NICA at EXIST CLEF Tasks 2024

Notebook for the NICA group at EXIST Lab at CLEF 2024

Aylin Naebzadeh[1,*], Melika Nobakhtian[2] and Sauleh Eetemadi[3]

[1] *Student at School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.*
[2] *Student at Tehran Institute for Advanced Studies (TeIAS), Khatam University, Tehran, Islamic Republic Of Iran.*
[3] *Assistant Professor of Computer Science, School of Computer Engineering, Iran University of Science and Technology, Tehran, Islamic Republic Of Iran.*

Abstract

In this paper, we introduce the models developed by the NICA group for the sEXism Identification in Social neTworks (EXIST) Shared Task at CLEF 2024. Our participation spanned across five tasks: Sexism Identification in Tweets (Task 1), Source Intention in Tweets (Task 2), Sexism Categorization in Tweets (Task 3), Sexism Identification in Memes (Task 4), and Source Intention in Memes (Task 5). For the first three tasks, we utilized various multi-lingual transformer models to detect sexism in English and Spanish tweets. For tasks 4 and 5, we employed the CLIP model, which leverages both image and corresponding text data to identify sexist elements. Our final model, as demonstrated through a comparative analysis with other transformer-based models, effectively leverages the multi-lingual transformer models to achieve competitive performance with hard labels. Notably, our model also yields promising results for the fourth and fifth subtasks, showcasing the efficacy of CLIP as a multi-modal Vision and Language (V&L) model. The EXIST shared task proposed two evaluation methods: Hard-Hard, and Soft-Soft, comparing the system's output with the ground truth. Our team secured the 4^{th} position in Task 5 for the Soft-Soft evaluation method, and the 9^{th} position in Task 4 for the Soft-Soft evaluation among the participants, achieving 0.3370 and 0.4299 of the ICM metric respectively.

Keywords

Sexism Identification, Transformers, Multi-lingual Transformers, CLIP

1. Introduction

Gender-based discrimination, particularly sexism, remains a significant challenge in digital interactions, affecting the inclusivity of online spaces. The proliferation of social media has intensified the spread of sexist content, underscoring the need for automated detection and classification methods.

To address this issue, a series of scientific events called EXIST has been established with the objective of comprehending sexism in its widest scope. Unlike its predecessors, EXIST 2024 expands its scope to include image-based content, such as memes, recognizing the diverse manifestations of sexism, from overt misogyny to subtle, implicit behaviors.

At EXIST 2024, participants are tasked with six challenges: identifying sexism in tweets and memes, determining the author's intention behind sexist content, and categorizing the aspects of women targeted by sexist messages. An important proposal of the task is "*The learning with disagreement paradigm*"[1] where the organizers propose to build systems that are able to consider the different perspectives that people have when identifying sexism. For this reason, task organizers propose two evaluation methods (Hard-Hard, and Soft-Soft). This paper details the NICA team's approach to the first five tasks of EXIST 2024, marking our inaugural participation. This is our first time participating in the EXIST competition. We utilized multi-lingual transformer models for text-based sexism detection

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

[*] Corresponding author.

a_naeb@comp.iust.ac.ir (A. Naebzadeh); M.Noubakhtian.5740@Khatam.ac.ir (M. Nobakhtian); sauleh@iust.ac.ir (S. Eetemadi)
https://github.com/AylinNaebzadeh (A. Naebzadeh); https://github.com/MelikaNobakhtian (M. Nobakhtian); http://www.sauleh.ir/ (S. Eetemadi)
0000-0001-9788-2764 (M. Nobakhtian)

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

and the CLIP model for categorizing sexism in images. This is our first time participating in the EXIST competition. The source code supporting our findings is available at this Github repository.

This work is structured as follows: Section 2 briefly provides a description of several earlier studies. Section 3 will then present an explanation of tasks. Following that, Section 4 and 5 will outline the experimental methodology and evaluation results respectively. Finally, in Section 6, we will present the key findings and conclusions of our studies, as well as some potential directions for future research.

2. Related Works

Fundamentally, sexism identification is categorised as a subtask of abusive language detection. It shares a close relationship with a number of abusive language detection, including racism, hate speech, personal attacks, and others. We consider sexism identification a problem of classification, where the models will classify which predefined labels a given content belongs to. For example, EVALITA [2] and AMI [3] focus on the identification of misogyny, while HateEval [4] focuses on the detection of hate speech directed against women and immigrants. In addition, shared tasks such as EDOS [5] aim to develop more accurate and explainable systems for sexism detection, and EXIST [6, 7, 8, 9] attempts to classify sexism according to the different facets of women that are affected. The efforts of increasing the scope of sexism detection, contribute to a more complete understanding of the different types of sexism and how they are expressed.

Until the recent past, machine learning techniques, like Recurrent Neural Networks (RNNs) and more classic algorithms, have been widely adopted for the classification of social media posts as sexist or non-sexist by capturing sequential dependencies in the text [10, 11].

With the advent of transformers since the seminal work of "Attention is All You Need" [12], the utilization of transformers for classification tasks has garnered significant attention. Transformers have demonstrated superior accuracy compared to their predecessors. Notably, XLM models have recently achieved state-of-the-art performance in various benchmark tasks [13, 14] and they have emerged as a powerful tool for text classification tasks, particularly in the domain of cross-lingual language modeling. XLM, which stands for Cross-lingual Language Model, is specifically designed to effectively handle multiple languages [14]. This cross-lingual capability enables XLM transformers to generalize well to languages with limited training data, facilitating effective knowledge transfer from high-resource to low-resource languages [15].

In EXIST 2022, ensembles of different language specific transformer models, including BERTweet-large [16], RoBERTa [17], DeBERTa v3 [18] for English, and BETO, BERTIN [19], and RoBERTuito [20] for Spanish, achieved the best results.

In the previous edition of EXIST, the Roh_Neil [21] team was able to win the 1^{st} rank for Hard-Hard evaluation method by applying XLM-RoBERTa-Large-Twitter. There were also other applications of RoBERTuito and BETO [22] for solving this challenge which have again achieved comparative results especially for Spanish tweets.

Sexism identification in textual data presents a significant challenge, and the complexity increases further when dealing with multimodal content. Memes, a prominent source of misogyny and hate speech, exemplify this challenge as they combine textual and image data. Effectively incorporating both modalities for sexism detection remains an ongoing effort.

The introduction of SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification (MAMI) marked a significant step forward, offering one of the first meme datasets for misogyny detection [23]. Analyzing submitted approaches revealed two primary strategies:

- Text-based models: Primarily utilized BERT [24] and RoBERTa [17] for textual analysis.
- Image-based models: Mostly adopted VisualBERT [25] for image analysis.

A smaller number of submissions explored multimodal approaches, leveraging models like CLIP [26] and ViLBERT [27] to jointly learn from both text and image data.

3. Tasks Description in EXIST 2024

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. All the five tasks in the hierarchy are classification tasks. Tasks 1, 2, 3, 4 and 5 are in the area of classification (binary and multi-class classification) and the tasks 3 and 6 are multi-label classification (categorization). Figure 1 shows an overview of tasks in EXIST 2024.

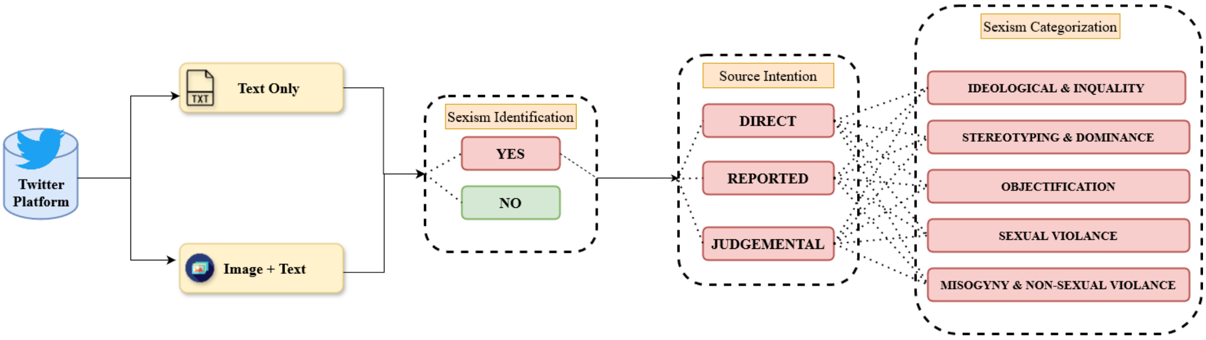


Figure 1: An Overview of Tasks in EXIST 2024

3.1. Task 1 - Binary Classification in Tweets

The first subtask is a binary classification. The systems must decide whether a given tweet contains sexist expressions or behaviours (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour), and classify it according to two categories: **YES** and **NO**.

3.2. Task 2 - Multi-class Classification in Tweets

The second subtask is a multi-class classification. For the tweets that have been predicted as sexist, the second task aims to classify each tweet according to the intention of the person who wrote it. One of the three following categories must be assigned to each sexist tweet:

- DIRECT
- REPORTED
- JUDGEMENTAL

3.3. Task 3 - Multi-label Classification in Tweets

The third subtask is a multi-label classification. For the tweets that have been predicted as sexist, the third task aims to categorize them according to the type of sexism. We propose a five-class classification task: This is a multi-label task, so that more than one of the following labels may be assigned to each tweet:

- IDEOLOGICAL-INEQUALITY
- STEREOTYPING-DOMINANCE
- OBJECTIFICATION
- SEXUAL-VIOLENCE
- MISOGYNY-NON-SEXUAL-VIOLENCE

3.4. Task 4 - Binary Classification in Tweets in Memes

This is a binary classification task consisting of deciding whether or not a given meme is sexist.

3.5. Task 5 - Multi-class Classification in Memes

As in task 2, this task aims to categorize the meme according to the intention of the author, which provides insights in the role played by social networks on the emission and dissemination of sexist messages. Due to the characteristics of the memes, the **REPORTED** label is virtually null, so in this task systems should only classify memes with **DIRECT** or **JUDGEMENTAL** labels.

3.6. Task 6 - Multi-label Classification in Memes

This subtask is a multi-label classification. This task aims to classify sexist memes according to the categorization provided for Task 3: (i) **IDEOLOGICAL AND INEQUALITY**, (ii) **STEREOTYPING AND DOMINANCE**, (iii) **OBJECTIFICATION**, (iv) **SEXUAL VIOLENCE**, and (v) **MISOGYNY AND NON-SEXUAL VIOLENCE**.

4. Methodology

Our approach can be divided into two groups based on the tasks and the datasets used. For Tasks 1, 2, and 3, which involve the detection and categorization of sexism in tweets, we utilized transformer models. Conversely, Tasks 4 and 5 required handling a completely different dataset comprising memes the use of a multi-modal model. The methodologies for these two groups of tasks are detailed in the following sections and illustrated in Figure 2.

4.1. Data

To conduct our experiments, we used the dataset provided by the organizers for EXIST 2024. This dataset includes comments extracted from Twitter that may contain popular sexist expressions and terms in both English and Spanish. A total of six annotators with diverse socio-demographic backgrounds, such as gender (male or female) and age groups (18-22, 23-45, or 46+), labeled the dataset. Instead of providing a single gold label for each text, the organizers supplied the labels assigned by each annotator along with their personal information, such as gender and age, for all tasks. This approach aims to capture the diversity of perspectives in a subjective task like sexism detection. Additionally, this version of EXIST includes a new dataset for the newly added Tasks 4 and 5, which consist of memes—images that contain a piece of text. The organizers provided gold labels for the training and development sets, which we concatenated based on unique ID values to maintain consistency.

4.1.1. Tasks 1-2-3

The data is split into a training and development set with the former containing 6,920 and 1,038 respectively. The test set contains 2,076 tweets in English, Spanish, and a mix of both languages.

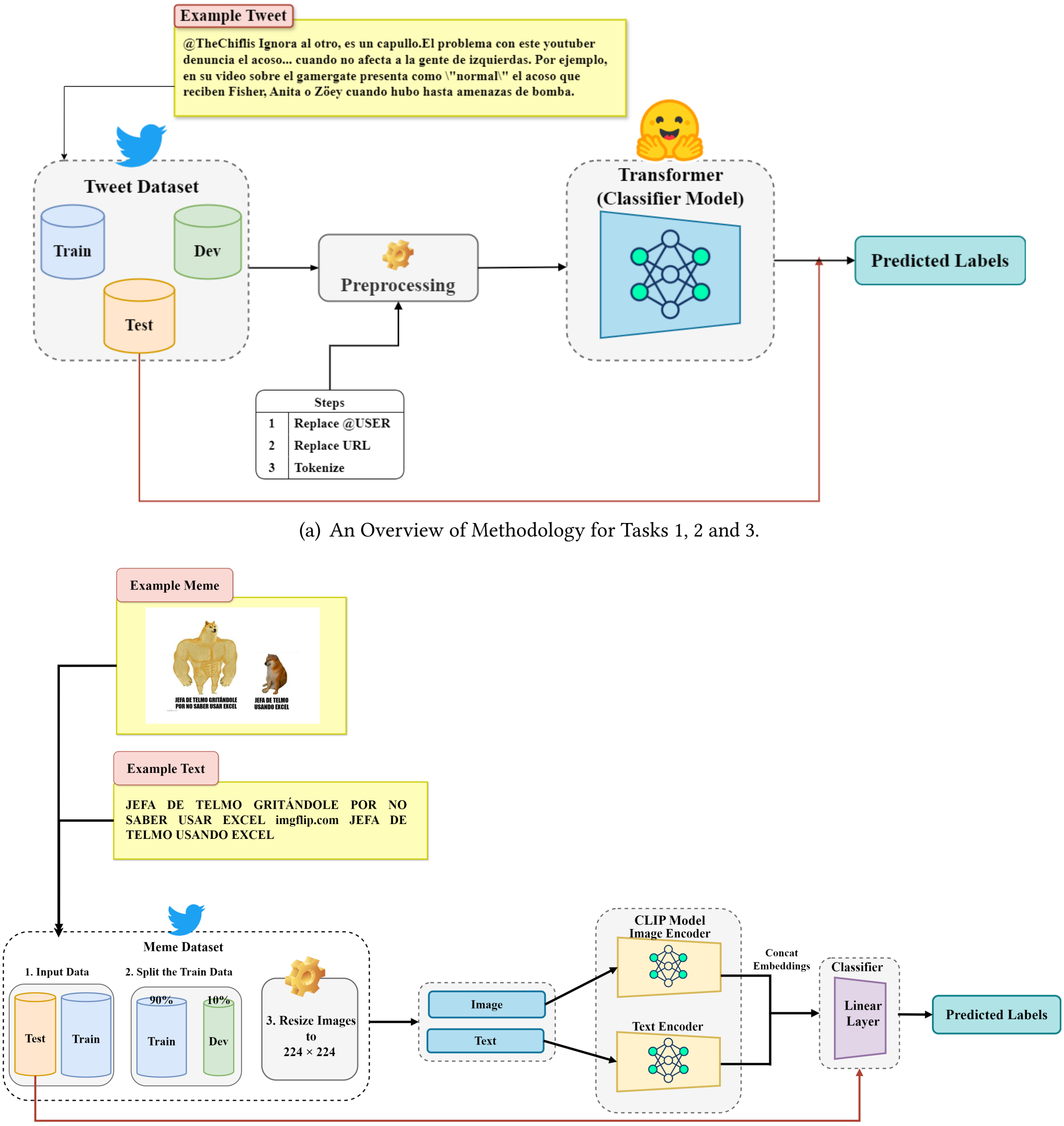
4.1.2. Tasks 4-5

Our data was divided into training and development sets for both tasks using a 90/10 split. This ensured a representative distribution of English and Spanish data in both sets.

4.2. Pre-Processing

4.2.1. Tasks 1-2-3

In our study, we adopt a straightforward pre-processing approach, which aligns with the recommended practices outlined in the usage guide of the XLM-T-10-L and other transformer based models. The pre-processing method focuses on handling tags and URLs present in the tweets. Any user handle encountered in the tweets is replaced with “@USER”, while URLs are replaced with “#HTTPURL”. We refrain from implementing additional pre-processing steps in order to retain the intricacies and unique



@TheChiflis Ignora al otro, es un capullo.El problema con este youtuber denuncia el acoso... cuando no afecta a la gente de izquierdas. Por ejemplo, en su video sobre el gamergate presenta como \"normal\" el acoso que reciben Fisher, Anita o Zöey cuando hubo hasta amenazas de bomba.

(a) An Overview of Methodology for Tasks 1, 2 and 3.

JEFA DE TELMO GRITÁNDOLE POR NO SABER USAR EXCEL imgflip.com JEFA DE TELMO USANDO EXCEL

(b) An Overview of Methodology for Tasks 4 and 5.

Figure 2: An Overview of Methodology with respect to Tasks.

characteristics inherent in the tweet data. By avoiding excessive pre-processing, we aim to preserve the originality and nuances of the tweet content, allowing the models to capture the genuine nature of the tweets during the subsequent analysis and classification stages.

4.2.2. Tasks 4-5

For text data, we do not apply any specific pre-processing. However, for image data, we resize the images to dimensions of 224 x 224 pixels to ensure compatibility with the CLIP model.

Table 1
Best Results Summary

Task No.	Version	Language	Rank	# Run
Task 5	Soft-Soft	All	4	1
		ES	4	1
		EN	4	1
	Hard-Hard	All	6	1
		ES	4	1
		EN	7	1
Task 4	Soft-Soft	All	9	1
		ES	10	1
		EN	9	1
	Hard-Hard	All	11	1
		ES	13	1
		EN	12	1
Task 3	Soft-Soft	All	12	2
		ES	13	2
		EN	12	2
	Hard-Hard	All	13	2
		ES	17	1
		EN	16	2
Task 2	Soft-Soft	All	28	2
		ES	31	2
		EN	28	2
	Hard-Hard	All	15	2
		ES	16	2
		EN	12	2
Task 1	Soft-Soft	All	37	1
		ES	38	1
		EN	37	2
	Hard-Hard	All	19	1
		ES	18	1
		EN	21	1

4.3. Model

4.3.1. Tasks 1-2-3

We decided to exclusively utilize pre-trained Transformer models based on their impressive performance in previous editions of EXIST. To ensure both robustness and ease of implementation, we selected HuggingFace's [28] Trainer API for minimizing the need for custom code development. Our experimentation involved a comprehensive evaluation of existing General Language and Tweet-specific language (NLP) models. We tried several multi-lingual transformers including sdadas/xlm-roberta-large-twitter, google-bert/bert-base-multilingual-uncased, distilbert/distilbert-base-multilingual-cased[29], and FacebookAI/xlm-roberta-base, with common hyperparameters that are considered best practices. However, hardware limitations posed significant challenges. For example, we were unable to load the ai-forever/mGPT [30] model, as our code had crashed during the training phase due to its extensive number of parameters.

Table 2
Hyperparameter Settings for Our Systems in Tasks 1, 2 and 3

Task	Run	Model	Epochs	Train Batch size	Eval Batch size	LR	Weight Decay	Val F1
1	-	DBMLC	5	8	8	$2e-5$	0.01	0.7761
1	3	DBMLC	5	8	8	$3e-5$	0.01	0.7812
1	-	DBMLC	5	8	8	$4e-5$	0.01	0.7471
1	-	XLM-B	5	8	5	$2e-5$	0.01	0.8268
1	1	XLM-L-T	4	8	16	$2e-5$	0.01	0.8527
1	-	BBMUC	5	8	16	$2e-5$	0.01	0.8123
1	2	BBMUC	5	8	8	$3e-5$	0.01	0.8980
1	-	BBMUC	5	8	8	$4e-5$	0.01	0.8621
2	1	XLM-L-T	4	8	16	$2e-5$	0.01	0.4894
2	2	BBMUC	4	8	8	$1e-5$	0.01	0.7342
3	1	XLM-L-T	4	8	16	$2e-5$	0.01	0.5449
3	2	BBMUC	4	8	8	$3e-5$	0.01	0.5849

*DBMLC → DistilBert Base Multilingual Cased.
*XLM-L-T → XLM Roberta Large Twitter.
*XLM-B → XLM Roberta Base.
*BBMUC → Bert Base Multilingual Uncased.

For these three tasks, we send the following runs:

- Run 1: The model for this run is *sdadas/xlm-roberta-large-twitter*. The hyper-parameter setting for this model is $2e-5$ for learning rate, 0.01 for weight decay, 8 and 16 for train and validation batch size, respectively.
- Run 2: The model for this run is *google-bert/bert-base-multilingual-uncased*. The hyper-parameter setting for this model is $3e-5$ for learning rate, 0.01 for weight decay, 8 and 8 for both train and validation batch sizes.
- Run 3: The model for this run is *distilbert/distilbert-base-multilingual-cased*. The hyper-parameter setting for this model is $3e-5$ for learning rate, 0.01 for weight decay, 8 and 8 for both train and validation batch sizes.

Table 2 shows the hyperparameter values for each experiment with the computed F1-Score on validation dataset. One of the key observations from our experiments is that model *google-bert/bert-base-multilingual-uncased* consistently outperforms other models, especially in tasks 2 and 3. This performance discrepancy can largely be attributed to the simplicity and reduced number of parameters in model *google-bert/bert-base-multilingual-uncased*. Given the limited amount of input data available for training, simpler models with fewer parameters tend to generalize better, as they are less prone to overfitting. Overfitting occurs when a model learns the noise and details in the training data to the extent that it negatively impacts the model's performance on new, unseen data.

4.3.2. Tasks 4-5

We employed the CLIP model [26] to achieve joint learning from textual and image modalities. CLIP encodes both the image and its corresponding text into embedding vectors. These embeddings were then concatenated to form a combined representation. A subsequent linear layer was applied to this combined representation to predict the most probable class for the given task.

For Task 4, we adopted the openai/clip-vit-base-patch32 model, while Task 5 employed the openai/clip-vit-large-patch14 model.

5. Evaluation

The shared task results are evaluated in three different scenarios, we focus on HARD-HARD for the first three tasks, but for the tasks 4 and 5 we concentrate on Hard-Hard and Soft-Soft. For the HARD-HARD evaluation, the gold label consists of the class annotated by the majority of annotators. Any items with no majority class are removed from the evaluation. Here, the model produces a single label, which is compared to the gold label. In the SOFT-SOFT evaluation, the system provides probabilities for each class, and this distribution is evaluated against the distribution of the annotator decisions. The official metric for the shared task is the “Information Contrast Measure” (ICM) [31]. A description of evaluation metrics can be viewed in Table 3. Our models achieved notable success, ranked 4^{th} place in Task 5 and 9^{th} place in Task 4 for the Soft-Soft evaluation, with ICM scores of 0.3370 and 0.4299, respectively. A brief summary of our best results for each task can be found in Table 1, but a complete overview of the final results of this study’s submissions with all the computed metrics can be found in tables 4 to 8 which have been provided by organizers in leader boards.

Table 3
Description of Evaluation Metrics

Metric	Description
ICM	Information Contrast Measure (ICM) is a similarity function used to evaluate the outputs of classification systems in hierarchical classification tasks. It generalizes Pointwise Mutual Information (PMI) and measures the resemblance between the system’s output and the ground truth labels. Higher values of ICM indicate better performance.
ICM-Soft	ICM-Soft is an evaluation metric that compares the categories assigned by the system with the probabilities assigned to each category in the ground truth. It considers the distribution of labels and the number of annotators assigned to each instance to determine the probability of the classes. Higher values of ICM-Soft indicate better performance.
ICM-Hard	ICM-Hard evaluation involves comparing the system’s "hard" output with the hard ground truth labels. A probabilistic threshold is employed to extract the hard-labels from the ground truth, considering the approval of multiple annotators for each task. Only the most popularly labeled classes are included in this evaluation. Higher values of ICM-Hard indicate better performance.
ICM-Soft Norm	ICM-Soft Norm is a normalized version of ICM-Soft that takes into account the number of annotators assigned to each instance and adjusts the probabilities accordingly. It handles instances labeled as "UNKNOWN" by reducing the number of annotators considered based on the count of "UNKNOWN" labels associated with them. Higher values of ICM-Soft Norm indicate better performance.
ICM-Hard Norm	ICM-Hard Norm is a normalized version of ICM-Hard that adjusts the hard-labels based on the number of annotators assigned to each instance. It considers instances labeled as "UNKNOWN" and adjusts the threshold for label extraction accordingly. Higher values of ICM-Hard Norm indicate better performance.
F1 Score	F1 Score is a commonly used evaluation metric in classification tasks. It is the harmonic mean of precision and recall, weighted by the same values. The F1 score treats false positives and false negatives equally, assuming that both types of errors have the same consequences. Higher values of F1 Score indicate better performance.
Cross Entropy	Cross Entropy is a metric used to measure the difference between the predicted probabilities and the true probabilities. It quantifies the average amount of information needed to identify the true class given the predicted probabilities. Lower values of Cross Entropy indicate better model performance.

6. Conclusion and Future Works

Sexism in digital interactions remains a widespread issue, significantly impacting the creation of inclusive and respectful online environments. The proliferation of social media has amplified the spread of sexist content, highlighting the urgent need for automated detection and classification methods. Our work in the EXIST CLEF 2024 shared task showcases the potential of multi-lingual transformer models and the CLIP model in identifying and understanding sexism and source intention in social media content. Despite the challenges and limitations of our hardware resources, our models achieved notable success. Moving forward, we plan to implement various pre-processing techniques, data augmentation strategies, and explore newer transformer models, as well as innovative classification approaches like few-shot [32] and zero-shot learning [33]. Additionally, we will consider using more interpretable machine learning algorithms, such as XGBoost [34], and ensemble methods [35]. By evaluating and comparing these diverse methods, we hope to enhance the robustness and accuracy of our models.

Acknowledgments

We would like to express our deep appreciation to the organizers of EXIST CLEF 2024 for providing us with the opportunity to participate in this esteemed event. It was all a great and valuable experience to work with the provided data, and gain new knowledge. We are truly grateful for the chance to present our work and contribute to the progress of our field. We would like to once again acknowledge and appreciate the support and encouragement received from all individuals involved, as their contributions have been instrumental in our accomplishments.

References

[1] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, , M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 38 (2021) 1385–1470.

[2] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, et al., Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy, 2023.

[3] 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, 2024, pp. 1864–1871.

[4] T. Sen, A. Das, M. Sen, Hatetinyllm: Hate speech detection using tiny large language models, arXiv preprint arXiv:2405.01577 (2024).

[5] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: explainable detection of online sexism, arXiv preprint arXiv:2303.04222 (2023).

[6] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023–learning with disagreement for sexism identification and characterization, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 316–342.

[7] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

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

[9] 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.

[10] A. Kalra, A. Zubiaga, Sexism identification in tweets and gabs using deep neural networks, arXiv preprint arXiv:2111.03612 (2021).

[11] M. Buzzell, J. Dickinson, N. Singh, S. Kübler, Iu-nlp-jedi: investigating sexism detection in english and spanish, Working Notes of CLEF (2023).

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[13] 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, arXiv preprint arXiv:1911.02116 (2019).

[14] G. Lample, A. Conneau, Cross-lingual language model pretraining, arXiv preprint arXiv:1901.07291 (2019).

[15] M. Artetxe, S. Ruder, D. Yogatama, On the cross-lingual transferability of monolingual representations, arXiv preprint arXiv:1910.11856 (2019).

[16] D. Q. Nguyen, T. Vu, A. T. Nguyen, Bertweet: A pre-trained language model for english tweets, arXiv preprint arXiv:2005.10200 (2020).

[17] 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 preprint arXiv:1907.11692 (2019).

[18] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).

[19] J. De la Rosa, E. G. Ponferrada, P. Villegas, P. G. d. P. Salas, M. Romero, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, arXiv preprint arXiv:2207.06814 (2022).

[20] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, arXiv preprint arXiv:2111.09453 (2021).

[21] R. Koonireddy, N. Adel, Roh_neil@ exist2023: detecting sexism in tweets using multilingual language models, Working Notes of CLEF (2023).

[22] H. Asnani, A. Davis, A. Rajanala, S. Kübler, Tlatlamiztli: fine-tuned robertuito for sexism detection, Working Notes of CLEF (2023).

[23] 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-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74. doi:10.18653/v1/2022.semeval-1.74.

[24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[25] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019. arXiv:1908.03557.

[26] 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, 2021. arXiv:2103.00020.

[27] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. arXiv:1908.02265.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[29] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[30] O. Shliazhko, A. Fenogenova, M. Tikhonova, V. Mikhailov, A. Kozlova, T. Shavrina, mgpt: Few-shot learners go multilingual, arXiv preprint arXiv:2204.07580 (2022).

[31] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.

[32] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, ACM computing surveys (csur) 53 (2020) 1–34.

[33] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–37.

[34] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[35] T. G. Dietterich, Ensemble methods in machine learning, in: International workshop on multiple classifier systems, Springer, 2000, pp. 1–15.

A. Online Resources

The source code and the final submission files can be accessed through the following official GitHub repository for EXIST2024:

- Github for EXIST2024 NICA

B. Evaluation Results Provided in Leader board

Table 4
Task 1 Evaluation Results

# Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm	Cross Entropy	F1-Score	Lang
EXIST2024-test_gold	0	3.1182	-	1.0000	-	0.5472	-	All
3	37	-2.8848	-	0.0374	-	1.5286	-	All
2	38	-2.8848	-	0.0374	-	1.3862	-	All
1	39	-2.8848	-	0.0374	-	1.2301	-	All
EXIST2024-test_gold	0	-	0.9948	-	1.0000	-	1.0000	All
1	19	-	0.5214	-	0.7621	-	0.7642	All
3	37	-	0.4358	-	0.7191	-	0.7429	All
2	43	-	0.3750	-	0.6885	-	0.7263	All
EXIST2024-test_gold	0	3.1177	-	1.0000	-	0.5208	-	ES
1	38	-2.8670	-	0.0402	-	1.1916	-	ES
3	39	-2.8671	-	0.0402	-	1.5302	-	ES
2	40	-2.8671	-	0.0402	-	1.3733	-	ES
EXIST2024-test_gold	0	-	0.9999	-	1.0000	-	1.0000	ES
1	18	-	0.5156	-	0.7578	-	0.7852	ES
3	36	-	0.4223	-	0.7112	-	0.7627	ES
2	44	-	0.3517	-	0.6759	-	0.7419	ES
EXIST2024-test_gold	0	3.1141	-	1.0000	-	0.5770	-	EN
2	37	-2.9063	-	0.0334	-	1.4007	-	EN
3	38	-2.9064	-	0.0333	-	1.5268	-	EN
1	39	-2.9064	-	0.0333	-	1.2734	-	EN
EXIST2024-test_gold	0	-	0.9798	-	1.0000	-	1.0000	EN
1	21	-	0.5122	-	0.7614	-	0.7362	EN
3	35	-	0.4393	-	0.7242	-	0.7176	EN
2	38	-	0.3871	-	0.6975	-	0.7057	EN

Table 5
Task 2 Evaluation Results

# Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm	Cross Entropy	F1-Score	Lang
EXIST2024-test_gold	0	6.2057	-	1.0000	-	0.9128	-	All
2	28	-5.7592	-	0.0360	-	2.7026	-	All
EXIST2024-test_gold	0	-	1.5378	-	1.0000	-	1.0000	All
2	15	-	0.1506	-	0.5490	-	0.4738	All
1	40	-	-0.9504	-	0.1910	-	0.1603	All
EXIST2024-test_gold	0	6.2431	-	1.0000	-	0.8926	-	ES
2	31	-5.7501	-	0.0395	-	2.6715	-	ES
EXIST2024-test_gold	0	-	1.6007	-	1.0000	-	1.0000	ES
2	16	-	0.1567	-	0.5490	-	0.4904	ES
1	40	-	-1.0391	-	0.1754	-	0.1545	ES
EXIST2024-test_gold	0	6.1178	-	1.0000	-	0.9354	-	EN
2	28	-5.7285	-	0.0318	-	2.7374	-	EN
EXIST2024-test_gold	0	-	1.4449	-	1.0000	-	1.0000	EN
2	12	-	0.1213	-	0.5420	-	0.4516	EN
1	39	-	-0.8529	-	0.2048	-	0.1667	EN

Table 6
Task 3 Evaluation Results

# Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm	Cross Entropy	F1-Score	Lang
EXIST2024-test_gold	0	9.4686	-	1.0000	-	-	-	All
2	12	-4.4324	-	0.2659	-	-	-	All
EXIST2024-test_gold	0	-	2.1533	-	1.0000	-	1.0000	All
2	13	-	-0.2383	-	0.4447	-	0.4564	All
1	19	-	-0.3258	-	0.4243	-	0.3867	All
EXIST2024-test_gold	0	9.6071	-	1.0000	-	-	-	ES
2	13	-4.5491	-	0.2632	-	-	-	ES
EXIST2024-test_gold	0	-	2.2393	-	1.0000	-	1.0000	ES
1	17	-	-0.3007	-	0.4329	-	0.4023	ES
2	18	-	-0.3611	-	0.4194	-	0.4262	ES
EXIST2024-test_gold	0	9.1255	-	1.0000	-	-	-	EN
2	12	-4.3081	-	0.2640	-	-	-	EN
EXIST2024-test_gold	0	-	2.0402	-	1.0000	-	1.0000	EN
2	16	-	-0.1145	-	0.4719	-	0.4837	EN
1	19	-	-0.3659	-	0.4103	-	0.3645	EN

Table 7
Task 4 Evaluation Results

# Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm	Cross Entropy	F1-Score	Lang
EXIST2024-test_gold	0	3.1107	-	1.0000	-	0.5852	-	All
1	9	-0.4360	-	0.4299	-	0.9278	-	All
EXIST2024-test_gold	0	-	0.9832	-	1.0000	-	1.0000	All
1	11	-	0.0767	-	0.5390	-	0.7248	All
EXIST2024-test_gold	0	3.1360	-	1.0000	-	0.6160	-	ES
1	10	-0.5939	-	0.4053	-	0.9610	-	ES
EXIST2024-test_gold	0	-	0.9815	-	1.0000	-	1.0000	ES
1	13	-	-0.0086	-	0.4956	-	0.7137	ES
EXIST2024-test_gold	0	3.0794	-	1.0000	-	0.5528	-	EN
1	9	-0.2959	-	0.4520	-	0.8929	-	EN
EXIST2024-test_gold	0	-	0.9848	-	1.0000	-	1.0000	EN
1	12	-	0.1612	-	0.5818	-	0.7379	EN

Table 8
Task 5 Evaluation Results

# Run	Rank	ICM-Soft	ICM-Hard	ICM-Soft Norm	ICM-Hard Norm	Cross Entropy	F1-Score	Lang
EXIST2024-test_gold	0	4.7018	-	1.0000	-	0.9325	-	All
1	4	-1.5329	-	0.3370	-	1.4664	-	All
EXIST2024-test_gold	0	-	1.4383	-	1.0000	-	1.0000	All
1	6	-	-0.2881	-	0.3999	-	0.3837	All
EXIST2024-test_gold	0	4.8140	-	1.0000	-	0.9365	-	ES
1	4	-1.7405	-	0.3192	-	1.4800	-	ES
EXIST2024-test_gold	0	-	1.4356	-	1.0000	-	1.0000	ES
1	4	-	-0.2668	-	0.4071	-	0.3771	ES
EXIST2024-test_gold	0	4.5834	-	1.0000	-	0.9282	-	EN
1	4	-1.3812	-	0.3493	-	1.4521	-	EN
EXIST2024-test_gold	0	-	1.4409	-	1.0000	-	1.0000	EN
1	7	-	-0.3123	-	0.3916	-	0.3860	EN