PINK at EXIST2024: A Cross-Lingual and Multi-Modal Transformer Approach for Sexism Detection in Memes

Notebook for the EXIST Lab at CLEF 2024

Giulia Rizzi[1,2,*,†], David Gimeno-Gómez[2,†], Elisabetta Fersini[1,†] and Carlos-D. Martínez-Hinarejos[2,†]

[1] *DISCo, Università degli Studi di Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy*
[2] *PRHLT, Universitat Politècnica de València, Camino de Vera, s/n, 46022 Valencia, Spain*

Abstract

Warning: This paper contains examples of language and images which may be offensive.

With the increasing influence of social media platforms, new forms of expression have gained popularity, encouraged by their immediate communication and sharing capabilities. Unfortunately, this accessibility has also enabled the dissemination of hateful messages, including those rooted in historical prejudices like misogyny, often manifesting in memes. The development of automated systems capable of detecting instances of sexism and other hateful expressions in this context poses significant challenges due to the multimodal nature of memes, the presence of irony, diverse categories of hate, and varied author intentions, particularly within the learning with disagreements regime. This paper presents the PINK team's participation in the EXIST (sEXism Identification in Social neTworks) Lab at CLEF 2024. Focused on Task 4, which addresses sexism identification and characterization in memes under the learning with disagreements paradigm, we proposed a unified, multi-modal Transformer-based architecture capable of dealing with multiple languages, namely English and Spanish. Our approach reached the 10th and 20th places in the final ranking for soft- and hard-label evaluations, respectively. This has been possible thanks to the use of well-established, state-of-the-art multilingual models, such as mBERT and CLIP, for feature extraction, as well as comprehensive ablation studies and the design of various model ensemble strategies. The source code of our approaches is publicly available at https://github.com/giulia95/PINK-at-EXIST2024/.

Keywords

Sexism Characterization, Learning with Disagreements, Perspectivism, Memes, Ensemble,

1. Introduction

Sexism in online content presents a significant challenge, considering the ease of sharing on social media and the various forms in which abusive messages can be represented (text, image, video, meme, ...). Moreover, the subjectivity of the tasks, typical for hate-related tasks, requires considering different interpretations and perspectives of the conveyed message that might be influenced by cultural differences and beliefs [1, 2]. Traditional sexism detection systems usually rely on predefined labels derived from a fixed definition of sexism that only represents a single perspective and, therefore, are not able to capture the complexity and the subjectivity of the task. Although detecting sexism is crucial, it becomes even more challenging due to its subjectivity and the various forms in which such messages can be expressed, such as memes. In memes, in fact, a hateful message might be represented by the image, the text, or via their combination. Moreover, due to the presence of irony, the conveyed message might appear harmless at first glance [3, 4, 5, 6].

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

[*] Corresponding author.

[†] These authors contributed equally.

g.rizzi10@campus.unimib.it (G. Rizzi); dagigo1@dsic.upv.es (D. Gimeno-Gómez); elisabetta.fersini@unimib.it (E. Fersini); cmartine@dsic.upv.es (Carlos-D. Martínez-Hinarejos)
https://www.unimib.it/giulia-rizzi (G. Rizzi); https://www.prhlt.upv.es/david-gimeno/ (D. Gimeno-Gómez); https://www.unimib.it/elisabetta-fersini (E. Fersini); https://personales.upv.es/carmarhi/ (Carlos-D. Martínez-Hinarejos)
0000-0002-0619-0760 (G. Rizzi); 0000-0002-7375-9515 (D. Gimeno-Gómez); 0000-0002-8987-100X (E. Fersini); 0000-0002-6139-2891 (Carlos-D. Martínez-Hinarejos)

An important contribution in this field that focuses on the problem of sexism identification under the paradigm of learning with disagreements in memes is represented by Task 4 at EXIST 2024: sEXism Identification in Social neTworks [7, 8]. In this paper, we address the task by proposing a unified, multi-modal Transformer-based architecture capable of dealing with multiple languages (English and Spanish). The proposed approach exploits well-established, state-of-the-art multilingual models (i.e., mBERT and CLIP) and ensemble strategies.

The paper is organized as follows. An overview of the state of the art is provided in Section 2 focusing on Sexism Detection. The proposed method is described in Section 3, including both details about the shared task and dataset, and the description of the architecture of the proposed model. In Section 4, the results achieved by the proposed approaches are reported. Finally, conclusions and future research directions are summarized in Section 5.

2. Related Work

Social media platforms provide a fertile environment for users to share their opinions and ideologies through various forms of expression, including text, memes, and videos. Often encouraged by anonymity, elements that convey hateful messages are just as easily represented and disseminated. As a consequence, hateful messages, also linked to historical aversions (e.g., hatred towards women), have found new ways of expression, for example, in memes [9]. Likewise, the interest of researchers have as a consequence, researchers' interest has expanded to consider such representative forms.

Hateful content identification. The research field dedicated to the identification of hateful content towards women is articulated into the identification of misogynistic or sexist content, also considering the different ways in which this type of hate can be expressed (e.g., by means of stereotypes, objectification, etc.). In particular, the majority of work in sexism/misogyny detection focuses on text (mostly Tweets) [10, 11, 12, 13], and only in recent years it has expanded to include multimodal content, for instance, memes [14, 15, 16]. A first insight to counter sexist memes was proposed in [14], in which both unimodal and multimodal approaches are investigated to evaluate the contribution of textual and visual modality in sexism identification. Similarly, in [15], the authors also evaluate the information content introduced by both modalities, identifying the visual component as more informative.

As a consequence of the growing attention of researchers in the sector, challenges and benchmark datasets dedicated to this research area have been proposed. An initial dataset, proposed in [17], is composed of 800 memes gathered from the most popular social media platforms, labeled both by domain experts and by annotators from a crowdsourcing platform. A similar benchmark has been realized for the *MAMI* shared task at SemEval 2022 [18]. This benchmark is composed of 11k memes divided into train and test, allowing the investigation of two different tasks: (i) the identification of misogynistic memes, and (ii) the recognition of the misogynistic type among Shaming, Stereotype, Objectification, and Violence. The majority of the works in this context exploited pre-trained models-based approaches [19, 20, 21, 22, 16] and/or investigated ensemble strategies [23, 24, 25, 26].

Learning with Disagreements. For what concerns the learning with disagreements paradigm, up to our knowledge, EXIST 2024 [7, 8] represents the first insight for multimodal sexism detection in memes, including perspectivism. Previous works under the learning with disagreement paradigm consider only textual expression. In this area, the main contribution is represented by [27], and by the previous edition of EXIST [28]. In particular, in the Learning With Disagreements (LeWiDi) challenge [27], four different datasets with different characteristics in terms of types, languages, goals, and annotation methods are proposed, including, for each instance, both hard labels (an aggregated *hateful/non-hateful* label) and soft labels (representing agreement among annotators). Similarly, the previous edition of EXIST [28] aimed at capturing sexism in all its forms while considering the perspective of the learning with disagreements paradigm. The challenge, articulated in different tasks, addressed sexism identification at different granularities and perspectives. The approach proposed by the participants of such challenges mostly relied on fine-tuned pre-trained models and/or ensemble methods [29, 30, 28].

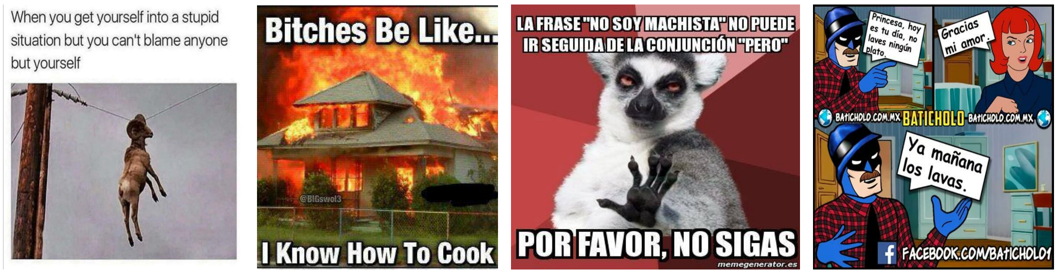
3. Method

3.1. Task Description

The *sEXism Identification in Social neTworks* task at CLEF 2024 [7, 8] aims at addressing the problem of sexism identification at a broad spectrum of sexism manifestation in social networks. Unlike previous editions of the same challenge [28], which focused solely on detecting and classifying sexist textual messages, the current edition also incorporates new tasks that address the same problems in a different form of representation: memes. As for the previous edition, the challenge embraces the Learning with Disagreements paradigm. For both tweets and memes, three different tasks are proposed that address the problem of sexism identification at different granularities, addressing (i) the identification of sexist messages, (ii) the author's intentions, and (iii) sexism categorization. This paper only focuses on **Task 4**, a binary classification task that consists of deciding whether or not a given meme is sexist.

3.2. Dataset

The meme dataset proposed for EXIST 2024 is composed of more than 3000 memes per language (considering English and Spanish). Memes that compose the dataset were collected via search query on Google Images exploiting 250 terms with varying degrees of use in both sexist and non-sexist contexts, all centered around women. Examples of memes are reported in Figure 1.


When you get yourself into a stupid situation but you can't blame anyone but yourself
Bitches Be Like...
@BIGswol3
I Know How To Cook
LA FRASE "NO SOY MACHISTA" NO PUEDE IR SEGUIDA DE LA CONJUNCIÓN "PERO"
POR FAVOR, NO SIGAS
memegenerator.es
Princesa, hoy es tu día, no laves ningún plato.
Gracias mi amor.
BATICHOLO.COM.MX BATICHOLO BATICHOLO.COM.MX
Ya mañana los lavas.
FACEBOOK.COM/BATICHOLO1

(a) Non-Sexist (En) (b) Sexist (En) (c) Non-Sexist (Sp) (d) Sexist (Sp)

Figure 1: Examples of memes in the EXIST meme dataset, showcasing the inherent challenges of the task. En and Sp refer to English and Spanish, respectively.

The challenge also approaches the sexism identification task from the perspective of the Learning with Disagreements paradigm. For each meme, it provides labels and personal meta-data information (e.g. age, gender, etc.) gathered from six annotators, introducing two different evaluation settings:

- **Hard Evaluation**: Systems performances are evaluated considering a hard label derived from the majority class among the different annotators' labels.
- **Soft Evaluation**: Systems performances are evaluated through a soft-soft evaluation that considers the probability distribution derived from the set of human annotators.

Unlike tasks based on Tweets, automatic sexism detection in memes, such as Task 4 addresses, did not come with a predefined validation dataset. Therefore, to conduct our hyperparameter optimization experiments, we partitioned the official training split to create a validation set, providing a basis for our initial research steps. The official training dataset comprises 4044 samples, evenly balanced in terms of sexism and non-sexism instances across both languages. From this dataset, we selected 20% of the samples for our validation set, using the remaining data for training our models. Once we identified the best-performing settings, we trained our models using the entire official training set for submission. The results discussed in Section 4 are based on the performance of these models on the official test set of the challenge.

3.3. Model Architecture

This section describes the modules that compose the cross-lingual, multi-modal, Transformer-based model proposed in this paper, whose overall architecture is depicted in Figure 2.

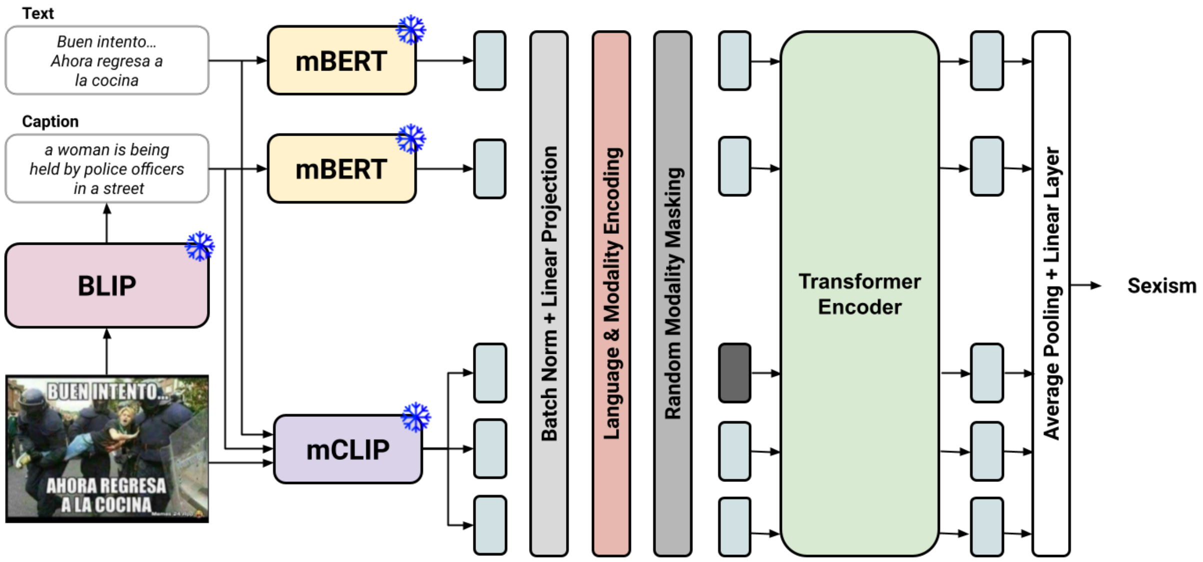

Text
Buen intento... Ahora regresa a la cocina
Caption
a woman is being held by police officers in a street
BLIP
BUEN INTENTO...
AHORA REGRESA A LA COCINA
mBERT
mBERT
mCLIP
Batch Norm + Linear Projection
Language & Modality Encoding
Random Modality Masking
Transformer Encoder
Average Pooling + Linear Layer
Sexism

Figure 2: Our proposed Cross-Lingual, Multi-Modal Transformer-based architecture extracts high-level features using large-scale, pre-trained models that are kept frozen during training. These features are then normalized and projected into the same dimensional space. They are then conditioned based on the language of the sample and its modality before being processed by a Transformer encoder backbone. The final classification is predicted through average pooling and a linear projection.

Input Data. As described above, the dataset provides both the raw image of the meme and its corresponding text. However, this text does not have to be directly and clearly related to the content represented in the image; in many cases, this relationship depends on subtle and complex social and contextual details. Therefore, to further inform the model, we also incorporated a caption of the image content as an additional input. We used the state-of-the-art BLIP[1] [31] model to automatically generate these captions. Consequently, our model processes three types of input data: the raw image, the OCR-based text, and the image caption.

Feature Extraction Frontend. We considered pre-trained, state-of-the-art models for feature extraction, namely the multilingual mBERT[2] [32] and mCLIP[3] [33]. These or similar models have been widely studied in the context of sexism detection [19, 20, 21, 22, 16] due to the multi-modality nature of memes, where the image and text are closely related or complementary, thus supporting our proposed approach. mBERT was used to extract feature embeddings for the OCR-based text and image-content captions separately, while all three types of input data, including the raw image, were processed by mCLIP to obtain their corresponding latent feature representations. As a result, we extracted five different input modalities. Note that mBERT and mCLIP remained frozen during the training process, so they were not adapted to the task. The rationale behind this decision was not only to lower the computational costs of our proposed approach, but also to investigate how robust this general, task-agnostic, pre-trained representations are in the context of sexism detection without fine-tuning.

Random Modality Masking. We also employed a modality masking strategy. By randomly masking one of the five input modalities, the model is forced to extract relevant information from all modalities, thus preventing complete reliance on any single modality. This approach enhances the model's robust-

[1] https://huggingface.co/Salesforce/blip-image-captioning-large

[2] https://huggingface.co/google-bert/bert-base-multilingual-cased

[3] https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

ness against scenarios where a modality might be absent. The effectiveness of this strategy is supported by our experimental results. Note that not always one of the input modalities has to be masked.

Transformer Encoder Backbone. The extracted feature representations are first processed with 1D batch normalization layer followed by a linear layer to project all the feature embeddings to a 256-dimensional space. Furthermore, these projected features are conditioned via learnable embeddings to enable the model to differentiate between the five input modalities and their corresponding languages. This conditioning strategy, inspired by numerous works in the field of natural language processing [32], allows the model to effectively handle multiple languages and modalities simultaneously into a unified framework. Specifically, all these input modalities are concatenated and then processed by an encoder backbone based on the Transformer architecture [34], where the model performs what can be considered as a soft cross-attention between modalities. Finally, the final classification is predicted through an average pooling of the encoder output sequence, followed by a linear layer projection.

3.4. Implementation Details

All our models were developed using the open-source PyTorch library, as well as the corresponding HuggingFace pre-trained models for feature extraction. Experiments were conducted on a GeForce RTX 2080 GPU with 8GM memory.

Model Architecture. Hyper-parameter optimization was conducted to determine the best-performing model architecture for both hard- and soft-label evaluation approaches. This process involved exploring a range of learning rates, numbers of encoder layers, attention heads, and feed-forward layer sizes. For the hard-label approach, the optimal architecture was defined as a 1-layer encoder with a self-attention module of 8 attention heads and a modality masking probability of 0.8. For the soft-label approach, we found that a 4-layer encoder with 4 attention heads and a modality masking probability of 0.7 yielded optimal results. In both cases, the latent dimension of the Transformer encoder was set to 256, while the inner hidden representation of its feed-forward modules was laid on a 2048-dimensional space.

Training Settings. In all our experiments, we used the AdamW optimizer [35] and a one-cycle linear scheduler with a batch size of 16 samples. Based on our prior experiments, we determined the optimum settings for each task. For hard-label evaluation, the model was trained for 5 epochs with a learning rate of 0.0002. For soft-label evaluation, the best results were achieved by training the model for 7 epochs with a learning rate of 0.0005. In both cases, the dropout rate was set to 0.1 and we employed the cross-entropy loss as the objective function.

Evaluation Metrics. All the reported metrics were computed using the official PyEvALL library[4]. Depending on the type of labels considered for the addressed task, different metrics were used to evaluate our models. According to the challenge instructions, we employed the F_1-Score, ICM, and ICM-Norm metrics for hard-label evaluation, and the cross entropy, ICM_{soft}, and ICM_{soft}-Norm metrics for soft-label evaluation.

Soft-Label Evaluation Postprocessing. As described in Subsection 3.2, each sample of the dataset was annotated by six annotators. Consequently, the ground truth probabilities can take one of seven possible values. Therefore, in soft-label settings, the raw logits predicted by our models were rounded to match one of the actual possible output values.

3.5. Proposed Approaches

In this work, we proposed three different approaches to address the automatic detection of sexism in memes both for the hard- and soft-label evaluation scenarios. Similar to previous studies [23, 24, 25, 26], we explored not only single-model approaches, but also adopted two model ensemble strategies to provide a more stable and robust solution:

[4] https://github.com/UNEDLENAR/PyEvALL

- **Single Model (SM).** This approach relies on a single model to obtain the final predictions. For each evaluation setting, we consider the corresponding best-performing model architecture determined through the hyperparameter optimization process described above.
- **Majority Voting Ensemble (MVE).** To provide a more stable and robust approach that does not rely solely on one specific training procedure, we adopted a model ensemble strategy. We trained five models using the best-performing architecture, each initialized with a different random seed. The final predictions were then aggregated using a majority voting ensemble. For soft-label evaluation settings, we converted the continuous soft-label predictions into discrete classes by selecting the class with the highest predicted value.
- **Average Probability Ensemble (APE).** Similarly, this model ensemble strategy involves training five models with different seeds based on the best-performing architecture. However, instead of working with discrete classes, we consider the raw logits predicted by the models and aggregate them via a non-weighted average. After combining these probabilities, we obtain the discrete classes by selecting the class with the highest predicted value.

4. Results & Discussion

Preliminary hyperparameter optimization experiments were carried out using the validation split described in Subsection 3.2. Once we determined the best-performing settings for each one of our three proposed approaches, we used the entire training set for final model estimation and submission in both hard- and soft-label evaluation scenarios. Tables 1, 2, and 3 compare our proposed approaches to the challenge baselines on the official test set in the context of the EXIST2024 Task 4. These challenge baselines are described as follows:

- **Gold.** Given that the ICM measure is unbounded, this baseline provides the best possible reference by using an oracle that perfectly predicts the ground truth.
- **Majority Class.** A non-informative baseline that classifies all instances according to the majority class based on the six annotators.
- **Minority Class.** A non-informative baseline that classifies all instances according to the minority class based on the six annotators.

Note that the ICM-Norm, both for hard- and soft-label scenarios, is considered the official evaluation metric of the challenge.

4.1. Overall Results Analysis

Table 1
Overall analysis of our proposed approaches for EXIST2024 Task 4 for both English and Spanish instances. Results reported for the test set, using variants of ICM (↑), F_1-score (↑), and cross entropy (↓) as evaluation metrics. Best results among our approaches are highlighted in **bold** for each evaluation metric.

	Hard Labels			Soft Labels		
	ICM	ICM-Norm	F_1-score	ICM$_{soft}$	ICM$_{soft}$-Norm	Cross Entropy
EXIST$_{2024}$ (Gold)	0.9832	1.0000	1.0000	3.1107	1.0000	0.5852
EXIST$_{2024}$ (Majority Class)	−0.4038	0.2947	0.6821	−2.3568	0.1212	4.4015
EXIST$_{2024}$ (Minority Class)	−0.6468	0.1711	0.0000	−3.5089	0.0000	5.5672
PINK_1 (SM)	**0.0076**	**0.5039**	0.7044	−0.4537	0.4271	**0.9282**
PINK_2 (MVE)	−0.0346	0.4824	**0.7102**	**−0.4396**	**0.4293**	0.9375
PINK_3 (APE)	−0.0053	0.4973	0.7006	−0.6378	0.3975	0.9318

Table 1 provides a comparison of the overall performance of our submitted approaches on the official test set, considering both English and Spanish instances. Our best-performing approaches achieved

10th place for soft-label evaluation settings and 20th place for hard-label evaluation settings. In general terms, the model ensemble strategy based on majority voting (MVE) stands as the most robust approach. However, the performance of the single model (SM) does not substantially differ, suggesting that this simpler approach might be more appealing for deployment as it does not require considering multiple model decisions and their associated time cost. Furthermore, the SM approach was the one providing the best results for hard-label settings in terms of ICM-Norm. This reflects the greater complexity of the soft-label prediction task and its reliance on model ensembles to offer a more accurate outcome.

4.2. Language-Specific Results Analysis

Tables 2 and 3 provide a comparison of the overall performance of our submitted approaches on the official test set for the English and Spanish instances, respectively. For Spanish, our best-performing approaches achieved 5th place for soft-label evaluation settings and 13th place for hard-label evaluation settings. For Spanish, we achieved 11th and 21st places for soft- and hard-label settings, respectively. One of the most notable findings is not only the performance gap between both languages, but also the fact that the best-performing approach varies depending on whether we are dealing with English or Spanish memes. While English mostly benefits from model ensembles in all cases, the Spanish language shows a substantial improvement when addressing the hard-label scenario with a single-model approach. For soft-label settings, both languages reflect the same trend toward reliance on ensembles.

Table 2
Overall analysis of our proposed approaches for EXIST2024 Task 4 for the English language. Results reported for the test set, using variants of ICM (↑), F_1-score (↑), and cross entropy (↓) as evaluation metrics. Best results among our approaches are highlighted in **bold** for each evaluation metric.

	Hard Labels			Soft Labels		
	ICM	**ICM-Norm**	**F_1-score**	**ICM_{soft}**	**ICM_{soft}-Norm**	**Cross Entropy**
$EXIST_{2024}$ (Gold)	0.9848	1.0000	1.0000	3.0794	1.0000	0.5528
$EXIST_{2024}$ (Majority Class)	−0.4076	0.2931	0.6880	−2.2236	0.1390	4.4798
$EXIST_{2024}$ (Minority Class)	−0.6381	0.1761	0.0000	−3.1235	0.0000	5.4888
PINK_1 (SM)	0.1413	0.5717	0.7270	−0.2760	0.4552	**0.8975**
PINK_2 (MVE)	0.1574	0.5799	**0.7422**	**−0.2703**	**0.4561**	0.9119
PINK_3 (APE)	**0.1699**	**0.5862**	0.7310	−0.5164	0.4162	0.9097

One possible explanation for the performance gap between English and Spanish could be attributed to the discrepancy in data availability used to estimate our pre-trained, state-of-the-art feature extractors. Despite these models were designed to be multi-lingual, they may have been better optimized with English data due to its abundance. Consequently, when applied to Spanish memes, these models may not generalize as effectively, resulting in lower performance. This discrepancy highlights the need for further research to address the challenges posed by multiple languages in automatic sexism detection.

Table 3
Overall analysis of our proposed approaches for EXIST2024 Task 4 for the Spanish language. Results reported for the test set, using variants of ICM (↑), F_1-score (↑), and cross entropy (↓) as evaluation metrics. Best results among our approaches are highlighted in **bold** for each evaluation metric.

	Hard Labels			Soft Labels		
	ICM	**ICM-Norm**	**F_1-score**	**ICM_{soft}**	**ICM_{soft}-Norm**	**Cross Entropy**
$EXIST_{2024}$ (Gold)	0.9815	1.0000	1.0000	3.1360	1.0000	0.6160
$EXIST_{2024}$ (Majority Class)	−0.4001	0.2962	0.6765	−2.4997	0.1014	4.3270
$EXIST_{2024}$ (Minority Class)	−0.6557	0.1660	0.0000	−3.9408	0.0000	5.6416
PINK_1 (SM)	**−0.1262**	**0.4357**	**0.6846**	−0.6524	0.3960	0.9575
PINK_2 (MVE)	−0.2267	0.3845	0.6833	**−0.6276**	**0.3999**	0.9618
PINK_3 (APE)	−0.1805	0.4080	0.6746	−0.7709	0.3771	**0.9529**

5. Conclusions

In this work, we proposed a cross-lingual and multi-modal Transformer-based approach for sexism identification under the paradigm of learning with disagreements. Although achieved results did not show a significant improvement in performances by the introduction of ensemble methods, more complex aggregation strategies might be investigated for future work to aggregate models with different input configurations. Additionally, more sophisticated ensemble strategies can be explored, such as Bayesian Model Averaging (BMA) [36]. Finally, the discrepancy in performance between different languages highlights the need for further research to effectively handle multiple languages in the context of automatic sexism detection.

Acknowledgements

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. The work of D. Gimeno-Gómez and C.-D. Martínez Hinarejos was partially supported by Grant CIACIF/2021/295 funded by Generalitat Valenciana and by Grant PID2021-124719OB-I00 under project LLEER funded by MCIN/AEI/10.13039/501100011033/ and by ERDF, EU A way of making Europe.

References

[1] Y. Sang, J. Stanton, The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation, in: Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I, Springer, 2022, pp. 425–444.

[2] M. Sandri, E. Leonardelli, S. Tonelli, E. Jezek, Why don't you do it right? analysing annotators' disagreement in subjective tasks, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2428–2441. URL: https://aclanthology.org/2023.eacl-main.178. doi:10.18653/v1/2023.eacl-main.178.

[3] T. E. Ford, M. A. Ferguson, Social consequences of disparagement humor: A prejudiced norm theory, Personality and social psychology review 8 (2004) 79–94.

[4] T. E. Ford, C. F. Boxer, J. Armstrong, J. R. Edel, More than "just a joke": The prejudice-releasing function of sexist humor, Personality and Social Psychology Bulletin 34 (2008) 159–170.

[5] B. Chulvi, L. Fontanella, R. Labadie-Tamayo, P. Rosso, et al., Social or individual disagreement? perspectivism in the annotation of sexist jokes, in: CEUR Workshop Proceedings, volume 3494, 2023.

[6] R. Labadie Tamayo, M. A. Chulvi-Ferriols, P. Rosso, Everybody hurts, sometimes overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter, Procesamiento del lenguaje natural (2023) 383–395.

[7] Plaza, Laura and Carrillo-de-Albornoz, Jorge and Ruiz, Víctor and Maeso, Alba and Chulvi, Berta and Rosso, Paolo and Amigó, Enrique and Gonzalo, Julio and Morante, Roser and Spina, Damiano , 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.

[8] 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 Multilin-

guality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[9] L. Fontanella, B. Chulvi, E. Ignazzi, A. Sarra, A. Tontodimamma, How do we study misogyny in the digital age? a systematic literature review using a computational linguistic approach, Humanities and Social Sciences Communications 11 (2024) 1–15.

[10] M. Anzovino, E. Fersini, P. Rosso, Automatic identification and classification of misogynistic language on twitter, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2018, pp. 57–64.

[11] M. A. Bashar, R. Nayak, N. Suzor, Regularising lstm classifier by transfer learning for detecting misogynistic tweets with small training set, Knowledge and Information Systems 62 (2020) 4029–4054.

[12] H. T. Ta, A. B. S. Rahman, L. Najjar, A. Gelbukh, Transfer learning from multilingual deberta for sexism identification, in: CEUR Workshop Proceedings, volume 3202, CEUR-WS, 2022.

[13] R. Calderón-Suarez, R. M. Ortega-Mendoza, M. Montes-Y-Gómez, C. Toxqui-Quitl, M. A. Márquez-Vera, Enhancing the detection of misogynistic content in social media by transferring knowledge from song phrases, IEEE Access 11 (2023) 13179–13190.

[14] 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 (ACIIW), 2019, pp. 226–231.

[15] B. O. Sabat, C. C. Ferrer, X. G. i Nieto, Hate speech in pixels: Detection of offensive memes towards automatic moderation, 2019. arXiv:1910.02334.

[16] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474.

[17] F. Gasparini, G. Rizzi, A. Saibene, E. Fersini, Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content, Data in brief 44 (2022) 108526.

[18] 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: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549.

[19] Z. Zhou, H. Zhao, J. Dong, N. Ding, X. Liu, K. Zhang, DD-TIG at semeval-2022 task 5: Investigating the relationships between multimodal and unimodal information in misogynous memes detection and classification, in: The 16th International Workshop on Semantic Evaluation, 2022, pp. 563–570.

[20] 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: The 16th International Workshop on Semantic Evaluation, 2022, pp. 636–641.

[21] S. Hakimov, G. S. Cheema, R. Ewerth, TIB-VA at SemEval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes, in: The 16th International Workshop on Semantic Evaluation, 2022, pp. 756–760.

[22] Jin Mei Zhi and Zhou Mengyuan and Mengfei Yuan and Dou Hu and Xiyang Du and Lianxin Jiang and Yang Mo and XiaoFeng Shi, PAIC at SemEval-2022 Task 5: Multi-Modal Misogynous Detection in MEMES with Multi-Task Learning And Multi-model Fusion, in: The 16th International Workshop on Semantic Evaluation, 2022, pp. 555–562.

[23] G. Balducci, G. Rizzi, E. Fersini, Bias mitigation in misogynous meme recognition: A preliminary study, in: CEUR Workshop Proceedings, volume 3596, CEUR-WS. org, 2023.

[24] W. Yu, B. Boenninghoff, J. Röhrig, D. Kolossa, Rubcsg at semeval-2022 task 5: Ensemble learning for identifying misogynous memes, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 626–635.

[25] C. Tao, J. J. Kim, taochen at semeval-2022 task 5: Multimodal multitask learning and ensemble learning, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 648–653.

[26] R. Sivanaiah, S. A. Deborah, S. M. Rajendram, T. T. Mirnalinee, TechSSN at SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification using Deep Learning Models, in: Proceedings

of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 571–574. doi:10.18653/v1/2022.semeval-1.78.

[27] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2304–2318.

[28] Plaza, Laura and Carrillo-de-Albornoz, Jorge and Morante, Roser and Amigó, Enrique and Gonzalo, Julio and Spina, Damiano and Rosso, Paolo, Overview of EXIST 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.

[29] J. A. García-Díaz, R. Pan, G. Alcaráz-Mármol, M. J. Marín-Pérez, R. Valencia-García, Umuteam at SemEval-2023 task 11: Ensemble learning applied to binary supervised classifiers with disagreements, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1061–1066.

[30] A. de Paula, G. Rizzi, E. Fersini, D. Spina, AI-UPV at EXIST 2023–Sexism Characterization Using Large Language Models Under The Learning with Disagreements Regime, in: CEUR Workshop Proceedings, volume 3497, CEUR-WS, 2023, pp. 985–999.

[31] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proc. of the 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, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[33] 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: Proc. of the 38th International Conference on Machine Learning (ICML), volume 139 of *Proc. of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, NeurIPS 30 (2017) 6000–6010. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[35] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: Proc. of ICLR, 2019.

[36] E. Fersini, E. Messina, F. A. Pozzi, Sentiment analysis: Bayesian Ensemble Learning, Decision Support Systems 68 (2014) 26–38.