Semantica at EXIST 2026: A Two-Stage Multimodal Approach for Sexism Categorization in Memes

EXIST 2026 at CLEF 2026 - Task 2, Subtask 3: Sexism Categorization in Memes

Mabrouka Bessghaier[1,*], Wajdi Zaghouani[1]

[1]*Northwestern University in Qatar, Doha, Qatar*

Abstract

This paper describes the system submitted by our team Semantica to the EXIST shared task at CLEF 2026, specifically Task 2, Subtask 3: Sexism Categorization in Memes. The task requires multi-label classification of multilingual memes (English and Spanish) into five fine-grained categories of sexism: Ideological & Inequality, Stereotyping & Dominance, Objectification, Sexual Violence, and Misogyny & Non-Sexual Violence. We propose a two-stage multimodal pipeline combining XLM-RoBERTa for multilingual text encoding, a frozen CLIP vision encoder for image feature extraction, CLIP's text encoder for cross-modal alignment signals, and handcrafted enriched bilingual lexicon features. Two runs were submitted sharing the same architecture but differing in training strategy: Semantica_1 uses inverse-frequency class weighting with per-class threshold optimization, while Semantica_2 additionally applies oversampling of rare categories and boosted class weights. In the soft-soft evaluation, Semantica_1 ranked 24th out of 115 systems , substantially outperforming the majority baseline. Hard-label performance was more modest, with both runs ranking in the lower half of the hard-hard leaderboard, revealing a gap between well-calibrated probability distributions and reliable discrete predictions. Our results highlight the persistent challenge of detecting rare sexism categories, particularly Misogyny & Non-Sexual Violence, and motivate future work on sensorial data integration and more targeted data augmentation strategies.

Keywords

sexism detection, multimodal classification, meme analysis, XLM-RoBERTa, CLIP, EXIST 2026, CLEF 2026

1. Introduction

Online sexism increasingly manifests in multimodal and implicit forms, with internet memes presenting a particularly challenging medium for automatic detection. Memes combine visual imagery and embedded text in ways that can amplify, ironize, or recontextualize sexist content, making them resistant to text-only or image-only analysis [1]. The EXIST (sEXism Identification in Social neTworks) shared task series [2, 3, 4, 5] has provided a growing benchmark for systems tackling this problem across multiple languages and media types.

EXIST 2026 [5, 6] introduces three key novelties compared to prior editions: a focus on exclusively multimedia content (memes and TikTok videos), bilingual data in English and Spanish, and the integration of Human-Centered AI principles. Specifically, selected subjects were exposed to the multimedia stimuli while their physiological and behavioral responses were recorded, producing multimodal sensorial signals (including eye tracking (ET), heart rate (HR), and electroencephalography (EEG)) that enrich traditional annotation labels and provide insight into how people unconsciously process and react to sexist content [7]. This Human-Centered framework distinguishes EXIST 2026 from all prior editions.

Among the six subtasks offered, we participate in **Task 2, Subtask 3: Sexism Categorization in Memes**, which requires assigning memes to one or more of five fine-grained sexism categories, given that a meme has already been identified as sexist. This subtask presents several compounding challenges. The five categories are not mutually exclusive (i.e., a single meme can belong to multiple categories simultaneously). The label distribution is heavily skewed, with some categories (notably Misogyny & Non-Sexual Violence) comprising only a small fraction of the training data. Furthermore,

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

*Corresponding author.

mabrouka.bessghaier@northwestern.edu (M. Bessghaier); wajdizaghouani@northwestern.edu (W. Zaghouani)

understanding a meme requires reasoning over both its visual and textual content jointly, since neither modality alone captures the full meaning [8].

We propose a **two-stage multimodal pipeline** in which Stage 1 performs binary sexism detection and Stage 2 assigns fine-grained categories to memes identified as sexist. Both stages use a shared fusion architecture combining XLM-RoBERTa [9], a frozen CLIP [10] vision encoder, CLIP's text encoder for cross-modal alignment, and a handcrafted bilingual enriched lexicon. We submitted two runs sharing this architecture but differing in their Stage 2 training strategies.

The paper is organized as follows. Section 2 describes the task and dataset. Section 3 details our system. Section 4 describes the submitted runs. Section 5 presents and analyzes the results. Section 6 concludes with directions for future work.

2. Task and Dataset Description

2.1. Task Definition

Task 2, Subtask 3 of EXIST 2026 is a multi-label hierarchical sexism categorization task applied to memes. The label space has two levels: a top level distinguishing sexist (YES) from non-sexist (NO) content, and a second level with five mutually non-exclusive sexism categories applicable to sexist memes:

- **IDEOLOGICAL-INEQUALITY**: discredits the feminist movement, rejects gender inequality, or presents men as victims of gender-based oppression.
- **STEREOTYPING-DOMINANCE**: expresses false ideas about women's suitable roles (domestic, submissive, etc.) or claims male superiority.
- **OBJECTIFICATION**: presents women as objects, hypersexualizes female attributes, or enforces beauty standards.
- **SEXUAL-VIOLENCE**: makes sexual suggestions, requests sexual favors, or references sexual harassment or assault.
- **MISOGYNY-NON-SEXUAL-VIOLENCE**: expresses hatred or non-sexual violence toward women.

Systems may submit **hard labels** (a single predicted label set per instance) or **soft labels** (a probability distribution over all categories). The evaluation framework follows a *learning with disagreement* paradigm [4]: rather than collapsing annotator votes into a single gold label, the full distribution of annotator labels is retained. Hard evaluation uses ICM as the official metric; soft evaluation uses ICM-Soft, both computed via the PyEvALL library.

2.2. Dataset

The EXIST 2026 Memes Dataset [5] builds upon the EXIST 2025 dataset, extending it with sensor-based physiological annotations and removing a small number of duplicate instances detected after the 2025 release. The dataset comprises **3,984 training memes** and **1,053 test memes** in English and Spanish. The test set contains 1,053 memes with balanced language distribution.

The category distribution is substantially imbalanced. *IDEOLOGICAL-INEQUALITY* and *STEREOTYPING-DOMINANCE* are the most frequent categories, whereas *MISOGYNY-NON-SEXUAL-VIOLENCE* is the least represented. To mitigate the effects of class imbalance, we employed inverse-frequency class weighting. The weight assigned to each category c was computed as:

$$w_c = \frac{\frac{1}{n_c+1}}{\min_{c'} \left(\frac{1}{n_{c'}+1} \right)}, \tag{1}$$

where n_c denotes the number of training instances labeled with category c. This normalization ensures that the most frequent category, *STEREOTYPING-DOMINANCE*, receives a weight of $1.0\times$, while the rarest category, *MISOGYNY-NON-SEXUAL-VIOLENCE*, is assigned a weight of $2.58\times$.

In addition to textual and visual content, EXIST 2026 provides **physiological sensor data** for each meme, collected while subjects viewed the stimuli specifically Eye Tracking (24 features per subject), Heart Rate (4 features per subject), and EEG (80 features per subject).

3. System Description

3.1. Overall Pipeline

Our system follows a two-stage cascade that mirrors the hierarchical label structure of the task:

1. **Stage 1: Sexism Detection:** A binary classifier determines whether a meme is sexist (YES) or not (NO).
2. **Stage 2: Sexism Categorization:** For memes classified as sexist, a multi-label classifier assigns one or more of the five sexism categories.

This design ensures that no meme simultaneously receives the NO label and a sexism category, avoiding logical contradictions in the output. Each stage trains an independent model with the same underlying architecture but different output dimensionality (1 unit for Stage 1, 5 units for Stage 2).

3.2. Feature Representations

3.2.1. Multilingual Text Encoding (XLM-RoBERTa)

Meme text is encoded using **XLM-RoBERTa** [9] , a transformer pre-trained on 100 languages. The 768-dimensional [CLS] token embedding serves as the text representation. XLM-RoBERTa is fine-tuned during training, allowing it to adapt to the multilingual and domain-specific language of sexist meme content.

In preliminary experiments using XLM-RoBERTa alone, the model struggled to distinguish between sexism categories, as meme text is often short, ambiguous, or ironic, and textual content alone is insufficient to capture the full meaning conveyed by the image-text combination. This motivated the addition of visual and cross-modal features described below.

3.2.2. Visual Encoding (Frozen CLIP Vision Encoder)

Meme images are encoded using the **CLIP ViT-B/32** vision encoder [10], producing a 512-dimensional embedding via the visual projection layer. CLIP is pre-trained on large-scale image-text pairs, making it well-suited to the visual content of internet memes. We keep the vision encoder frozen throughout training to reduce computational cost and avoid overfitting on the relatively small dataset.

3.2.3. CLIP Text Encoder for Cross-Modal Alignment

We additionally incorporate **CLIP's text encoder** to produce a separate 512-dimensional text embedding projected into the same space as the image embedding. Because CLIP is trained to align image and text representations, the relationship between a meme's CLIP image embedding and CLIP text embedding encodes an implicit cross-modal alignment signal. This is especially informative for ironic or sarcastic memes, where visual and textual content are semantically divergent, which is a configuration that is common in sexist meme content [8]. The CLIP text encoder is also frozen. CLIP embeddings for both modalities are pre-computed and cached before training to avoid repeated forward passes.

Table 1
Enriched lexicon: keyword counts and representative English examples per category.

Category	EN	ES	Example keywords (EN)
IDEOLOGICAL-INEQUALITY	24	22	*feminazi, misandry, mansplaining, patriarchy, victimhood*
STEREOTYPING-DOMINANCE	39	35	*kitchen, housewife, submissive, nagging*
OBJECTIFICATION	39	38	*cleavage, curvy, virginity*
SEXUAL-VIOLENCE	35	34	*rapist, groped, nonconsensual, stalking, catcalling*
MISOGYNY-NON-SEXUAL-VIOLENCE	54	49	*subhuman, worthless, slag, vermin, psycho*

3.2.4. Enriched Bilingual Lexicon Features

A key contribution of our system is a carefully designed **enriched bilingual lexicon** covering English and Spanish, built specifically for this task. The lexicon has three components: (1) category-specific keyword lists, (2) syntactic phrase patterns, and (3) global cross-category cue sets.

Keyword Lists. For each of the five sexism categories, we curate a list of discriminative keywords in both languages, chosen to reflect the specific semantic register of each category. The design principle was to avoid generic terms that would create false positives across categories, instead targeting vocabulary that is characteristic of each type of sexist expression. Table 1 summarizes the keyword counts per category and provides keyword examples.

MISOGYNY-NON-SEXUAL-VIOLENCE intentionally receives the largest keyword set (54 EN / 49 ES) to compensate for its extreme underrepresentation in the training data. Its keywords cover violent threats (*kill, stab, choke*), dehumanizing slurs (*vermin, subhuman, slag*), and misogynistic insults (*worthless, psycho*), as well as their Spanish equivalents. Keywords for IDEOLOGICAL-INEQUALITY deliberately exclude broad neutral terms such as *rights*, *equal*, or *gender*, which would cause false positives for content that discusses gender issues without being sexist.

Phrase Patterns. Beyond individual keywords, we define **regular expression patterns** that capture frequent multi-word sexist constructions in both languages, applied to lowercased accent-normalized text. Phrase patterns are particularly important for capturing constructions that individual keywords cannot reliably identify: *"deserve"* alone is ambiguous, but *"she deserves to die"* is unambiguously MISOGYNY; *"consent"* alone may appear in educational contexts, but *"without her consent"* can be a SEXUAL-VIOLENCE signal. Other illustrative examples include *"go back to the kitchen"* and *"make me a sandwich"* (STEREOTYPING-DOMINANCE), *"feminism is cancer"* and *"men are the real victims"* (IDEOLOGICAL-INEQUALITY), *"rate her body"* (OBJECTIFICATION), and *"sexual assault"* and *"she was asking for it"* (SEXUAL-VIOLENCE), along with their Spanish equivalents such as *"calladita te ves mas bonita"* and *"sin su consentimiento"*. In total, 58 patterns are defined across the five categories.

Global Cue Features. Three additional features capture general emotionally charged language that may signal sexism regardless of specific category. We define three dedicated word sets: a **negative cue set** (e.g., *hate, worthless, disgusting*), a **violence cue set** (e.g., *kill, beat, stab, burn, matar, golpear*), and a **sexual cue set** (e.g., *rape, harass, consent, violar, acosar*). For each set, we count how many cue tokens appear in the meme text and normalize by text length, yielding three scalar features. These complement the more specific per-category lexicon features by providing a coarse signal of overall hostility or threat in the text.

Table 2
Model architecture summary. Both stages and both runs use the identical architecture.

Component	Dim	Trainable
XLM-RoBERTa ([CLS])	768	Yes
CLIP Vision Encoder	512	No (frozen)
CLIP Text Encoder	512	No (frozen)
Lexicon features	28	–
Concatenated input	1820	–
Fusion layer (ReLU + Dropout)	256	Yes
Classifier head	1 / 5	Yes

Feature Extraction. For each meme text, we apply Unicode normalization (lowercasing and accent removal) before matching. For each category c, we extract five sub-features: (1) frequency of matched lexicon tokens normalized by text length, (2) unique matched-token coverage normalized by text length, (3) a binary trigger flag indicating any lexicon match, (4) log-scaled count of unique matched tokens ($\log(1+n)$), and (5) log-scaled count of matched phrase patterns. Combined with the three global cue features, this yields a **28-dimensional** feature vector ($5 \times 5 + 3$). The log scaling reduces the influence of texts that match many keywords of the same type, preventing high-frequency memes from dominating the gradient.

These lexicon features provide interpretable, explicit prior knowledge that the neural components may not reliably acquire from limited training data, particularly for the rare categories. Ablation experiments during development confirmed that removing the lexicon features reduced the validation macro-F1 score, demonstrating their substantial contribution to system performance.

3.3. Model Architecture

The four feature vectors are concatenated to form a joint 1820-dimensional multimodal representation:

$$\mathbf{h} = [\mathbf{h}_{\text{XLM-R}};\ \mathbf{h}_{\text{CLIP-img}};\ \mathbf{h}_{\text{CLIP-txt}};\ \mathbf{h}_{\text{lex}}] \tag{2}$$

where $\mathbf{h}_{\text{XLM-R}} \in \mathbb{R}^{768}$, $\mathbf{h}_{\text{CLIP-img}} \in \mathbb{R}^{512}$, $\mathbf{h}_{\text{CLIP-txt}} \in \mathbb{R}^{512}$, and $\mathbf{h}_{\text{lex}} \in \mathbb{R}^{28}$. This is passed through a fusion layer:

$$\mathbf{z} = \text{Dropout}(\text{ReLU}(\mathbf{W}_f\mathbf{h} + \mathbf{b}_f)), \quad \mathbf{z} \in \mathbb{R}^{256} \tag{3}$$

followed by a linear classifier head. Both runs and both stages share this architecture identically; the differences between runs lie exclusively in training strategy (Section 4).

3.4. Training Setup

Both stages use **Binary Cross-Entropy with Logits (BCEWithLogitsLoss)**, which naturally handles multi-label outputs and allows soft annotator proportion distributions to serve directly as training targets, consistent with the learning-with-disagreement paradigm of EXIST 2026. The optimizer is **AdamW** [11] with weight decay 0.01 and gradient clipping at max norm 1.0. A linear learning rate warmup covers the first 10% of training steps.

Training labels for Stage 2 are aggregated from annotator responses following the task's hard-label derivation rule: a category is assigned a hard label of 1 if annotated by more than one annotator. If no category satisfies this threshold, the most frequently annotated category is used as a fallback to ensure every sexist training instance contributes a learning signal. Soft labels are computed as the proportion of valid annotators who selected each category.

We apply an 85/15 train/validation split, yielding 3,386/598 instances for Stage 1 and 1,702/301 for Stage 2. During validation, gold files are filtered to include only validation IDs before passing to PyEvALL, ensuring the evaluation correctly reflects our held-out split.

Table 3
Validation results at the best Stage 2 checkpoint. Stage 1 F1 is binary (sexist/non-sexist). Stage 2 category F1 scores are at the best checkpoint epoch; threshold F1 is after per-class threshold optimization.

Metric	Semantica_1	Semantica_2
Stage 1 F1 (binary)	0.7944	0.7927
IDEOLOGICAL-INEQUALITY	0.66	0.68
STEREOTYPING-DOMINANCE	0.62	0.65
OBJECTIFICATION	0.76	0.74
SEXUAL-VIOLENCE	0.56	0.55
MISOGYNY-NON-SEXUAL-VIOLENCE	0.37	0.36
Stage 2 Macro F1 (checkpoint)	0.5961	0.5999
Stage 2 Macro F1 (per-class thresh.)	0.6122	0.6124

3.5. Threshold Optimization

Stage 2 converts predicted probabilities to hard labels via threshold. We conduct two searches on the validation set: (1) a global threshold search over $\{0.05, 0.10, \ldots, 0.50\}$ selecting the value that maximizes validation macro-F1, and (2) a per-class threshold search that independently optimizes each category's threshold over the same grid. If the per-class strategy yields higher macro-F1 than the best global threshold, it is adopted. A fallback rule ensures that if no category exceeds its threshold, the highest-probability category is assigned.

4. Submitted Runs

We submitted two runs, each providing both hard and soft label predictions for all 1,053 test memes. Both runs use the 1820-dimensional architecture described above. The differences are exclusively in Stage 2 training strategy.

Semantica_1 (Run 1): Stage 1 trains for 5 epochs at LR 2×10^{-5}, dropout 0.3. Stage 2 trains for 15 epochs at LR 2×10^{-5}, dropout 0.3, using inverse-frequency class weights without additional boosting. Class weights (normalized): IDEOLOGICAL-INEQUALITY $1.22\times$, STEREOTYPING-DOMINANCE $1.00\times$, OBJECTIFICATION $1.09\times$, SEXUAL-VIOLENCE $2.15\times$, MISOGYNY-NON-SEXUAL-VIOLENCE $2.58\times$. Per-class threshold optimization selects thresholds [IDEO=0.40, STER=0.15, OBJE=0.25, SEXU=0.20, MISO=0.25], achieving validation macro-F1 of 0.6122 vs. global threshold F1 of 0.6033.

Semantica_2 (Run 2): Stage 1 trains identically to Run 1 (5 epochs, LR 2×10^{-5}). Stage 2 trains for 20 epochs at a lower LR of 1×10^{-5}, dropout 0.3, with two additional class imbalance strategies: (1) **oversampling** of rare categories in the training set (MISOGYNY instances are duplicated $2\times$ and SEXUAL-VIOLENCE instances are duplicated $1\times$ additional copy, expanding Stage 2 training from 1,702 to 2,768 instances; (2) **boosted class weights**) the MISOGYNY weight is multiplied by $2.0\times$ and SEXUAL-VIOLENCE by $1.5\times$ beyond the inverse-frequency baseline. Per-class thresholds [IDEO=0.30, STER=0.35, OBJE=0.25, SEXU=0.35, MISO=0.35] are selected.

5. Results and Analysis

Validation Results Table 3 reports Stage 1 binary F1 and per-category Stage 2 F1 scores on the internal validation set at the best checkpoint for each run, alongside the per-class threshold macro-F1.

Both runs achieve similar validation performance. The key improvement over our early development runs (which scored F1 = 0.0 on MISOGYNY-NON-SEXUAL-VIOLENCE) is attributable to the enriched lexicon with category-specific phrase patterns and the class weighting strategies, which together enabled the model to partially learn this highly underrepresented category.

Official Test Results Our best run, Semantica_1, ranked 24th out of 115 systems in the soft-soft evaluation (ICM-Soft: -5.63, ICM-Soft Norm: 0.20), well above the majority baseline (rank 74, ICM-Soft Norm: 0.00). In the hard-hard evaluation, Semantica_1 ranked 156th out of 184 systems (ICM-Hard: -2.35, F1: 0.385), falling below the majority baseline on the ICM scale despite achieving a substantially higher F1. Semantica_2 ranked lower in both tracks (39th soft, 158th hard), confirming that the aggressive oversampling and boosted weights did not generalize to the test set.

Analysis. Despite Semantica_2 achieving marginally higher validation macro-F1 ($+0.003$ at checkpoint), Semantica_1 outperforms it across all evaluation tracks and both languages. The oversampling and boosted weights in Semantica_2, combined with the lower learning rate (1×10^{-5} vs. 2×10^{-5}) and inflated training set (2,768 vs. 1,702 instances), appear to overfit to the validation distribution without generalizing to the test set. The oversampling did not add new information, while the boosted weights compounded this bias, pushing the model toward over-predicting MISOGYNY-NON-SEXUAL-VIOLENCE and SEXUAL-VIOLENCE even when the evidence was ambiguous. A notable exception occurs in the Spanish hard-hard evaluation, where Semantica_2 slightly outperforms Semantica_1 (rank 158 vs. 165), suggesting the oversampling strategy may have marginal benefit for Spanish-specific rare-category patterns.

In early development runs without enriched lexicon features and class weighting, this category consistently scored F1 = 0.000. The enriched lexicon and class weighting strategies improved validation F1 to 0.37–0.36. However, these improvements did not fully generalize to the test set. This failure is rooted in extreme class imbalance and the subtlety of misogynistic expressions, which often lack distinctive surface-level markers and overlap with general offensive language.

The enriched lexicon proved essential to system performance. Ablation experiments during development showed that removing the lexicon features reduced validation macro-F1 from 0.5961 to approximately 0.4126. The phrase patterns were particularly valuable for MISOGYNY-NON-SEXUAL-VIOLENCE, which scored F1 = 0.000 in early runs without them, improving to 0.37 once category-specific patterns and weighted cues were introduced. The global cue features (capturing general negative, violent, and sexual language) provided a complementary coarse signal that helped the model flag emotionally charged memes even when specific category keywords were absent.

6. Conclusion

We presented the Semantica system for EXIST 2026 Task 2, Subtask 3: a two-stage multimodal pipeline combining XLM-RoBERTa, frozen CLIP visual and text encoders, and an enriched bilingual lexicon for fine-grained sexism categorization in multilingual memes. Our soft submissions achieved competitive rankings (24th and 39th out of 115), demonstrating well-calibrated probability distributions relative to annotator soft labels. Hard-label performance was more modest, revealing a gap between distributional calibration and reliable discrete predictions, particularly at the top level of the label hierarchy.

Acknowledgments

The authors thank the EXIST 2026 organizers for providing the dataset, sensorial annotations, evaluation framework, and leaderboard infrastructure. Our work was supported by the grant NPRP14C0916-210015, awarded by the Qatar Research, Development and Innovation Council (QRDI).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with code development and debugging. After using this tool, the authors reviewed and edited all content as needed and take full responsibility for the publication's content.

References

[1] D. Kiela, H. Firooz, A. Nandan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 2611–2624.

[2] 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.

[3] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024—learning with disagreement for sexism identification and characterization in tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117.

[4] 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: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 266–289.

[5] 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.

[6] 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. B.-C. no, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS.org, 2026.

[7] I. Arcos, P. Rosso, E. Gomis, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and EEG signals, in: Proceedings of the Language Resources and Evaluation Conference (LREC 2026), 2026.

[8] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, MEMEWEAVER: Inter-meme graph reasoning for sexism and misogyny detection, in: Findings of the Association for Computational Linguistics: EACL 2026, 2026.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, Édouard Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 8440–8451.

[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: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[11] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Proceedings of the 7th International Conference on Learning Representations, 2019.