TrankilTwice at EXIST2025: Detecting Sexism in Memes under Multi-Lingual Settings

Notebook for the EXIST Lab at CLEF 2025

Paolo Italiani[1,3,*,†], Fariha Maqbool[2,3,†], David Gimeno-Gómez[3,†], Elisabetta Fersini[2] and Carlos-D. Martínez-Hinarejos[3]

[1]Dep. of Computer Science and Engineering, University of Bologna, Bologna, Italy
[2]DISCo, Università degli Studi di Milano-Bicocca, Milan, Italy
[3]PRHLT research center, Universtitat Politècnica de València, Valencia, Spain

Abstract

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

The rapid spread of information through social media has amplified new forms of expression, particularly through multimodal content such as memes. While these formats offer creative and engaging means of communication, they have also become vehicles for disseminating harmful messages. Misogynistic content often hides behind humor, irony, or ambiguity, often reflecting a wide spectrum of author intentions. This variability presents a key challenge for automated detection systems, especially under learning frameworks like Learning with Disagreements. This paper presents the TrankilTwice team's participation in the EXIST (sEXism Identification in Social neTworks) Lab at CLEF 2025. Focused on Task 2.1, which addresses sexism identification in Spanish and English memes, we proposed an end-to-end system integrating LLM-based prompting strategies, cross-modal language encoding, and graph-based modeling at meme level. Our approach reached the 1st and 5th places in the final ranking for soft- and hard-label evaluations, respectively. While these results demonstrate the effectiveness of our approach, further analysis revealed performance gaps across languages, pointing to the need for more robust multilingual handling.

Keywords

Sexism Detection, Memes, Multimodal Large Language Models, Graph Convolutional Networks.

1. Introduction

The widespread use of social media platforms has changed the way of communication making it easier for people to share their thoughts and opinions. Although these platforms support free speech, some people misuse this freedom to abuse others based on gender, religion, or race [1]. One of the most prevalent forms of online hate is sexism [2], a type of gender-based bias that promotes the belief that one gender is superior to another. Although it can affect individuals of any gender, women are most often its primary targets [3] which have been shown to harm women's self-esteem, cognitive performance, career ambitions, and encourage traditional gender roles and dependency behaviors [4].

These harmful attitudes are increasingly spreading online through various media, including text, images, and videos. Among these, memes have become a common medium for disseminating sexist content. They combine text and images that may seem harmless separately, but when paired together, can convey offensive messages. This subtle interplay between modalities makes automated sexism detection in memes especially challenging because it requires a thorough understanding of the context of both image and text.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
paolo.italiani@studio.unibo.it (P. Italiani); f.maqbool@campus.unimib.it (F. Maqbool); dagigo1@dsic.upv.es (D. Gimeno-Gómez); elisabetta.fersini@unimib.it (. Fersini); cmartine@dsic.upv.es (Carlos-D. Martínez-Hinarejos)
0000-0002-9710-3748 (P. Italiani); 0009-0008-2587-9417 (F. Maqbool); 0000-0002-7375-9515 (D. Gimeno-Gómez); 0000-0002-8987-100X (. Fersini); 0000-0002-6139-2891 (Carlos-D. Martínez-Hinarejos)

To build safer and more inclusive online spaces, researchers are actively working on automated methods for detecting harmful content in memes. Early studies primarily relied on unimodal approaches using traditional machine learning techniques with hand-crafted features, such as TF-IDF or Bag-of-words for text [5], which lacked the capacity to capture deeper contextual cues. The adoption of deep learning methods [6, 7, 8], and later transformer-based models [9, 10], marked a substantial improvement in feature representation and classification accuracy. However, unimodal models often find it difficult to fully interpret the semantic alignment between modalities. To address this, recent research has been shifted towards multimodal approaches, ranging from basic fusion methods [11] to sophisticated vision-language models such as VisualBERT [12], CLIP [13] , and BLIP [14]. These models learn joint representations of text and image, enabling a more nuanced understanding of multimodal content. Several studies have successfully applied these models for hate and offensive content detection in memes and related tasks [15]. Furthermore, attention-based and transformer-driven architectures have enhanced the ability to model complex interactions between modalities [16, 17]. Recently, Graph Neural Networks (GNNs) have also been explored to capture relational structures within and across modalities, offering contextual reasoning for improved detection performance [18, 19].

This challenge of sexism detection extends beyond the English language, as millions of users worldwide employ social media as a vehicle for spreading hate. However, since most published methodologies and resources for detecting offensive language and hate speech were designed for English [20], researchers tried to generate resources for cross-lingual and cross-cultural perspectives. In this sense, shared tasks such as DA-VINCIS [21], HOMO-MEX [22], DIMEMEX [23] have contributed in the advancement of sexism detection in the Spanish language. These initiatives highlight the ongoing efforts being made to identify and moderate harmful content across various languages and content types.

The subjective nature of hate speech, however, makes this problem even more challenging, as perceptions vary based on culture, experience, and context [24]. What one person perceives as offensive may seem harmless to another which leads to disagreements among annotators. Despite these differences in opinion, the annotations are usually merged into a single "gold standard" dataset, which is then used to train and evaluate machine learning models. However, relying on a single "correct" answer overlooks the inherent subjectivity and complexity of many tasks. This approach tends to favor simplified, low-risk evaluations, which can ultimately hinder progress in the field [24]. To address this, efforts like the Learning With Disagreement (LeWiDi) paradigm [25] promote frameworks that preserve annotator disagreement, aiming to take into consideration different viewpoints as opposed to depending solely on a strict definition of hate.

The sEXism Identification in Social neTworks (EXIST 2025) shared task [26, 27] incorporates this learning with disagreement paradigm, which leverages these multiple perspectives, improving the ability of models to generalize across different interpretations of harmful content. Furthermore, by including both Spanish and English languages, the task also addresses the challenges of multilinguality and cultural differences. Therefore, this paper presents our participation in Task 2.1 of EXIST 2025 challenge and is organized as follows: Section 2 describes our methodology, including description of the task, dataset, and the architecture of our model. Section 3 details the experimental setup, and Section 4 presents and discusses our results. Finally, Section 5 concludes the paper.

2. Method

This section covers several key aspects, including the contextualization of the EXIST2025 challenge, an overview of the dataset used in this work focused on sexism detection in memes, and a description of our proposed model architecture.

2.1. Task Description

The *sEXism Identification in Social neTworks* task at CLEF 2025 [26, 27] aims at addressing the problem of sexism identification at a broad spectrum of sexism manifestation in social networks. Overall, the competition introduces three subtasks, each targeting a different level of granularity in sexism

identification: (i) detecting whether a message is sexist, (ii) inferring the author's intent, and (iii) categorizing the specific type of sexism expressed. For all cases, the challenge also embraces the LeWiDi paradigm to better address the subjective nature of this task. Another relevant aspect of the 2025 edition is its expanded scope. Unlike previous editions of the challenge [28, 29], which focused solely on detecting and classifying sexist content in text and memes, the current edition introduces new tasks that address the same issues through a different form of representation: videos. It is important to note, however, that **this paper focuses exclusively on Task 2.1**, a binary classification task aimed at determining whether a given meme is sexist or not.

2.2. Dataset Overview

The meme dataset proposed for EXIST 2025 Task 2.1 includes over 3,000 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. Figure 1 exemplifies the types of memes we can find.

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

Figure 1: Examples of memes in the EXIST 2025 meme dataset, showcasing the inherent challenges of the task. En and Sp refer to English and Spanish, respectively.

In addition to the conventional binary, discrete classification framework, the challenge also adopts the LeWiDi paradigm to approach the task of sexism identification. Each meme is annotated by six individuals, with labels and corresponding personal metadata (e.g., age, gender, etc.) collected for each annotator. This setup enables two distinct 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 considering the probability distribution for each class derived from all annotators' labels.

Table 1
Data distribution across splits for EXIST 2025 dataset

	Language	Class Label		Total
		Sexist	Non-Sexist	
Training	*Spanish*	1,111	536	1,647
	English	1,013	575	1,588
	Total	**2,124**	**1,111**	**3,235**
Validation	*Spanish*	144	49	193
	English	128	83	211
	Total	**272**	**132**	**404**
Test	*Spanish*	140	54	194
	English	126	85	211
	Total	**266**	**139**	**405**

Unlike tasks based on Tweets, automatic sexism detection in memes did not come with a predefined validation dataset. Therefore, to conduct our hyperparameter optimization experiments, **we partitioned the official training split to create both our internal validation and test sets**, providing a basis for our initial experimental 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 10% of the samples for our validation and test set, using the remaining data for training our models. Table 1 reports the data stastics of each our internal data sets in terms of language and class.

Overall, we observe that the distribution across class labels in all data sets is notably imbalanced, with a higher prevalence of sexist memes —an aspect that might impact the optimal estimation of our automatic models. Nonetheless, it worth noting that the training set remains balanced in terms of language representation. Further analyses in this regard can be found in Section 4.

2.3. Model Architecture

This section describes the modules that compose our proposed multimodal approach for the automatic detection of sexism in memes presented in this paper, whose overall architecture is depicted in Figure 2.

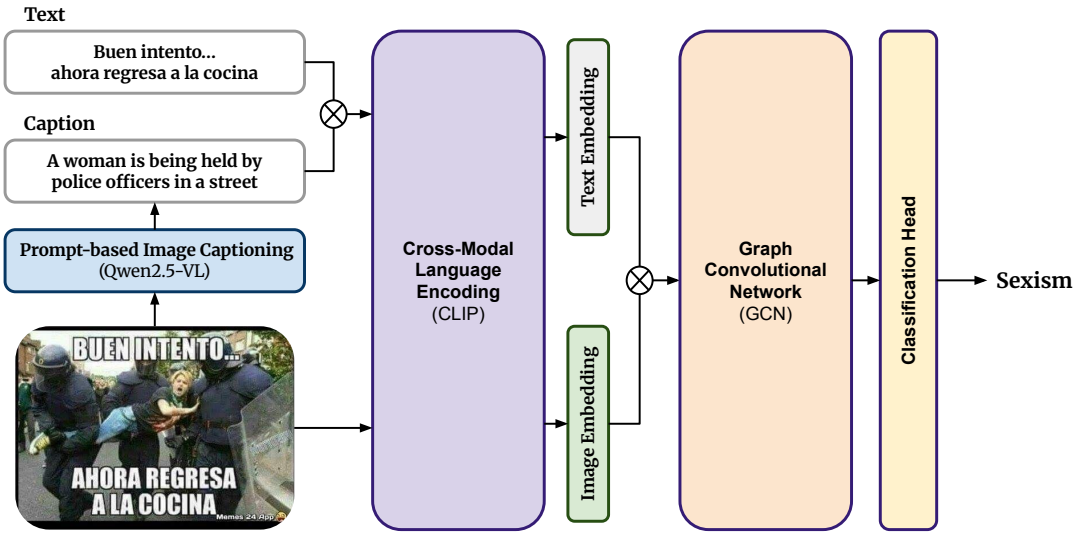

Figure 2: Overall schema of our proposed approach for sexism detection in memes.

Prompt-based Image Captioning. As a first step, we employed a multimodal Large Language Model (LLM), specifically Qwen2.5-V [30], to generate descriptive captions for each meme image. To do so, we input the image into the model using the following prompt: ``Describe this image without including what text reads and credit sources". Similar strategies have been adopted in recent state-of-the-art studies [18]. Note that, regardless of the meme's original language, all generated image captions were produced in English.

Cross-Modal Language Encoding. In the next step, we extracted feature embeddings for both the text and image modalities using CLIP [31], a cross-modal language encoder that has been extensively explored in the context of sexism detection [32, 33, 10]. For the visual modality, embeddings were obtained by directly inputting the raw meme image. For the textual modality, we constructed a composite input by concatenating the prompt-based image caption (generated by the LLM) with the textual transcription of the meme itself, separated by a special delimiter token. This process yielded two distinct embeddings: one representing the visual content and the other the textual information.

Graph Convolutional Network (GCN). Inspired by prior work demonstrating the effectiveness of graph-based frameworks for the automatic identification of hate speech [18], we incorporated a GCN module into our architecture. After concatenating the two embeddings produced by CLIP, the GCN is employed to further enrich this multimodal representation by modeling the relationships and potential

contextual cues among memes.

Classification Head. Finally, the latent and enriched representation produced by the GCN module is passed through a classification head consisting of a linear projection layer followed by a softmax activation function.

3. Experimental Setup

In this section, we first provide details on our task domain adaptation using a larger meme-based hate speech dataset, followed by preliminary experiments on hyperparameter tuning. We then present the three approaches submitted to the challenge, describe the official baselines used for comparison, and review the evaluation metrics employed for model performance assessment. Finally, we summarize key aspects of our run submissions and implementation setup.

3.1. Task Domain Pre-Training

As we described in Subsection 2.3, memes are represented using latent embeddings extracted from a pre-trained CLIP model. While it is true that CLIP was pre-trained on large-scale data, its training was oriented toward general-purpose tasks. Therefore, we explored the inclusion of additional datasets that, although broader in scope —such as general hate speech rather than misogyny specifically— could potentially improve performance. In preliminary experiments, we found the inclusion of the MAMI dataset [20] particularly beneficial. Derived from SemEval-2022 Task 5, this dataset focuses on detecting misogynistic content across five categories: not misogynistic, shaming, stereotype, objectification, and violence. It contains approximately 11,000 annotated samples.

3.2. Hyperparameter Optimization

To identify the optimal model configuration, we conducted a series of preliminary hyperparameter tuning experiments using the validation split described in Subsection 2.2. Notably, these experiments were conducted exclusively under the hard-label setup. Once the best-performing settings were determined for each of our three proposed approaches, we applied the same configurations to train the corresponding soft-label models without additional tuning.

Specifically, we explored a range of values for key hyperparameters, including the learning rate, the number of layers in the GCN, and the batch size. Given the particular sensitivity of GCNs to batch size, we carried out this exploration both during the pre-training phase with the MAMI dataset and the subsequent fine-tuning on the EXIST dataset.

3.3. Proposed Approaches

We selected the three best-performing model configurations for our challenge run submission. Based on the specific values of their most relevant hyperparameters, we can distinguish the following models:

- **TrankilTwice_1**. This model uses a learning rate of 5×10^{-6} and a batch size of 64 both for the MAMI pretraining and EXIST fine-tuning. It incorporates a single GCN layer, as deeper architectures showed no benefit, likely due to the strong representational power of the initial embeddings.
- **TrankilTwice_2**. Similar to the previous setup, this configuration also uses a learning rate of 5×10^{-6} and a batch size of 64, but decreases the MAMI pretraining batch size to 20. It likewise includes one single GCN layer.
- **TrankilTwice_3**. Although it also uses a learning rate of 5×10^{-6}, this approach employs a batch size of 100 for the EXIST training, and 64 for the MAMI pretraining. Consistent with the other models, its best performance was achieved with a single GCN layer.

3.4. Baselines

We compared our proposed approaches against the official baselines provided by the challenge, which include the following:

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

3.5. Run Submission

It is also important to note that, for run submission, we chose not to retrain the models on the full dataset. This decision was motivated by the concern that, since our models were already pretrained and further adapted on the MAMI dataset, training on additional data could increase the risk of overfitting. Therefore, we simply performed inference on the official test set provided by the challenge.

3.6. Evaluation Metrics

For the evaluation of our model approaches, we followed the official guidelines of the challenge. Depending on the label setup, the evaluation metric differs: the F_1-score is used for hard labels, while cross-entropy is applied for soft labels. Nonetheless, the core evaluation metric for both scenarios —albeit with slight adaptations— is the so-called Information Contrast Measure (ICM) [34], along with its normalized variant, ICM-Norm.

3.7. Implementation Details

All runs were tracked with Weights & Biases and executed on a single Nvidia GeForce RTX3090 GPU of 24 GB of dedicated memory, 64 GB of VRAM, and an Intel® Core™ i9-10900X1080 CPU @ 3.70GHz. The operating system is Ubuntu 20.04.3 LTS. To enhance consistency and portability, our development environment is built on top of a docker container with a NVIDIA image.

All components of our proposed architecture, except for the LLM, were locally trained and, when necessary, quantized using standardized procedures.

4. Results & Discussion

Our approach reached the 1st and 5th places in the final challenge ranking for soft- and hard-label evaluations, respectively. Given the lack of ground truth for official test set, this section focuses, however, on the results obtained through our experiments with our internal data split. While Subsection 4.1 discusses the overall performance results of our proposed appraoch, Subsection 4.2 addresses the analysis of these performances depending on the meme's language.

4.1. Overall Performance Analysis

By examining the results reported in Table 2, we observe that our proposed approach substantially outperforms the official majority- and minority-based baselines provided by the challenge. A comparison across our three selected models highlights the importance of optimal hyperparameter selection, as performance varies notably depending on configuration. Interestingly, we can observe how the worst model configuration for the hard-label setup (`TrankilTwice_3`) is, however, the best-performing one for the soft-label. This finding highlights the relevance of the LeWiDi paradigm [35] by demonstrating how modeling inter-annotator subjectivity can lead to different outcomes.

Table 2
Overall analysis of our proposed approaches for EXIST2025 Subtask 2.1 for both English and Spanish instances. Results reported for our internal test set, using variants of ICM (↑), macro 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
$EXIST_{2025}$ (Gold)	+0.9279	1.0000	1.0000	+3.1597	1.0000	0.6373
$EXIST_{2025}$ (Majority)	−0.3393	0.3171	0.3964	−2.7354	0.0671	4.4227
$EXIST_{2025}$ (Minority)	−0.8821	0.0246	0.2555	−3.9249	0.0000	5.5460
TrankilTwice_1	**+0.1292**	**0.5696**	**0.7236**	−0.2636	0.4583	1.0894
TrankilTwice_2	+0.0971	0.5523	0.7094	−0.2410	0.4619	1.3574
TrankilTwice_3	+0.0562	0.5303	0.6942	**−0.2198**	**0.4652**	**1.0394**

4.2. Language-Specific Performance Analysis

When analyzing our model performances for each language separately, as reported in Tables 3 and 4, we can infer two key findings:

Table 3
Overall analysis of our proposed approaches for EXIST2025 Subtask 2.1 for English-only instances. Results reported for our internal test set, using variants of ICM (↑), macro 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
$EXIST_{2025}$ (Gold)	+0.9726	1.0000	1.0000	+3.1633	1.0000	0.6503
$EXIST_{2025}$ (Majority)	−0.3839	0.3027	0.3739	−3.1768	0.0000	4.8345
$EXIST_{2025}$ (Minority)	−0.6991	0.1406	0.2872	−3.4555	0.0000	5.1341
TrankilTwice_1	**+0.2231**	**0.6147**	**0.7458**	−0.1692	0.4733	1.0911
TrankilTwice_2	+0.1858	0.5955	0.7311	−0.1200	0.4810	1.3938
TrankilTwice_3	+0.1831	0.5941	0.7294	**+0.1126**	**0.5178**	**1.0034**

Table 4
Overall analysis of our proposed approaches for EXIST2025 Subtask 2.1 for Spanish-only instances. Results reported for our internal test set, using variants of ICM (↑), macro 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
$EXIST_{2025}$ (Gold)	+0.8532	1.0000	1.0000	+3.1989	1.0000	0.6232
$EXIST_{2025}$ (Majority)	−0.3049	0.3213	0.4192	−2.4436	0.1181	3.9748
$EXIST_{2025}$ (Minority)	−1.1575	0.0000	0.2177	−4.7657	0.0000	5.9939
TrankilTwice_1	**−0.0133**	**0.4922**	**0.6885**	**−0.4640**	**0.4275**	1.0876
TrankilTwice_2	−0.0417	0.4756	0.6723	−0.4750	0.4258	1.3177
TrankilTwice_3	−0.1276	0.4252	0.6406	−0.6782	0.3940	**1.0786**

- First, we observe that performance is consistently higher on English memes compared to their Spanish counterparts. Although LLM-based prompt responses were generated in English for both cases, this performance gap may still be attributed to the overall English-centric bias inherent in

most of the pre-trained models we employed, as well as potential differences in linguistic nuance or meme structure between the two languages.
- The second finding concerns the label setup. For English, we observe the same pattern noted in our overall performance analysis: the model that performed worst under the hard-label setup (`TrankilTwice_3`) emerges as the best under the soft-label setup. However, in the case of Spanish memes, the best-performing model remains consistent across both scenarios, with `TrankilTwice_1` outperforming or on pair with the others configurations.

To identify potential biases in the data, we examined the distributions within our internal split. As shown in Table 1, despite the notable class imbalance, there is no clear differentiation across languages. This observation supports our hypothesis that performance discrepancies between languages may stem from linguistic nuances or subtle meme structures that are more easily captured by the predominantly English-centric pre-trained models used in our approach.

5. Conclusions

In this paper, we presented our contribution to the EXIST 2025 challenge, specifically addressing Task 2.1, which focuses on the automatic classification of sexism in memes for both Spanish and English. We described our proposed system, which integrates LLM-based prompting strategies, cross-modal language encoding, and graph-based modeling at the meme level. Our experimental findings, particularly the performance discrepancies across languages, underscore the need for further research into multilingual approaches for automatic sexism detection.

Acknowledgments

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 PCI2022-135008-2 under project MARTINI funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGenerationEU/PRTRPI. The work of P. Italiani was partially supported by (i) Artificial Intelligence for Public Administration Connected (AI-PACT), CUP B47H22004450008 and B47H22004460001, PNRR, mission 4, component 2, investment 2.3, (ii) DigitAl lifelong pRevEntion (DARE), CUP B53C22006450001, code PNC0000002, Complementary National Plan PNC-I.1, (iii) Future Artificial Intelligence Research (FAIR), Spoke 8, PNRR—M4C2—Investment 1.3, Extended Partnership PE00000013, (iv) Chips JU TRISTAN project, G.A. 101095947, European Commission and Italian MIMIT. We thank Maggioli Group for co-funding the Ph.D. scholarship to P. Italiani. The work of F. Maqbool and E. Fersini was supported by PNRR ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing (CN00000013), under the NRRP MUR program funded by the NextGenerationEU. Finally, note that the work of P. Italiani and F. Maqbool was carried out during their internship at the PRHLT research center.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4.5 in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. I. Montero, N. Laforgue-Bullido, D. Abril-Hervás, Hate speech: A Systematic Review of Scientific Production and Educational Considerations, Revista Fuentes 24 (2022) 222–233.

[2] J. Fox, C. Cruz, J. Y. Lee, Perpetuating Online Sexism Offline: Anonymity, Interactivity, and the Effects of Sexist Hashtags on Social Media, Computers in Human Behavior 52 (2015) 436–442.

[3] M. Duggan, Online Harassment, Pew Research Center (2017).
[4] A. I. Kural, M. Kovács, Attachment security schemas to attenuate the appeal of benevolent sexism: The effect of the need to belong and relationship security, Acta Psychologica 229 (2022) 103671.
[5] S. Akuma, T. Lubem, I. T. Adom, Comparing Bag of Words and TF-IDF with Different Models for Hate Speech Detection From Live Tweets, Int. J. Inf. Technol. 14 (2022) 3629–3635.
[6] Z. Zhang, D. Robinson, J. Tepper, Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web, Springer, Cham, 2018, pp. 745–760.
[7] T. Deshpande, N. Mani, An Interpretable Approach to Hateful Meme Detection, in: Proceedings of the International Conference on Multimodal Interaction, ICMI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 723–727. doi:10.1145/3462244.3479949.
[8] M. D. Belete, G. K. Alitasb, Identification of Hateful Amharic Language Memes on Facebook using Deep Learning Algorithms, Systems and Soft Computing 7 (2025) 200258. doi:10.1016/j.sasc.2025.200258.
[9] S. Chinivar, M. S. Roopa, J. S. Arunalatha, K. R. Venugopal, Identification of Misogynistic Memes Using Transformer Models, in: S.-L. Peng, A. Mondal, V. R. Kagita, J. L. Sarkar (Eds.), Proceedings of International Conference on Advanced Communications and Machine Intelligence, Springer Nature Singapore, Singapore, 2024, pp. 107–116.
[10] G. Rizzi, D. Gimeno-Gómez, E. Fersini, C.-D. Martínez-Hinarejos, PINK at EXIST2024: a cross-lingual and multi-modal transformer approach for sexism detection in memes, Working Notes of CLEF (2024).
[11] E. Fersini, F. Gasparini, S. Corchs, Detecting sexist meme on the web: A study on textual and visual cues, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 226–231. doi:10.1109/ACIIW.2019.8925199.
[12] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, What Does BERT with Vision Look At?, in: Proceedings of the 58th annual meeting of the Association for Computational Linguistics (ACL), 2020, pp. 5265–5275.
[13] 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.
[14] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-Training for Unified Vision-Language Understanding and Generation, in: Proceedings of the International Conference on Machine Learning, 2022, pp. 12888–12900.
[15] P. Kapil, A. Ekbal, A transformer based multi task learning approach to multimodal hate speech detection, Natural Language Processing Journal 11 (2025) 100133. doi:10.1016/j.nlp.2025.100133.
[16] E. F. Ayetiran, Özlem Özgöbek, An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection, Information Systems 123 (2024) 102378. doi:10.1016/j.is.2024.102378.
[17] A. Indira Kumar, G. Sthanusubramoniani, D. Gupta, A. R. Nair, Y. A. Alotaibi, M. Zakariah, Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches, Egyptian Informatics Journal 30 (2025) 100683. doi:10.1016/j.eij.2025.100683.
[18] B. Xu, E. Yu, J. Zhou, H. Lin, L. Zong, HyperHatePrompt: A Hypergraph-based Prompting Fusion Model for Multimodal Hate Detection, in: Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, 2025, pp. 3825–3835.
[19] Y. Mu, J. Yang, T. Li, S. Li, W. Liang, HA-GCEN: Hyperedge-abundant Graph Convolutional Enhanced Network for Hate speech Detection, Knowledge-Based Systems 300 (2024) 112166. doi:10.1016/j.knosys.2024.112166.
[20] 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.

[21] H. J. Jarquín-Vásquez, D. I. Hernández-Farías, L. J. Arellano, H. J. Escalante, L. V. Pineda, M. Montes-y-Gómez, F. Sánchez-Vega, Overview of DA-VINCIS at IberLEF 2023: Detection of Aggressive and Violent Incidents from Social Media in Spanish, Proces. del Leng. Natural 71 (2023) 351–360.

[22] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of HOMO-MEX at Iberlef 2023: Hate speech detection in Online Messages directed Towards the MEXican Spanish speaking LGBTQ+ population, Proces. del Leng. Natural 71 (2023) 361–370.

[23] H. J. Jarquín-Vásquez, I. Tlelo-Coyotecatl, M. Casavantes, D. I. Hernández-Farías, H. J. Escalante, L. V. Pineda, M. Montes-y-Gómez, Overview of DIMEMEX at IberLEF 2024: Detection of Inappropriate Memes from Mexico, Proc. del Leng. Natural 73 (2024) 335–345.

[24] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We Need to Consider Disagreement in Evaluation, in: K. Church, M. Liberman, V. Kordoni (Eds.), Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, Online, 2021, pp. 15–21. doi:10.18653/v1/2021.bppf-1.3.

[25] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 Task 12: Learning with Disagreements, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021, Association for Computational Linguistics, 2021, pp. 338–347. doi:10.18653/V1/2021.SEMEVAL-1.41.

[26] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2025.

[27] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025.

[28] 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.

[29] 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.

[30] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-VL Technical Report, arXiv preprint arXiv:2502.13923 (2025).

[31] 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 ICML), volume 139 of *Proc. of Machine Learning Research*, PMLR,

2021, pp. 8748–8763.
[32] 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.
[33] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing Misogynous Memes: Biased Models and Tricky Archetypes, Information Processing & Management 60 (2023) 103474.
[34] E. Amigo, A. Delgado, Evaluating Extreme Hierarchical Multi-label Classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proc. of the 60th ACL, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. doi:10.18653/v1/2022.acl-long.399.
[35] 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 SemEval-2023, 2023, pp. 2304–2318.