A Contrastive Learning Based Approach to Detect Sexism in Memes

Notebook for the EXIST Lab at CLEF 2024

Fariha Maqbool[1], Elisabetta Fersini[1]

[1]*Dipartimento di informatica, sistemistica e comunicazione*
University of Milano-Bicocca
Viale Sarca 336, 20126 Milan, Italy

Abstract

The widespread use of social media has created a unique challenge in detecting and mitigating sexism in online content. In this paper, we present our approach for detecting sexism in memes, developed for Task 4 of the EXIST 2024 challenge. The task was based on binary classification problem to detect whether or not a meme is sexist, within the context of a learning with disagreement paradigm. In our approach, We used ResNet50 and m-BERT models finetuned on EXIST 2024 dataset to get image and text embeddings. These embeddings, along with the annotators' data, were subsequently used to train a model using contrastive learning. The results on the test data demonstrate the effectiveness of contrastive learning techniques in addressing multimodal tasks.

Keywords

Sexism Identification, Learning with disagreement, Contrastive Learning

1. Introduction

Sexism is a type of bias and prejudice that leads to detrimental sex-based stereotypes and societal expectations. It often involves a combination of gender-based beliefs, attitudes, and actions that result in uneven treatment of men and women. Historically and culturally pervasive, sexism against women is rooted in the notion of male supremacy, affecting various aspects of life such as the workplace, politics, society, and the family [1].

In today's digital age, the widespread use of social media has contributed to the alarming prevalence of sexist content. This content spreads rapidly, fueling more instances of sexism in various forms. However, detecting such content might be challenging due to the diverse ways it is expressed. Internet memes, in particular, have emerged as a notable medium for communicating these concepts in an engaging manner [2]. Detecting sexism and other forms of hateful content in memes poses a considerable challenge. Memes typically consist of an image accompanied by text, and while the visual and textual components are related, they may not convey the same meaning when viewed independently. Therefore, effectively identifying hateful memes requires a careful analysis of both the visual elements and the accompanying text.

In addition to the challenges of detecting sexism in memes, another challenge arises from the inherent subjectivity and disagreement among annotators when labeling such content. Different annotators may have varying perspectives on sexism or hate speech, influenced by their individual backgrounds, experiences, and cultural contexts. This disagreement can lead to inconsistent annotations, which must be carefully managed to train machine learning models. EXIST 2024 incorporates this learning with disagreement approach, which leverages these multiple perspectives, increasing dataset richness and improving the ability of models to generalize across different interpretations of harmful content.

In this paper, we describe the overview of the system we developed for sEXism Identification in Social neTworks (EXIST 2024) [3][4] shared task at CLEF 2024. Our team participated in Task-4 a binary classification task to detect whether or not the meme was sexist. We proposed a contrastive

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

f.maqbool@campus.unimib.it (F. Maqbool); elisabetta.fersini@unimib.it (E. Fersini)
0009-0008-2587-9417 (F. Maqbool); 0000-0002-8987-100X (E. Fersini)

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

learning-based approach to predict the hard labels for each meme that represents the label for each meme by the aggregation of perspectives of different annotators.

2. Related Work

Over the past few years, numerous academic events and shared tasks have focused on identifying misogyny [5][6][7] and detecting hate speech against immigrants and women [8]. It is important to note that sexism and misogyny are not always similar. Sexism encompasses a broad spectrum of oppression or prejudice against women that can range from overt hostility, such as misogyny, to more nuanced forms. Hence, while misogyny is a part of sexism, it certainly doesn't define its full extent.

To fill this research gap, sEXism Identification in Social neTworks (EXIST) shared tasks were proposed at the IberLEF forum [9][10] whose aim was to identify and classify sexism in textual data, from explicit or hostile to other subtle or even benevolent expressions that involve implicit sexist behaviours. In 2023, they again proposed the task with the adoption of the "learning with disagreements" paradigm for the development of the dataset and, optionally, for the evaluation of the systems [11]. EXIST 2024 [12] is the fourth edition of the sEXism Identification in Social neTworks challenge which presents shared tasks on sexism detection on social media. 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. Detection of sexism in memes is quite challenging because of the multimodality of memes. Elisabetta Fersini et al. [13] presented the first attempt to address the challenge of automatic detection of sexist memes. The study examined both unimodal and multimodal approaches to understand the role of textual and visual cues. They also released a benchmark dataset containing 800 memes, which include sexist and non-sexist content. Each meme is labeled based on visual and textual elements. The dataset comprises images along with their associated texts.

There have been contrastive learning techniques used for the detection of hateful, misogynous, or sexist content. Jason Angel et al.[14] presented an approach for multilingual sexism identification in tweets. They finetuned multilingual RoBERTa language model by integrating contrastive learning as an intermediate step. The competitive results achieved show the effectiveness of contrastive learning in sexism identification task in textual data. In addition to textual data, contrastive learning has also been used for vision language tasks. Charic F. Cuervo and Natalie Parde [15] used contrastive learning based model named CLIP [16] for the task of detecting misogynous memes. They slightly modified the CLIP model's approach such that the language content from the meme was used as the training text along with the correct label. Lei Chen and Hou W. Chou [17] also used CLIP model for feature extraction and Logistic regression in these extracted features to detect misogyny in memes. These studies collectively highlight the significant advancements made through the application of contrastive learning techniques in detecting hateful, misogynous, and sexist content across different modalities.

3. Task Description and Dataset

The dataset consists of 4044 memes for training and 1053 memes for testing in both English and Spanish. We split the training dataset to training, validation and test sets in the ratio of 80,10 and 10 respectively. The text of the memes has already been compiled by the organizers in a separate file. In Table 1 we described the details of the dataset. We worked only on the Task-4 to identify whether or not the memes are sexist for which binary labels were provided. The dataset also follows the learning with disagreement paradigm, in which each data point is labeled by multiple annotators, and disagreements between their annotations are retained for analysis. All demographic data, including gender, age, ethnicity, level of education, and country of residence, is carefully documented for every annotator. The labels assigned to memes by each annotator was also recorded in the dataset, and a hard label was determined through majority voting. In cases where there was an equal number of votes, the label 'unknown' was assigned. These particular samples, labeled as 'unknown', were excluded from our training data.

Table 1
Dataset description

Data Type	Total Size	Spanish	English	sexist	Non-sexist	unknown
Train	4044	2034	2010	2038	1382	624
Test	1053	540	513	–	–	–

4. System Overview

We implemented a contrastive learning based strategy to perform binary classification of memes to sexist and non-sexist. Figure 1 shows the workflow of our proposed system.

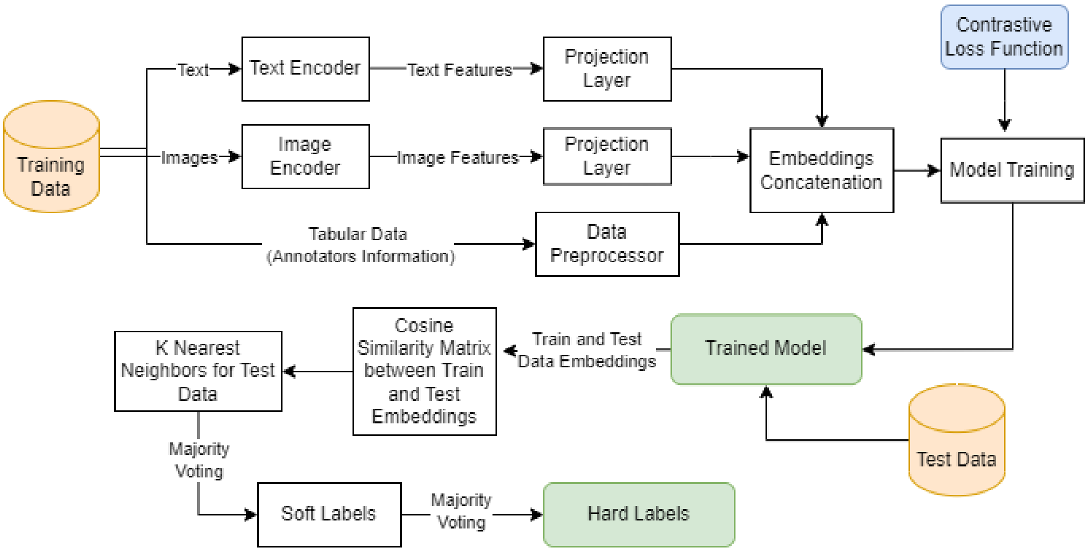

Figure 1: Workflow of the Proposed System Architecture

4.1. Image and Text Encoders

We used ResNet-50 as image encoder which is a widely used deep convolutional neural network architecture introduced by Kaiming He [18] known for its effectiveness in image recognition and computer vision tasks. It consists of 50 layers, including convolutional layers, batch normalization, and ReLU activation functions. We finetuned ResNet-50 model on the respective dataset before getting image embeddings. The torchvision library of python was used to load the pre-trained ResNet-50 model. We froze all its layers to retain their pre-trained weights and unfroze the last few layers to allow fine-tuning. To adapt the model for this specific task, we modified the head of the model by replacing the fully connected layer with additional layers, including a dropout layer for regularization. This model was finetuned for 30 epochs with adam optimizer and later used for feature extraction of images.

To get the text embeddings we finetuned transfomer based multilingual BERT model introduced by Jacob Devlin et al. [19]. To fintune this model, each text sample is tokenized using the BERT tokenizer, ensuring that the sequences are appropriately padded and truncated to the specified maximum length. The tokenized text data, along with the corresponding attention masks, token type IDs, and labels, are converted into tensors suitable for model input. This preparation is crucial for ensuring that the text data are in the correct format for BERT, facilitating efficient and effective training. The core of the model was initialized with pre-trained weights from the multilingual cased BERT and a linear layer was added at model's head that maps the BERT output to the desired number of output classes, which in this case is one (for binary classification). A sigmoid activation function was used at the output layer to facilitate binary classification.

4.2. Projection layer

After encoding the images and texts, they were passed through a projection layer for dimensionality reduction and feature transformation on input vectors. This projection layer consists of a Linear projection layer, Gaussian Error Linear Unit (GELU) activation function and fully connected layers. It ensures that embeddings from different modalities are effectively aligned and normalized, facilitating improved downstream learning and integration tasks.

4.3. Combining features

In order to combine the image and text features, we used the feature interaction matrix (FIM) introduced by Gokul K. Kumar et al.[20] that directly models the correlations between each text and image. This matrix is obtained by computing the outer product of each text and image feature, but in order to reduce the dimensionality of the representations, the authors only considered the diagonal elements of FIM. We also followed the same strategy. The dimension of the vector obtained from this method was n. To include the annotator's information, the tabular data containing annotator's information was also concatenated with these image-text features.

4.4. Training and Testing

Our contrastive learning-based model is then trained on these combined features using infoNCE loss function with the objective of increasing the cosine similarity between the memes of similar classes and decreasing between dissimilar ones. To evaluate the trained model, we conducted tests on the evaluation dataset. Firstly, we extracted the image and text data features of the test samples using the model encoders. Next, we computed the cosine similarity between each test sample and all the training samples. For each test sample, we used the K-Nearest Neighbors (KNN) algorithm to select the 10 training embeddings with the highest cosine similarity to that sample. The label for each test sample was then assigned based on the most common label among these 10 training samples. We predicted the labels for each annotator separately, then we applied majority voting on these labels to find the final hard label for each sample.

5. Experimentation and Results

In our implementation, we used PyTorch library in Python. After feature extraction and concatenation, the model was trained using contrastive loss for 50 epochs with Adam optimizer and a batch size of 32. We used the transformers library to train our model with learning rate set to 1e-5 for image encoder and 1e-4 for text encoder. A dropout layer was also added for regularization. We save the model with the lowest contrastive loss on the validation set during training. We then use the saved model to make predictions on the unseen test set.

The challenge uses the ICM metric [21] to evaluate the performance. This metric is a similarity function that extends the concept of Pointwise Mutual Information (PMI) to measure the similarity between the model's predictions and the ground truth categories. The normalized ICM is calculated by considering the "Minority class" baseline, which assigns all instances to the minority class, as the lowest score, and the "Gold standard" as the highest score.

Table 2 shows the results of our system based on hard hard evaluation. Hard-Hard evaluation means that the final hard labels of the samples are compared with the gold labels of test set. The model was able to achieve the best score on samples with English text with ICM Normalized score of 0.277 and F1 score of 0.5816 for positive samples.

Table 2
Official results of our system on Hard-Hard Evaluation

Language	Model	ICM-Hard	ICM-Hard Norm	F1_YES
All	Baseline	0.9832	1.0000	1.0000
	Proposed Model	-0.4986	0.2465	0.5674
English	Baseline	0.9848	1.0000	1.0000
	Proposed Model	-0.4377	0.2778	0.5816
Spanish	Baseline	0.9815	1.0000	1.0000
	Proposed Model	-0.5591	0.2152	0.5537

6. Conclusion

In this paper, we present our approach and the results obtained for Task 4 of the sEXism Identification in Social neTworks (EXIST 2024) challenge. This task involves a binary classification problem, where the goal is to distinguish between sexist and non-sexist memes, incorporating a learning with disagreement paradigm. We employed ResNet-50 and mBERT models to encode the visual and multilingual textual data of the memes, respectively. After obtaining the embeddings from these models, we concatenated the data and trained a contrastive learning-based model on these embeddings. The performance of our model was evaluated on the test data using the ICM metric for hard labels, achieving ICM scores of 0.2778 for English, 0.2152 for Spanish, and 0.2465 for the combined dataset. These results demonstrate the effectiveness of our approach in addressing the challenge of sexism detection in a multilingual and multimodal context.

7. Acknowledgments

We acknowledge the support of the PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] A. ElBarazi, How social media affects people's ideas on sexist behaviours and gender-based violence (2023). doi:10.19080/GJIDD.2023.12.555838.

[2] C. Jennifer, F. Tahmasbi, J. Blackburn, G. Stringhini, S. Zannettou, E. D. Cristofaro, Feels bad man: Dissecting automated hateful meme detection through the lens of facebook's challenge (2022). doi:10.36190/2022.65.

[3] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[4] 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.

[5] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Z. Margetts, An expert annotated dataset for the detection of online misogyny, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics:

Main Volume, EACL 2021, Online, April 19 - 23, 2021, Association for Computational Linguistics, 2021, pp. 1336–1350. doi:10.18653/V1/2021.EACL-MAIN.114.

[6] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018, in: P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, J. C. de Albornoz (Eds.), Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018, volume 2150 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 214–228.

[7] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 533–549. doi:10.18653/V1/2022.SEMEVAL-1.74.

[8] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019, Association for Computational Linguistics, 2019, pp. 54–63. doi:10.18653/V1/S19-2007.

[9] F. J. Rodríguez-Sanchez, J. Carrillo-de-Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of EXIST 2021: sexism identification in social networks, Proces. del Leng. Natural 67 (2021) 195–207.

[10] F. J. Rodríguez-Sanchez, J. Carrillo-de-Albornoz, L. Plaza, A. Mendieta-Aragón, G. M. Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sexism identification in social networks, Proces. del Leng. Natural 69 (2022) 229–240.

[11] L. Plaza, J. Carrillo-de-Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023: sexism identification in social networks, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 593–599. doi:10.1007/978-3-031-28241-6_68.

[12] 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: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, volume 14612 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 498–504. doi:10.1007/978-3-031-56069-9_68.

[13] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist MEME on the web: A study on textual and visual cues, in: 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACII Workshops 2019, Cambridge, United Kingdom, September 3-6, 2019, IEEE, 2019, pp. 226–231. doi:10.1109/ACIIW.2019.8925199.

[14] J. Angel, S. T. Aroyehun, A. F. Gelbukh, Multilingual sexism identification using contrastive learning, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 855–861.

[15] C. F. Cuervo, N. Parde, Exploring contrastive learning for multimodal detection of misogynistic memes, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 785–792. doi:10.18653/V1/2022.SEMEVAL-1.109.

[16] 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.

[17] L. Chen, H. W. Chou, RIT boston at semeval-2022 task 5: Multimedia misogyny detection by using coherent visual and language features from CLIP model and data-centric AI principle, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022, Association for Computational Linguistics, 2022, pp. 636–641.

[18] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[19] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/V1/N19-1423.

[20] G. K. Kumar, K. Nandakumar, Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features, CoRR abs/2210.05916 (2022). doi:10.48550/ARXIV.2210.05916.

[21] E. Amigó, A. D. 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), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 5809–5819. doi:10.18653/V1/2022.ACL-LONG.399.