes sexist memes into various types, we fine-tuned both a RoBERTa model for text and a Google Vision Transformer (ViT) for images, subsequently training a LightGBM classifier on the concatenated embeddings. Our results demonstrate that the combination of CLIP and LightGBM outperforms other models in binary classification, while an ensemble of text and image models enhances performance in multiclass categorization. These findings highlight the potential of leveraging advanced machine learning models and multi-modal embeddings for effective sexism detection and classification in memes, addressing a critical issue in combating online misogyny.

Keywords

Sexism detection, meme classification, binary classification, multiclass classification, ViT, Transformers, Gradient Boosting Methods

1. Introduction

According to Glick and Fiske's concept of "sexism" [3], sexism is a multidimensional construct that encompasses two sets of sexist attitudes: hostile and benevolent. While hostile sexism communicates a clear antipathy toward women, benevolent sexism takes the form of seemingly positive but in fact patronizing beliefs about women. These beliefs often involve traditional stereotyping and masculine dominance, ultimately leading to a restriction of women's roles and damaging consequences for gender equality.

EXIST is a series of scientific events and shared tasks on sexism identification in social networks.Building upon previous EXIST challenges (EXIST 2021, EXIST 2022, and EXIST 2023), this edition, EXIST2024 [5] aims to capture sexism in a broad sense, from explicit misogyny to subtle expressions involving implicit sexist behaviors. While the three previous editions focused solely on detecting and classifying sexist textual messages, this new edition incorporates new tasks that center around images, particularly memes. Memes are images, typically humorous in nature, that are spread rapidly by social networks and Internet users. The challenge consists of five tasks in two languages, English and Spanish.

2. Tasks performed

In the EXIST2024 competition, our team chose to participate in Task 4, which focuses on identifying sexism in memes through binary classification, determining whether a given meme is sexist or not. Additionally, we took part in Task 6, which involves categorizing sexist memes into various types of sexism, such as ideological and inequality, stereotyping and dominance, objectification, sexual

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

[†]These authors contributed equally.

agarcuc@inf.upv.es (A. G. i. Cucó); mobrrei@inf.upv.es (M. O. Reina)

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

violence, and misogyny and non-sexual violence. This categorization is based on the classification system provided for Task 3.

3. Main Objectives of Experiments

In the EXIST2024 competition, our primary objective is to investigate the effectiveness of machine learning algorithms in identifying and categorizing sexist memes through binary classification and multiclass classification. Specifically, we aim to develop models that can accurately determine whether a given meme is sexist or not, and further classify sexist memes into various types of sexism.

Our research questions include:

- How effective are machine learning algorithms in identifying sexist memes through binary classification?
- Can we develop a model that accurately categorizes sexist memes into various types of sexism, such as ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence?
- What are the challenges and limitations of using machine learning algorithms for sexism detection in memes, and how can we address these issues?

To address these questions, our experiments will focus on the following tasks:

- Binary Classification: We will develop and evaluate machine learning models that can accurately classify memes as sexist or not sexist.
- Multiclass Classification: We will develop and evaluate machine learning models that can accurately categorize sexist memes into various types of sexism, as defined by the EXIST2024 competition's classification system.
- Model Generalization: We will evaluate the performance of our models on a diverse range of memes, including those from different cultural and linguistic backgrounds. This will help us assess the generalizability of our models and identify potential issues related to cross-cultural differences and language barriers.

By addressing these objectives, we aim to contribute to the development of more accurate, fair, and robust machine learning models for sexism detection in memes, ultimately helping to combat the spread of hateful content against women on social media platforms.

4. Approaches and methodology

4.1. Task 4

In the context of the EXIST2024 Task 4 for binary classification of memes, our methodology involved a two-stage approach to address the classification problem. The dataset was splitted into train/val test in order to evaluate the model.

Initially, we fine-tuned the pre-trained Contrastive Language-Image Pre-training (CLIP) model [6], a deep learning model developed by OpenAI that has demonstrated exceptional performance in vision and language tasks. CLIP is trained on a vast dataset of 400 million (image, text) pairs, enabling it to learn rich visual and textual representations.

During the first stage of our approach, we incorporated a fully connected network specifically designed for binary classification within the CLIP model. This allowed the model to learn task-specific features and improve its ability to differentiate between the two classes of memes. Once the fine-tuning process was complete, we discarded the fully connected layer, as our primary focus was on obtaining high-quality embeddings that could be used to train a more specialized classifier.

In the second stage, we employed a Light Gradient Boosting Machine (LightGBM) [4] classifier, a highly efficient and effective gradient boosting framework that uses tree-based learning algorithms. By training the LightGBM classifier with the CLIP embeddings obtained from the first stage, we aimed to further enhance the classification performance. This two-stage training strategy allowed us to leverage the strengths of both CLIP and LightGBM, resulting in improved binary classification of memes in the EXIST2024 Task 4. We tried a logistic regression classifer instead of LightGBM for the final classification (See table 1).

4.2. Task 6

For the EXIST2024 Task 6, after segmenting the dataset into train/val and test, ensuring a robust evaluation process, we meticulously crafted a methodology involving three stages: two fine tunings and a LightGBM training.

The classification problem we tackled was multifaceted, encompassing six categories. Initially, we had to take into account binary classification to distinguish between sexist and non-sexist memes. Then, if the meme was deemed sexist, we further dissected the binary category into five classes that could overlap: stereotyping, objectification, ideological, sexual violence, and misogyny. It's important to note that there was no overlap when the binary category was 'no '. Thus, we defined only the five sexism categories as labels, and if a meme didn't fall into any of these categories, we assumed it to be non-sexist.

Furthermore, this task was labelled by a group of experts who voted on the classification they considered optimal. Thus, we had not only the hard label but also the soft label given by those experts. We aimed to use all this data to configure an accurate model.

As stated previously, we had images and text for the meme data. This information could not be used as such, so we tried to obtain embeddings that would represent task-specific features. To do so, we fine-tuned a RoBERTa model [2] in the first stage, incorporating a final fully connected layer. Then, we retrained the model to learn the soft labels, using the Kullback-Leibler divergence as a loss function. In the second phase, we did the same for the image, using, in this case, a Google ViT [1] as the base model. The idea in those steps was to represent the memes in arrays that sum the features the experts use to discriminate them.

Finally, in the last stage of the process, we concatenated both embeddings (thus enhancing the information achieved in image and text), and we applied a LightGBM to finally learn to classify into hard labels.

5. Resources employed

To develop and evaluate our machine learning models for sexism detection in memes, we utilized various resources, including open-source datasets, pre-trained models, and computational platforms. Our primary focus was on utilizing Google Colab's free GPU resources to train and test our models efficiently. Here is a detailed breakdown of the resources we used for model development:

- Datasets: We relied on publicly available datasets, such as the EXIST2024 competition's training dataset, which provided a diverse collection of memes labeled as sexist or not sexist. Additionally, we used external datasets to augment our training data and improve model performance.
- Pre-trained Models: We leveraged pre-trained models, such as RoBERTa, Google ViT and CLIP, as feature extractors for our meme classification tasks. These models were fine-tuned on our training data to improve their performance on the sexism detection and categorization tasks.
- Google Colab: We used Google Colab's free GPU resources to train and test our models. This allowed us to efficiently experiment with various algorithms, hyperparameters, and feature engineering techniques without incurring significant computational costs.

By leveraging these resources, we were able to develop and evaluate machine learning models for sexism detection and categorization in memes. Our experiments focused on addressing the research questions outlined in the main objectives section, ultimately contributing to the development of more accurate, fair, and robust models for combating hateful content against women on social media platforms.

6. Results

6.1. Task 4

Table 1
Task 4 Results

Model	F1	MCC
CLIP$_{class}$	0.71	0.32
CLIP$_{LGBM}$	**0.72**	**0.34**
CLIP$_{LR}$	0.71	0.33

The results presented in Table1 showcase the performance of different models on the binary classification task for memes. The models evaluated include CLIP with a logistic regression (LR) classifier, CLIP with Light Gradient Boosting Machine (LGBM) classifier, and CLIP with a simple binary classifier ($CLIP_{class}$). The performance metrics used for evaluation are F1-score and Matthews Correlation Coefficient (MCC).

The results indicate that the CLIP model with the LightGBM classifier ($CLIP_{LGBM}$) outperforms the other models in both F1-score and MCC. The $CLIP_{LGBM}$ model achieves an F1-score of 0.72 and an MCC of 0.34, which is a slight improvement over the $CLIP_{class}$ model (F1-score of 0.71 and MCC of 0.32) and the $CLIP_{LR}$ model (F1-score of 0.71 and MCC of 0.33).

Table 2
Task 4 Rank

Task	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
Task 4 Hard-Hard ALL	26	-0.1159	0.4411	0.6632
Task 4 Hard-Hard ES	28	-0.1966	0.3999	0.6537
Task 4 Hard-Hard EN	27	-0.0355	0.4820	0.6739

In Table 2 we present the official results on the competition test set obtained for our best model for task 4, $CLIP_{LGBM}$ model.

6.2. Task 6

As for task 6, we observed the weighted F1-score when training with text had more potential than when training with image, as seen in table 3. The Google ViT did not learn to classify misogyny and sexual violence, thus causing a massive drop in its performance. However, the ensemble was the best performing so far, hence showing that the image data had relevant information for the classification.

Table 3
Task 6 Results

Model	Macro F1
RoBERTa	0.45
Google ViT	0.29
Ensemble	**0.49**

In Figure 1, we observe that even though the model predicts ideological inequality and objectification pretty well, it fails to predict accurately misogyny (non-sexual violence).

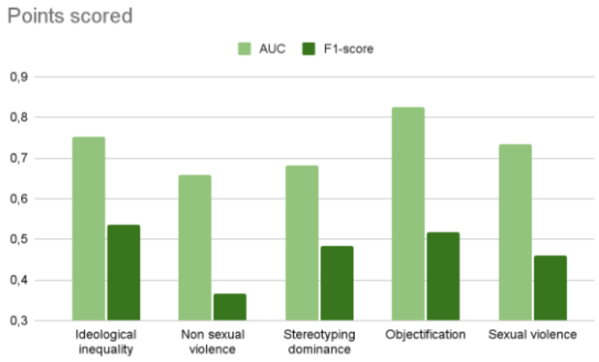

Figure 1: AUC and F1-score by category

Table 4
Task 6 Rank

Task	Rank	ICM-Hard	ICM-Hard Norm	MACRO F1
Task 6 Hard-Hard ALL	13	-1.6216	0.1636	0.3211
Task 6 Hard-Hard ES	2	-0.8701	0.3219	0.4687

Finally, in Table 4 we can see the official results for our submition for task 6 in EXIST2024.

7. Conclusions and Future work

The results showcase the effectiveness of the proposed two-stage training strategy, which combines the strengths of CLIP and LightGBM, resulting in improved binary classification performance. Future research could explore the potential of integrating other advanced classifiers or fine-tuning techniques to further enhance the classification capabilities of the CLIP model. Furthermore, we would like to experiment by applying data augmentation since we did not use the English data at all for task 6, and we could have included it in training by translating it, for example. We would also like to explore other ways to cross both embeddings rather than concatenating them, such as using attention mechanisms.

Building on the findings and methodologies presented in this paper, future research can explore several promising avenues to further enhance sexism detection and categorization in memes:

- Data Augmentation and Multilingual Training: Integrating additional data sources and employing data augmentation techniques, such as translating non-English memes, can improve model robustness and performance. This approach will help address the challenges of cross-cultural differences and language barriers.
- Advanced Embedding Techniques: Investigating alternative methods for combining text and image embeddings, such as attention mechanisms, can potentially yield richer and more contextually relevant features. This could enhance the model's ability to capture nuanced information from both modalities.
- User Feedback and Iterative Improvement: Incorporating user feedback into the model development cycle can help refine the models and make them more aligned with real-world expectations and nuances.
- Explainability and Interpretability: Enhance the interpretability of the models by incorporating explainable AI techniques. This would help in understanding the decision-making process of the models and identifying potential biases or errors.

By exploring these directions, future work can contribute to the development of more sophisticated, accurate, and ethical machine learning models for detecting and combating sexism in memes, ultimately fostering a safer and more inclusive online environment.

References

[1] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[2] Asier Gutiérrez Fandiño et al. "MarIA: Spanish Language Models". In: *Procesamiento del Lenguaje Natural* 68 (2022). ISSN: 1135-5948. DOI: 10.26342/2022-68-3. URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley.

[3] Peter Glick and Susan T. Fiske. "Ambivalent sexism". In: vol. 33. Advances in Experimental Social Psychology. Academic Press, 2001, pp. 115–188. DOI: https://doi.org/10.1016/S0065-2601(01)80005-8. URL: https://www.sciencedirect.com/science/article/pii/S0065260101800058.

[4] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[5] Laura Plaza et al. "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.

[6] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *CoRR* abs/2103.00020 (2021). arXiv: 2103.00020. URL: https://arxiv.org/abs/2103.00020.