Team Marsad at EXIST 2026: End-to-End Vision-Language Fine-Tuning for Sexism Identification in Multilingual Memes and TikTok Videos

Notebook for the EXIST Lab at CLEF 2026

Kholoud Khalil Aldous[1,*], Wajdi Zaghouani[1]

[1]*Northwestern University, Doha, Qatar*

Abstract

This paper describes the participation of the Marsad team in EXIST 2026 (Tasks 2.1 and 3.1), the sixth edition of the sEXism Identification in Social neTworks shared task at CLEF 2026. Both tasks require binary identification of sexist content under the Learning With Disagreement (LeWiDi) paradigm, which demands that systems predict soft annotator-vote distributions rather than a single hard label. For Task 2.1 (sexism identification in memes, 3,984 training instances), we evaluate a suite of models via stratified 5-fold cross-validation on the augmented training set, including TF-IDF with Logistic Regression, CLIP image-text fusion with Logistic Regression and SVM, and end-to-end fine-tuned CLIP ViT-L/14, submitting the top-three models by CV F1. For Task 3.1 (sexism identification in TikTok videos, 2,524 training instances), we extract 5 frames per video using OpenCV and encode them with CLIP ViT-B/32 and ViT-L/14 using average pooling, evaluating video-text fusion models with SVM and Logistic Regression alongside a fine-tuned CLIP ViT-L/14 and XLM-RoBERTa-large; the top-3 CV models are combined in a weighted soft-label ensemble as the primary run. Our best Task 2.1 run (Marsad_1, fine-tuned CLIP ViT-L/14) achieved ICM-Soft $= -0.1156$ (rank 17/141 overall), with notably stronger English performance (ICM-Soft $= 0.1440$, rank 11/141). Our best Task 3.1 run (Marsad_2, fine-tuned CLIP ViT-L/14) achieved ICM-Soft $= -0.3836$ (rank 35/60) and ICM-Hard $= 0.0774$ (rank 54/75). We analyze the gap between cross-validation and test-set performance, discuss the challenges of soft label prediction and multilingual generalization, and outline directions for future work.

Keywords

sexism detection, learning with disagreement, CLIP, multimodal fusion, meme classification, video understanding, soft labels, cross-validation

1. Introduction

Sexism is a widespread form of online discrimination that has increasingly migrated from text to multimodal formats, including internet memes and short-form social media videos. The automatic detection of such content is a critical challenge for content moderation systems and requires models that can jointly reason over visual and linguistic features. The EXIST shared task series [1, 2] has progressively advanced the state of the art in computational sexism detection since its first edition in 2021.

EXIST 2026, the sixth edition, introduces two key innovations. First, it adopts the Learning With Disagreement (LeWiDi) paradigm [3], which requires systems to predict the full distribution of annotator judgements rather than a single majority-vote label, evaluated using the Information Contrast Measure (ICM) [4]. Second, for the first time, the dataset includes physiological sensor signals — EEG, heart rate, and eye-tracking, collected during annotation, capturing unconscious affective responses to sexist content [1].

The Marsad team participated in Task 2.1 (sexism identification in memes) and Task 3.1 (sexism identification in TikTok videos). Our approach evaluates a suite of multimodal models under stratified 5-fold cross-validation, ranging from TF-IDF baselines to end-to-end fine-tuned CLIP ViT-L/14 and

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

kholoud.aldous@northwestern.edu (K. K. Aldous); wajdi.zaghouani@northwestern.edu (W. Zaghouani)

0000-0003-1188-5724 (K. K. Aldous); 0000-0003-1521-5568 (W. Zaghouani)

CLIP feature fusion with SVM. For Task 2.1, we submit three individually-trained models selected by CV F1; for Task 3.1, our primary run is a weighted soft-label ensemble of the top-3 CV models. Our best Task 2.1 run achieved rank 17/141 on ICM-Soft overall (rank 11/141 on English); our best Task 3.1 run achieved rank 35/60 on ICM-Soft.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the tasks and datasets. Section 4 presents our system. Section 5 describes the experimental setup. Section 6 presents official results. Section 7 discusses key findings, and Section 8 concludes.

2. Related Work

2.1. Sexism Detection

Automatic sexism detection has been studied extensively on text-based social media platforms [5]. Early approaches relied on lexica and surface features; contemporary systems leverage pre-trained transformers such as BERT [6] and multilingual models including XLM-RoBERTa [7]. The EXIST task series has systematically expanded the scope from binary text classification to multi-label categorization and, from 2025, to multimodal memes and TikTok video content [8, 2], establishing a progressively richer benchmark for computational sexism detection.

2.2. Multimodal Content Moderation

The detection of harmful content in memes has received sustained attention since the Hateful Memes Challenge [9], which demonstrated that unimodal models are insufficient and that joint image-text understanding is essential. CLIP [10] has emerged as a dominant backbone for multimodal classification tasks due to its large-scale contrastive pre-training on 400 million image-text pairs, producing semantically aligned visual and textual representations. CLIP-based approaches have been successfully applied to hate speech and offensive content detection in both static images and short video clips, with frame-level encoding followed by temporal pooling being a widely adopted strategy for video understanding. However, CLIP's textual encoder carries a well-documented English bias [11, 12] that can limit performance on non-English content, a limitation particularly relevant for multilingual tasks such as EXIST 2026.

2.3. Learning With Disagreement (LeWiDi)

The LeWiDi paradigm [3] challenges the assumption that a single gold label can adequately represent subjective annotation tasks, arguing that disagreement between annotators carries meaningful signal that should be preserved. EXIST 2026 adopts this paradigm, evaluating systems using the Information Contrast Measure (ICM-soft) [4], which generalizes Pointwise Mutual Information and measures the similarity between a system's predicted probability distribution and the observed annotator vote distribution.

3. Task and Dataset Description

3.1. Task 2.1: Sexism Identification in Memes

Task 2.1 is a binary classification task over internet memes. A meme is labelled sexist (YES) if it expresses or depicts a sexist situation, regardless of whether the author endorses it — content that criticises or denounces sexism is still annotated YES. Each instance carries a vector of individual annotator votes, from which a soft probability distribution $[p_{\text{NO}}, p_{\text{YES}}]$ is derived. The training set contains 3,984 memes in both English (en, $n = 2005$) and Spanish (es, $n = 1979$). The test set contains 1,053 instances. The hard-label distribution shows 2,003 YES and 1,367 NO instances, indicating a moderate class imbalance.

3.2. Task 3.1: Sexism Identification in Videos

Task 3.1 applies the same binary identification objective to TikTok short-form videos. The training set contains 2,524 videos, with a near-balanced hard-label split (1,306 NO, 1,202 YES) but a stronger language imbalance (es $n = 1524$, en $n = 1000$). Videos include spoken audio, text overlays, music, and visual content. The test set contains 674 instances.

4. System Description

Our approach consists of three stages: data preparation and augmentation, a comprehensive cross-validation model sweep, and submission generation from the best models.

4.1. Data Augmentation

Both tasks exhibit class imbalance in the hard-label distribution. For Task 2.1, the minority class (NO, $n = 1367$) is oversampled to approximately twice its original size (NO: $1367 \rightarrow 2734$), producing an augmented training set of 4,737 instances. For Task 3.1, the minority class is YES ($n = 1202$), oversampled to ≈ 2404, producing 3,710 augmented instances. Augmentation is performed via random word deletion applied to the text component of each instance.

4.2. Feature Extraction

Task 2.1 — Memes. We pre-compute two sets of CLIP embeddings for all 3,984 training instances and cache them for use across all classifier experiments: (1) **CLIP ViT-B/32** (base), producing 512-dimensional image and text embeddings; (2) **CLIP ViT-L/14** (large), producing 768-dimensional embeddings. For the fusion models, image and text embeddings are concatenated, yielding 1,024-dimensional (base) or 1,536-dimensional (large) feature vectors.

Task 3.1 — Videos. We extract $T = 5$ frames per video using OpenCV, sampling uniformly across the video duration. Frame extraction is applied to all 2,524 training videos and 674 test videos. Frame embeddings are averaged across the 5 frames to produce a single 512-dimensional (base) or 768-dimensional (large) video-level visual embedding. This is concatenated with the CLIP text embedding of the video title or caption to form the final feature vector.

4.3. Model Suite

We evaluate the following models in 5-fold cross-validation on the augmented datasets.

Task 2.1 Models.

- **Fine-tuned CLIP ViT-L/14 (e2e):** end-to-end fine-tuning of CLIP ViT-L/14 for 5 epochs with early stopping on validation loss (lr $= 10^{-5}$, batch size 64). Submitted as Marsad_1.
- **CLIP fusion large + LogReg:** concatenated image+text embeddings (1,536-dim) from CLIP ViT-L/14 with Logistic Regression. Submitted as Marsad_2.
- **CLIP fusion + SVM (tuned):** concatenated image+text embeddings (1,024-dim) from CLIP ViT-B/32 with an RBF-kernel SVM; C tuned via grid search (optimal $C = 1.0$). Submitted as Marsad_3.
- **Majority/Minority class baselines:** predict the majority (NO) or minority (YES) label for all instances; reported as lower-bound reference.

Table 1
5-fold CV results for submitted models — Task 2.1 Memes ($N = 4,737$). Ranked by CV F1 macro.

Model	CV F1	±std	CV Acc
Fine-tuned CLIP ViT-L/14 (e2e)	0.7915	0.0122	0.7950
CLIP fusion large + LogReg	0.7507	0.0087	0.7549
CLIP fusion + SVM (C=1.0)	0.7503	0.0081	0.7564
Majority class (NO)	0.3659	0.0002	0.5772
Minority class (YES)	0.2972	0.0002	0.4228

Table 2
5-fold CV results for submitted models — Task 3.1 Videos ($N = 3,710$). Ranked by CV F1 macro.

Model	CV F1	±std	CV Acc
Video fusion + SVM (C=10)	0.7820	0.0094	0.8116
Fine-tuned CLIP ViT-L/14 (e2e)	0.7755	0.0070	0.7960
TF-IDF + LogReg	0.7153	0.0063	0.7350
XLM-RoBERTa-large	0.5274	0.1645	0.6836
Majority class (YES)	0.3932	0.0002	0.6480
Minority class (NO)	0.2604	0.0003	0.3520

Task 3.1 Models.

- **Video fusion + SVM (tuned):** avg-pooled 5-frame CLIP ViT-B/32 embeddings concatenated with text embeddings, classified with an RBF-kernel SVM; grid search optimal $C = 10.0$. Component of Marsad_1 ensemble.
- **TF-IDF + LogReg:** TF-IDF features over bilingual (EN+ES) video captions with Logistic Regression. Component of Marsad_1 ensemble.
- **Fine-tuned CLIP ViT-L/14 (e2e):** end-to-end fine-tuning on frame-averaged video embeddings; 5 epochs. Component of Marsad_1 ensemble; submitted alone as Marsad_2.
- **XLM-RoBERTa-large:** text-only fine-tuned baseline on video captions. Submitted as Marsad_3.
- **Majority/Minority class baselines:** predict the majority (YES) or minority (NO) label for all instances; reported as lower-bound reference.

5. Experimental Setup

5.1. Implementation Details

All experiments were conducted on a CUDA-enabled GPU server using PyTorch 2.3.1 with batch size 64. Model selection uses stratified 5-fold cross-validation (`StratifiedKFold`, `random_state=42`) on the augmented training sets (Task 2.1: $N = 4{,}737$; Task 3.1: $N = 3{,}710$), ensuring class distribution is preserved across folds. CV scores are computed exclusively on held-out folds to prevent data leakage, and all models are refit on the full training set before test-set submission. Tables 1 and 2 report cross-validation results for the submitted models.

5.2. Run Configuration

Table 3 summarizes the three submitted runs for each task, mapping each Marsad run identifier to its model, CV F1, and submission status.

Task 2.1. Three successfully-trained models were selected for submission based on CV F1 ranking: fine-tuned CLIP ViT-L/14 as Marsad_1, CLIP fusion large + LogReg as Marsad_2, and CLIP fusion + SVM as Marsad_3.

Table 3
Submitted runs for Tasks 2.1 and 3.1. CV F1 macro is from 5-fold cross-validation on the augmented training set. All runs trained on bilingual EN+ES data.

Run	Model	CV F1	Notes
Task 2.1 — Memes (1,053 test instances)			
Marsad_1	Fine-tuned CLIP ViT-L/14 (e2e)	0.7915	Best single model by CV F1
Marsad_2	CLIP fusion large + LogReg	0.7507	CLIP ViT-L/14 img+txt embeddings (1,536-d)
Marsad_3	CLIP fusion + SVM (C=1.0, RBF)	0.7503	CLIP ViT-B/32 img+txt embeddings (1,024-d)
Task 3.1 — TikTok Videos (674 test instances)			
Marsad_1	Weighted ensemble (top-3 CV)	—[†]	SVM (w=0.349) + CLIP (w=0.334) + TF-IDF (w=0.317)
Marsad_2	Fine-tuned CLIP ViT-L/14 (e2e)	0.7755	Single best model; ensemble loaded successfully
Marsad_3	XLM-RoBERTa-large	0.5274	Text-only baseline on video captions

[†]No CV F1 available: the ensemble was constructed post-CV by combining the top-3 models weighted by their individual CV F1 scores. The weighted estimate is $0.349{\times}0.7820 + 0.334{\times}0.7755 + 0.317{\times}0.7153 \approx 0.759$.

Table 4
Task 2.1 official Soft-Soft evaluation. Rank out of 141 submitted runs.

Run	ALL		EN		ES	
	Rk	ICM-S	Rk	ICM-S	Rk	ICM-S
Marsad_1	**17**	$\mathbf{-0.1156}$	**11**	$\mathbf{0.1440}$	**30**	$\mathbf{-0.3995}$
Marsad_2	24	-0.2374	21	-0.0054	44	-0.4955
Marsad_3	50	-0.4212	57	-0.3281	48	-0.5292

Task 3.1. Marsad_1 is a weighted soft-label ensemble of the top-3 CV models — Video fusion SVM (F1 $= 0.782$), Fine-tuned CLIP ViT-L/14 (F1 $= 0.776$), and TF-IDF + LogReg (F1 $= 0.715$) — with weights proportional to CV F1 scores ($w_k = \text{F1}_k / \sum \text{F1}_j$), yielding normalised weights of 0.349, 0.334, and 0.317 respectively; the soft output is the weighted average of each model's `predict_proba` output. Marsad_2 submits the fine-tuned CLIP ViT-L/14 model alone to isolate its individual contribution relative to the ensemble. Marsad_3 provides a text-only upper bound using XLM-RoBERTa-large on video captions.

6. Results

6.1. Task 2.1: Sexism Identification in Memes

Tables 4 and 5 present the official results of Task 2.1. **Under the soft evaluation**, Marsad_1 (fine-tuned CLIP ViT-L/14) achieves the best performance, ranking 17th out of 141 systems overall (ICM-Soft $= -0.1156$). Marsad_1 performs better on English content (ICM-Soft $= 0.1440$, rank 11/141) than on Spanish content (ICM-Soft $= -0.3995$, rank 30/141), indicating a pronounced language gap. **Under the hard evaluation**, Marsad_3 (CLIP fusion SVM) achieves the best ICM-Hard overall, ranking 41st out of 214 systems (ICM-Hard $= 0.1218$). On English content, Marsad_1 leads with rank 37 (ICM-Hard $= 0.2139$, F1 $= 0.7223$), while Marsad_3 performs best on Spanish, ranking 53rd (ICM-Hard $= 0.0377$, F1 $= 0.7256$).

Table 5
Task 2.1 official Hard-Hard evaluation. Rank out of 214 submitted runs.

Run	ALL			EN			ES		
	Rk	ICM-H	F1	Rk	ICM-H	F1	Rk	ICM-H	F1
Marsad_1	47	0.1194	0.6876	**37**	**0.2139**	**0.7223**	60	0.0261	0.6522
Marsad_2	66	0.0611	0.7153	45	0.1750	0.7341	90	−0.0532	0.6988
Marsad_3	**41**	**0.1218**	0.7219	39	0.2043	0.7173	**53**	**0.0377**	**0.7256**

Table 6
Task 3.1 official Soft-Soft evaluation. Rank out of 60 submitted runs.

Run	ALL		EN		ES	
	Rk	ICM-S	Rk	ICM-S	Rk	ICM-S
Marsad_1 (ensemble)	50	−0.9274	48	−0.7095	53	−1.2977
Marsad_2 (CLIP e2e)	**35**	**−0.3836**	**35**	**−0.1248**	**35**	**−0.7843**
Marsad_3 (XLM-R-L)	55	−1.1539	55	−0.9584	56	−1.4934

Table 7
Task 3.1 official Hard-Hard evaluation. Rank out of 75 submitted runs.

Run	ALL			EN			ES		
	Rk	ICM-H	F1	Rk	ICM-H	F1	Rk	ICM-H	F1
Marsad_1 (ens.)	68	−0.0898	0.6227	66	0.0219	0.6720	68	−0.2461	0.5571
Marsad_2 (CLIP)	**54**	**0.0774**	**0.6656**	**43**	**0.1589**	**0.7025**	**55**	**−0.0391**	**0.6154**
Marsad_3 (XLM-R)	69	−0.0958	0.6007	65	0.0422	0.6355	69	−0.2602	0.5614

6.2. Task 3.1: Sexism Identification in Videos

Tables 6 and 7 present the official results of Task 3.1 . **Under the soft evaluation**, Marsad_2 (fine-tuned CLIP ViT-L/14) achieves the best performance, ranking 35th out of 60 systems overall (ICM-Soft $= -0.3836$). Marsad_2 performs better on English content (ICM-Soft $= -0.1248$, rank 35/60) than on Spanish content (ICM-Soft $= -0.7843$, rank 35/60), indicating a pronounced language gap consistent with Task 2.1. Under the hard evaluation, Marsad_2 again leads, ranking 54th out of 75 systems overall (ICM-Hard $= 0.0774$). On English content, Marsad_2 ranks 43rd (ICM-Hard $= 0.1589$, F1 $= 0.7025$), while on Spanish it ranks 55th (ICM-Hard $= -0.0391$, F1 $= 0.6154$).

7. Discussion

7.1. English vs. Spanish Performance Gap

All runs show a consistent and substantial decline from English to Spanish on both tasks. For Task 2.1, Marsad_1 achieves ICM-Soft $= 0.1440$ on English but -0.3995 on Spanish, a gap of approximately 0.54 ICM points. The same pattern holds under the hard evaluation, where Marsad_1 ranks 37th on English (ICM-Hard $= 0.2139$) but drops to 60th on Spanish (ICM-Hard $= 0.0261$). For Task 3.1, Marsad_2 achieves ICM-Soft $= -0.1248$ on English versus -0.7843 on Spanish, a gap of 0.66 ICM points, and ICM-Hard $= 0.1589$ on English versus -0.0391 on Spanish. This is consistent with a known limitation of CLIP, which was trained predominantly on English image-text pairs [10], leading to a strong English bias in its textual encoder that persists even when the model is applied to multilingual content [11, 12]. Additionally, TikTok content in Spanish is disproportionately representative of Latin American and Iberian cultural contexts where sexist expression may take culturally specific forms not well-captured by CLIP's pre-training data.

7.2. Generalization Gap

Both tasks show a consistent gap between cross-validation scores and official test-set results. On Task 2.1, Marsad_1 drops from CV F1 $= 0.7915$ to a test-set F1 of 0.6876, a drop of approximately 10 points. On Task 3.1, Marsad_2 similarly drops from CV F1 $= 0.7755$ to a test-set F1 of 0.6656, a drop of approximately 11 points. In both cases, the likely cause is the same: oversampling the minority class by duplicating slightly altered text instances inflates CV scores within the augmented distribution but does not generalize to unseen test instances.

7.3. Challenges of the LeWiDi Paradigm

None of our models were explicitly trained with a soft-label objective. All classification models were trained against hard majority-vote labels and converted post-hoc to soft predictions via probability calibration. This is a fundamental limitation: models optimized for hard-label accuracy may produce overconfident predictions that do not reflect genuine distributional uncertainty, leading to poor ICM-Soft scores even when hard-label F1 is reasonable. For instance, on Task 2.1 Marsad_3 achieves the best hard-evaluation rank (41st, F1 $= 0.7219$) yet ranks only 50th under the soft evaluation (ICM-Soft $= -0.4212$), directly illustrating this soft–hard trade-off. Future work should train models directly on annotator vote distributions rather than majority-vote labels, so that the model learns to reflect genuine annotation disagreement rather than converting hard predictions to soft probabilities after the fact.

8. Conclusion

We presented the Marsad team's participation in EXIST 2026 Tasks 2.1 and 3.1. For Task 2.1, we evaluated a suite of multimodal models under stratified 5-fold cross-validation on 3,984 bilingual memes augmented to 4,737 instances, submitting fine-tuned CLIP ViT-L/14 (Marsad_1), CLIP fusion large + LogReg (Marsad_2), and CLIP fusion + SVM (Marsad_3). For Task 3.1, we applied the same pipeline to 2,524 TikTok videos using 5-frame average pooling, submitting a weighted soft-label ensemble of the top-3 CV models (Marsad_1), fine-tuned CLIP ViT-L/14 alone (Marsad_2), and XLM-RoBERTa-large (Marsad_3). Our best Task 2.1 run achieved rank 17/141 on ICM-Soft overall and rank 11/141 on English; our best Task 3.1 run achieved rank 35/60 on ICM-Soft and rank 54/75 on ICM-Hard. Three findings stand out: (1) end-to-end fine-tuned CLIP ViT-L/14 consistently outperformed feature fusion and text-only baselines on both modalities; (2) both tasks showed a 10-point CV-to-test F1 drop, likely caused by minority-class augmentation inflating CV scores; and (3) a consistent English-Spanish performance gap of 0.54–0.66 ICM points highlights the limits of English-centric pre-training for multilingual sexism detection.

Future work will focus on: (1) training models directly on annotator vote distributions rather than majority-vote labels, so the model learns to reflect genuine annotation disagreement from the start; (2) language-specific fine-tuning or Spanish-augmented pre-training to close the EN–ES gap; (3) replacing random word-deletion augmentation with back-translation or paraphrase-based alternatives that do not inflate CV scores; and (4) integrating the physiological sensor data (EEG, heart rate, eye-tracking) provided by EXIST 2026 as auxiliary training signals.

9. Acknowledgments

This participation was supported by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI).

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 4.6 in order to: assist in drafting and structuring the paper, editing for clarity, and formatting the LaTeX source. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[2] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (extended overview), in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[3] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: https://aclanthology.org/2022.emnlp-main.731/. doi:10.18653/v1/2022.emnlp-main.731.

[4] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5809–5819. URL: https://aclanthology.org/2022.acl-long.399/. doi:10.18653/v1/2022.acl-long.399.

[5] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305/. doi:10.18653/v1/2023.semeval-1.305.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[7] 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, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[8] 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, A. García Seco de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, 2025, pp. 266–289.

[9] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes

challenge: Detecting hate speech in multimodal memes, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 2611–2624. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.

[10] 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, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[11] F. Carlsson, P. Eisen, F. Rekathati, M. Sahlgren, Cross-lingual and multilingual CLIP, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 6848–6854. URL: https://aclanthology.org/2022.lrec-1.739/.

[12] J. Wang, Y. Liu, X. Wang, Assessing multilingual fairness in pre-trained multimodal representations, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 2681–2695. URL: https://aclanthology.org/2022.findings-acl.211/. doi:10.18653/v1/2022.findings-acl.211.