AI Wizards at EXIST 2026: Hierarchical Soft-Label Learning for Multimodal Sexism Identification in Memes

Notebook for the EXIST Lab at CLEF 2026

Matteo Fasulo[1,*,†], Antonio Gravina[2,†], Luca Tedeschini[3,†] and Luca Babboni[4,†]

[1]Swiss Data Science Center, ETH Zürich, Andreasturm, Andreasstrasse 5, 8092 Zürich, Switzerland
[2]Everest Systems GmbH, Max-Jarecki-Straße 21, 69115 Heidelberg, Germany
[3]Villanova.ai S.P.A, Località Sa Illetta, SS 195 KM 2.3, 09123 Cagliari, Italy
[4]Independent researcher

Abstract

We present the **AI Wizards** submission to EXIST 2026 for multimodal sexism identification in memes. The task is composed of three, increasingly harder subtasks. We model them hierarchically as conditional soft-label prediction over empirical annotator distributions. Our system maps fixed Gemini Embedding 2 vision-language representations through a lightweight Gated MLP trained with KL divergence and homoscedastic uncertainty weighting. Our submissions ranked **first** on Task 2.3 and **fourth** on Tasks 2.1 and 2.2 on the official Soft-Soft leaderboards. The code is available at https://github.com/NLP-AI-Wizards/EXIST-2026.

Keywords

sexism identification, multimodal memes, learning with disagreement, hierarchical classification, soft labels

1. Introduction

Sexist content online reinforces women's marginalization and exclusion from digital spaces [1]. Digital platforms often amplify offline patriarchal structures, where hostile and benevolent sexism jointly legitimize gender inequality [2]. Exposure to such content causes measurable psychological harm (e.g., elevated anxiety, depression, and self-censorship), which silences women's participation in public discourse [3]. Robust automated detection is therefore a critical intervention for digital safety.

Memes pose unique challenges: their meaning emerges non-linearly from text-image interaction, often relying on irony, humor, or implicit cultural knowledge to provide plausible deniability [4, 5]. Detection systems must model how modalities jointly construct gendered meanings that may be implicit, ironic, or ambiguous.

EXIST 2026 [6, 7] structures sexism detection as three hierarchically nested subtasks, detailed in Section 3: binary sexism identification, source intention detection for sexist content, and fine-grained multi-label categorization across five sexism types. Critically, the task adopts the Learning with Disagreement (LeWiDi [8]) paradigm, requiring systems to predict the full empirical annotator distribution rather than a single majority label.

We address all three subtasks as hierarchical conditional multi-task learning over pre-computed vision-language embeddings, preserving semantic dependencies while directly optimizing for soft-label evaluation.

Our main contributions are as follows.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
matteo.fasulo@sdsc.ethz.ch (M. Fasulo); antonio.gravina@everest-erp.com (A. Gravina); ltedeschini@villanova.ai (L. Tedeschini); luca.babboni2@studio.unibo.it (L. Babboni)
https://matteofasulo.com (M. Fasulo); https://github.com/GravAnt (A. Gravina); https://github.com/LucaTedeschini (L. Tedeschini); https://github.com/ElektroDuck (L. Babboni)
0000-0002-7019-3157 (M. Fasulo); 0009-0007-9252-263X (A. Gravina); 0009-0006-0375-829X (L. Tedeschini); 0009-0001-5260-7467 (L. Babboni)

- **Multimodal representation:** Pre-trained Gemini Embedding 2 features processed through compact gated MLP blocks, avoiding fine-tuning of large generative models on limited data.
- **Dynamic multi-task weighting:** KL-divergence soft-label objective with learned homoscedastic uncertainty weights that automatically balance task contributions.
- **Hierarchical decoding:** Conditional loss masking and probabilistic decoding that respect the task taxonomy without architectural overhead.

2. Related Work

2.1. Multimodal Sexism Detection

Early detection systems focused on text-only platforms [9, 10]. The introduction of multimodal benchmarks (SemEval-2022 MAMI [4] and the Hateful Memes Challenge [5]) revealed that unimodal systems fail when toxicity arises only from text-image interaction [11]. The EXIST lab has tracked this evolution from text (2021–2023) to memes (2024) [12] to video (2025) [9]. As sexism detection has moved into the multimodal setting, it has increasingly drawn on architectures originally developed for the broader hate speech domain, which we review next.

2.2. Architectures and Representation

Contrastive vision-language models such as CLIP [13] and SigLIP [14] learn powerful joint image-text representations, but their training objective is general-purpose and not tailored to hateful content detection. Building on this foundation, Hate-CLIPper [15] extends the contrastive paradigm by introducing a bilinear fusion mechanism explicitly designed for hate speech detection, effectively specializing the general alignment learned by CLIP-like models into a task-specific framework for identifying hateful memes. In parallel, an alternative line of work has moved away from fusion-based specialization altogether, instead leveraging the emergent reasoning capabilities of large pretrained models: recent EXIST submissions explored LoRA fine-tuning and prompting of large models such as Claude Sonnet [16] to tackle hateful meme detection directly, trading architectural specialization for scale and generality. These two directions, however, share a common limitation for the sexism detection setting: neither explicitly models the annotation disagreement that characterizes subjective tasks like sexism labeling, nor the hierarchical structure of sexism sub-categories.

2.3. Learning with Disagreement and Hierarchy

Subjective tasks such as sexism detection exhibit high inter-annotator disagreement that majority voting erases, potentially silencing minority perspectives. The LeWiDi paradigm treats this disagreement as signal rather than noise, evaluating systems on their ability to predict full annotator distributions. Hierarchical classification for related tasks has been explored through cascaded architectures and conditional decoding, but few systems combine both hierarchy and soft-label prediction in a unified multi-task framework. Building on these three lines of work, our approach combines a compact contrastive backbone, following the efficiency principles of fusion-based architectures like Hate-CLIPper while avoiding the computational cost of large prompted models, with a disagreement-aware, hierarchical multi-task head that jointly predicts soft label distributions and sexism sub-categories. This design directly addresses the gap identified above: unlike prior CLIP-based or LLM-prompted systems, our model explicitly captures both annotator disagreement and hierarchical task structure within a single, efficient, and reproducible framework.

3. Dataset and Problem Formulation

EXIST 2026 formalizes the research problem through three hierarchically organized subtasks focused on multimodal sexist content:

- **Sexism Identification (Task 2.1):** A binary classification task aimed at determining whether a meme depicts a sexist situation or endorses sexist behavior (YES) or, conversely, does not contain sexist content (NO).
- **Source Intention Detection (Task 2.2):** A binary classification task defined exclusively for memes previously labeled as sexist. It distinguishes between DIRECT intention, where the communicative goal is to produce inherently sexist content or to incite sexist attitudes, and JUDGEMENTAL intention, where the content portrays sexist situations or behaviors with the explicit objective of condemning or criticizing them.
- **Sexism Categorization (Task 2.3):** A multi-label classification task that assigns each sexist meme to one or more of the following five categories:
    1. IDEOLOGICAL-INEQUALITY: The content delegitimizes the feminist movement, rejects or denies structural inequality between men and women, or portrays men as victims of gender-based oppression.
    2. STEREOTYPING-DOMINANCE: The content conveys essentialist or stereotypical beliefs about women that imply they are more suited to specific roles (e.g., mother, wife, family caregiver, faithful, tender, loving, submissive), are unfit for certain activities (e.g., driving, physically demanding work), or explicitly or implicitly asserts male superiority.
    3. OBJECTIFICATION: The content depicts women as objects, disregarding their dignity and personhood, or prescribes or describes physical attributes that women are expected to possess in order to conform to traditional gender roles (e.g., adherence to narrow beauty standards, hypersexualization of female bodies, representation of women's bodies as available for male use).
    4. SEXUAL-VIOLENCE: The content contains sexual insinuations, requests for sexual favors, or other forms of sexual harassment, including references to rape or sexual assault.
    5. MISOGYNY-NON-SEXUAL-VIOLENCE: The content expresses hatred, hostility, or non-sexual forms of violence directed toward women.

The dataset contains 3,984 training and 1,053 test memes in English and Spanish (Table 1). For the Soft-Soft track, each target represents the empirical annotator distribution. Tasks 2.2 and 2.3 are only semantically defined when sexism probability is non-zero, motivating our hierarchical conditional approach.

Table 1
EXIST 2026 dataset composition.

Language	Training	Test	Total
Spanish	1979	540	2519
English	2005	513	2518
Total	3984	1053	5037

3.1. Physiological Data and Modality Selection

EXIST 2026 provides auxiliary physiological recordings: EEG (16 channels, 5 frequency bands, 80 bandpower features), gaze data (200 Hz), and heart rate.

Among the three modalities, we prioritized EEG for quantitative analysis. EEG offers the highest feature dimensionality and the most direct, well-established link to implicit cognitive and evaluative processing in the psychophysiology literature [17], making it the most plausible candidate for capturing task-relevant variation among the provided physiological signals. Gaze and heart rate, by contrast, are lower-dimensional (respectively a 2D fixation trace and a single scalar signal) and reflect attentional and autonomic responses that are more directly tied to visual salience and general arousal than to the specific cognitive evaluation of sexist content. We therefore did not subject gaze and heart rate to the

same statistical testing procedure; their potential utility, particularly in combination with EEG or under non-linear modeling, remains untested and is left for future work.

To assess their discriminative utility, we applied the Principal Component Analysis (PCA) to retain 21 components explaining $\sim 95\%$ of total variance to the EEG features, then performed Multivariate Analysis of Variance (MANOVA) in the majority groups of soft-labels for each subtask (Table 2). Tasks 2.1 and 2.3 showed no significant linear separability. Although Task 2.2 reached statistical significance ($p_{\text{perm}} = 0.0010$), the effect was negligible (Wilks' $\lambda = 0.9946$, partial $\eta^2 < 0.01$). No individual EEG channel survived the Bonferroni correction ($\alpha = 6.25 \times 10^{-4}$).

Table 2
EEG separability analysis. PCA reduced features to 21 components. p_{perm} from 1,000-label permutation test.

Task	Sessions	Wilks' λ	p	p_{perm}	Sig. channels
Task 2.1	15,477	0.9988	0.6130	0.6034	0/80
Task 2.2	10,171	0.9946	7.9×10^{-5}	0.0010	0/80
Task 2.3	10,171	0.9918	0.4851	0.4855	0/80

Although non-linear biosignal integration remains an open direction, linear analysis revealed insufficient signal for the provided feature representations. We therefore prioritized robust semantic representations over physiological fusion, leaving deeper biosignal modeling for future work.

4. Methods

Our system rests on three principles: (1) extracting rich multimodal semantics without fine-tuning massive models on limited data; (2) predicting annotator disagreement distributions rather than forced labels; and (3) enforcing hierarchical constraints through conditional training and decoding.

4.1. Shared Backbone

Let $f_{\text{VLM}} : \mathcal{X} \to \mathbb{R}^{768}$ denote the frozen Gemini Embedding 2 encoder [18], producing a fixed embedding $\mathbf{x}$ per meme. We pass $\mathbf{x}$ through a lightweight Gated MLP backbone consisting of one SwiGLU block [19] with expansion factor 2 and dropout 0.2:

$$\text{SwiGLU}(\mathbf{x}) = (\mathbf{x}\mathbf{W}_1) \odot \text{SiLU}(\mathbf{x}\mathbf{W}_2), \tag{1}$$

$$\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \mathbf{W}_3 \left(\text{SwiGLU} \left(\text{LN}(\mathbf{h}^{(l)}) \right) \right). \tag{2}$$

The gating mechanism learns to suppress benign confounders and amplify hostile cross-modal contradictions. The resulting representation $\mathbf{h}$ feeds three independent linear heads producing logits for each subtask. The trainable portion contains only 3.5M parameters.

4.2. Soft-Label Training and Dynamic Task Weighting

The Soft-Soft evaluation expects systems to predict full annotator distributions. We train each head against empirical soft targets using KL divergence [20], which penalizes the model when it places mass on categories annotators collectively rejected. For Task 2.3, binary KL is applied independently per category and averaged.

Balancing three heterogeneous tasks—differing in label spaces, class imbalance, conditional support, and intrinsic difficulty—poses an optimization challenge. Fixed scalar weights require extensive tuning and cannot adapt as tasks converge at different rates. We instead adopt homoscedastic uncertainty weighting [21], introducing learnable variance parameters σ_i:

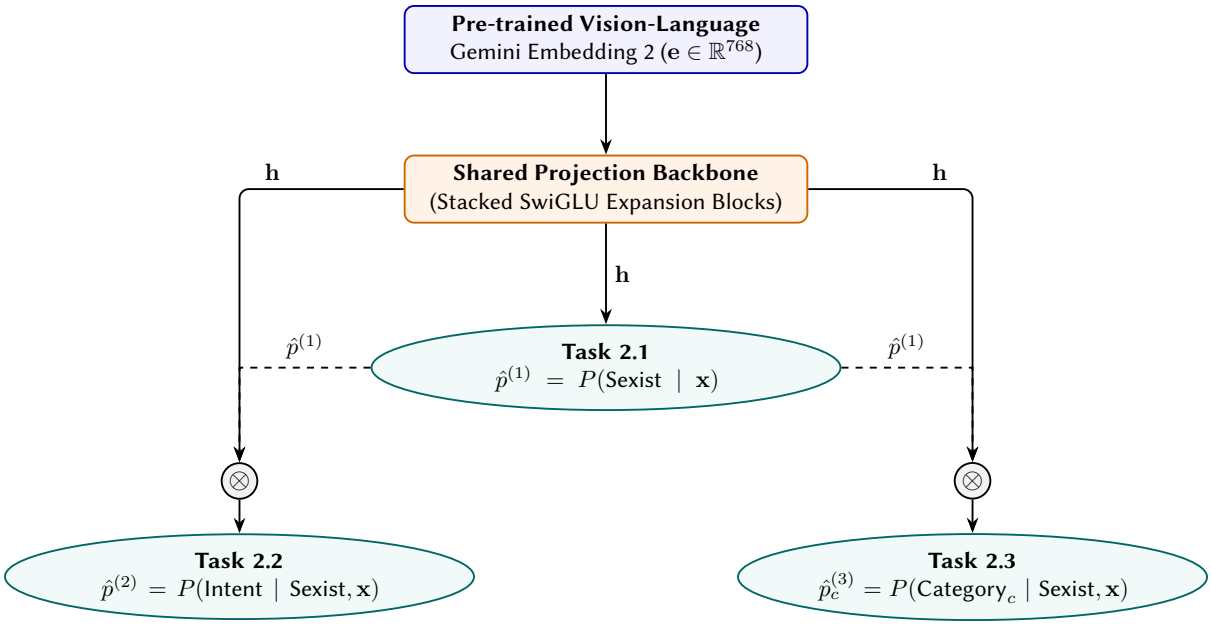

Figure 1: Architectural data flow illustrating the hierarchical conditional dependencies. Pre-trained Gemini 2 embeddings are mapped to a shared semantic representation $\mathbf{h}$ via SwiGLU blocks. This shared representation is routed to all classification heads (solid lines). The dashed lines and multiplication nodes ($\otimes$) represent the conditional dependency: structurally enforced via detached soft-gating in Run 3, and probabilistically enforced across all runs via joint probability decoding ($P(\text{Subtask}) = \hat{p}^{(1)} \cdot \hat{p}^{(\text{sub})}$) and masked multi-task loss.

$$\mathcal{L}(\mathbf{W}, \sigma_1, \sigma_2, \sigma_3) = \sum_{i=1}^{3} \left(\frac{1}{2\sigma_i^2} \mathcal{L}_i + \log \sigma_i \right). \tag{3}$$

This formulation enables self-paced learning: when task losses are high, σ_i increases to reduce that task's contribution; the $\log \sigma_i$ term prevents complete task abandonment. Finally, losses for Tasks 2.2 and 2.3 are masked to propagate gradients only for instances where downstream annotations exist, preventing non-sexist memes from forcing arbitrary updates on conditional heads.

4.3. Detached Soft-Gating (Ablation)

As an architectural ablation (submitted as Run 3), we tested explicit structural conditioning: before downstream heads, the shared representation is scaled by the predicted sexism probability with stopped gradients:

$$\mathbf{h}_{\text{downstream}} = \mathbf{h} \odot \text{sg}\left[\hat{p}^{(1)}\right]. \tag{4}$$

This suppresses downstream features when sexism probability is low. The stop-gradient operator prevents downstream losses from corrupting the Task 2.1 detector. The comparable performance of this variant (Section 6) suggests hierarchical constraints are adequately captured through conditional loss masking and probabilistic decoding alone, without architectural gating.

5. Training and Inference

We split training data into 80/10/10% partitions stratified by Task 2.1 majority label. Models are optimized with AdamW (learning rate 10^{-4}, weight decay 10^{-2}, batch size 8), OneCycleLR scheduling with 30% warm-up and cosine annealing, `bf16-mixed` precision, and early stopping on validation loss. We

submitted three runs: two base Gated MLP models with different random seeds (Runs 1–2) and the detached soft-gating variant (Run 3).

5.1. Inference

Soft-label decoding. For Task 2.2, the joint distribution over {JUDGEMENTAL, DIRECT, NO} is:

$$\left\{\hat{p}^{(1)}\hat{p}^{(2)},\ \hat{p}^{(1)}(1-\hat{p}^{(2)}),\ 1-\hat{p}^{(1)}\right\}. \tag{5}$$

For Task 2.3, each category probability is $\hat{p}^{(1)}\hat{p}_c^{(3)}$, with $1-\hat{p}^{(1)}$ assigned to the non-sexist complement.

Hard-label decoding (submitted to Hard-Hard track without threshold tuning). We threshold at 0.5: if $\hat{p}^{(1)} < 0.5$, all downstream labels are null. Otherwise, Task 2.2 uses $\hat{p}^{(2)} \geq 0.5$. For Task 2.3, the active set is $\{c \mid \hat{p}_c^{(3)} \geq 0.5\}$. To satisfy the ontological constraint that sexist memes must have ≥ 1 category, empty sets are resolved by selecting the maximum-probability category. No task-specific tuning was performed.

6. Results

6.1. Local Evaluation

Tables 3 and 4 report results on our 10% test split using official ICM metrics [22], with cross-entropy (CE) included for binary tasks.

Table 3
Soft-label evaluation on local test split.

	Task 2.1			Task 2.2			Task 2.3	
Model	ICM-S	ICM-S Nr	CE	ICM-S	ICM-S Nr	CE	ICM-S	ICM-S Nr
Run 1 (Base)	0.1648	0.5258	0.9340	**-0.6040**	**0.4367**	1.5011	-3.6935	0.3111
Run 2 (Base)	**0.2025**	**0.5317**	**0.9314**	-0.7807	0.4181	**1.4822**	**-3.2130**	**0.3356**
Run 3 (Gating)	0.1900	0.5297	0.9448	-0.8484	0.4110	1.5426	-3.2546	0.3335

Table 4
Hard-label evaluation on local test split. F1Y = F1 for YES, M F1 = macro F1.

	Task 2.1			Task 2.2			Task 2.3		
Model	ICM-H	ICM-H Nr	F1Y	ICM-H	ICM-H Nr	M F1	ICM-H	ICM-H Nr	M F1
Run 1 (Base)	**0.3308**	**0.6690**	**0.7524**	**0.1295**	**0.5445**	**0.5189**	-0.4355	0.4098	0.4045
Run 2 (Base)	0.2635	0.6346	0.7407	-0.1270	0.4563	0.4886	**-0.3319**	**0.4313**	**0.4530**
Run 3 (Gating)	0.2917	0.6490	0.7488	-0.0476	0.4836	0.4816	-0.4213	0.4128	0.4292

Local results reveal meaningful seed variance, especially for downstream tasks—consistent with the small-data, high-disagreement setting. Run 2 performed best for Task 2.3, while Run 1 led for Task 2.2. The gating variant (Run 3) performed comparably to the base architecture, confirming that hierarchical constraints are effectively captured through conditional losses and probabilistic decoding without architectural modification.

6.2. Official Leaderboard

Table 5 reports our best official Soft-Soft results per task. Our system ranked **1st** on Task 2.3 and **4th** on Tasks 2.1 and 2.2.

Table 6 reports our best Hard-Hard results, obtained without threshold tuning or hard-label-specific calibration. The gap between Soft-Soft and Hard-Hard rankings confirms that distribution prediction

Table 5
Official Soft-Soft leaderboard results.

Task	System	Rank	ICM-Soft	ICM-Soft Norm
2.1	**aiwizards_3**	**4**	**0.2323**	**0.5373**
2.1	aiwizards_1	5	0.2039	0.5328
2.1	aiwizards_2	6	0.1846	0.5297
2.2	**aiwizards_1**	**4**	**-0.6720**	**0.4285**
2.2	aiwizards_3	5	-0.7623	0.4189
2.2	aiwizards_2	7	-0.9496	0.3990
2.3	**aiwizards_2**	**1**	**-2.8881**	**0.3469**
2.3	aiwizards_3	2	-2.9507	0.3436
2.3	aiwizards_1	3	-3.2537	0.3276

and discrete decision quality are related but distinct objectives, with the latter likely benefiting from task-specific threshold optimization.

Table 6
Official Hard-Hard leaderboard results.

Task	System	Rank	ICM-Hard	ICM-Hard Norm
2.1	**aiwizards_3**	**14**	**0.2673**	**0.6359**
2.1	aiwizards_1	18	0.2208	0.6123
2.1	aiwizards_2	19	0.2188	0.6112
2.2	**aiwizards_3**	**12**	**0.0020**	**0.5007**
2.2	aiwizards_1	13	-0.0353	0.4877
2.2	aiwizards_2	17	-0.1248	0.4566
2.3	**aiwizards_3**	**8**	**-0.3953**	**0.4180**
2.3	aiwizards_2	9	-0.4224	0.4124
2.3	aiwizards_1	13	-0.5316	0.3897

Three observations emerge: (1) the system is better matched to soft-label prediction, for which it was designed; (2) performance is strongest on the most challenging subtask (Task 2.3); (3) probabilistic outputs retain discriminative structure sufficient for non-trivial hard-label results even with naive thresholding.

7. Limitations

Our methodology is subject to three principal limitations. First, we rely on Gemini Embedding 2, a proprietary model, which constrains the full reproducibility of the proposed pipeline. However, the modular architecture of our system allows straightforward substitution with open-weight alternatives in future work. Second, the hard-label decoding procedure employs a fixed decision threshold of 0.5; consequently, the Hard-Hard results should not be construed as representing an upper bound on attainable performance. Third, our statistical analysis of physiological signal separability was confined to EEG data; gaze and heart-rate signals were not empirically evaluated and were excluded a priori, based on considerations of feature dimensionality and hypothesized task relevance, rather than on evidence derived from measured separability. Future work might include EEG, eye-tracking, and heart-rate data as well as their combination [23, 24, 25, 26].

8. Conclusion

We presented a hierarchical multi-task system for soft-label sexism detection in memes. By combining frozen Gemini Embedding 2 representations, compact gated MLP blocks, KL divergence against annotator distributions, and learned homoscedastic uncertainty weighting, our approach achieves strong performance while maintaining efficiency (3.5M trainable parameters). Our submissions ranked 1st on Task 2.3 and 4th on Tasks 2.1 and 2.2 on the official Soft-Soft leaderboards. Future work should investigate open-weight embedding backbones, task-specific hard-label calibration, and deeper integration of physiological signals through non-linear fusion architectures.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI-ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

Acknowledgments

Luca Tedeschini was supported by the project "IPCEI-CIS - Progetto Villanova" (Beneficiary: Tiscali Italia S.p.A., Prog. n. SA. 102519 – CUP B29J24000850005).

References

[1] M. Nadim, A. Fladmoe, Silencing women? gender and online harassment, Soc. Sci. Comput. Rev. 39 (2021) 245–258. URL: https://doi.org/10.1177/0894439319865518. doi:10.1177/0894439319865518.

[2] P. Glick, S. Fiske, An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality, American Psychologist 56 (2001) 109–118. doi:10.1037/0003-066X.56.2.109.

[3] J. Im, S. Schoenebeck, M. Iriarte, G. Grill, D. Wilkinson, A. Batool, R. Alharbi, A. Funwie, T. Gankhuu, E. Gilbert, M. Naseem, Women's perspectives on harm and justice after online harassment, Proc. ACM Hum.-Comput. Interact. 6 (2022). URL: https://doi.org/10.1145/3555775. doi:10.1145/3555775.

[4] 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.

[5] 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: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[6] L. Plaza, J. Carrillo-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), 2026.

[7] L. Plaza, J. Carrillo-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. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026. Working Notes, 2026.

[8] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, J. Artif. Int. Res. 72 (2022) 1385–1470. URL: https://doi.org/10.1613/jair.1.12752. doi:10.1613/jair.1.12752.

[9] 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. doi:10.1007/978-3-031-88720-8_65.

[10] 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.

[11] R. Cao, M. S. Hee, A. Kuek, W.-H. Chong, R. K.-W. Lee, J. Jiang, Pro-cap: Leveraging a frozen vision-language model for hateful meme detection, in: Proceedings of the 31st ACM International Conference on Multimedia, MM '23, 2023, p. 5244–5252. URL: http://dx.doi.org/10.1145/3581783.3612498. doi:10.1145/3581783.3612498.

[12] 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: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024, pp. 93–117.

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

[14] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11941–11952. doi:10.1109/ICCV51070.2023.01100.

[15] G. K. Kumar, K. Nandakumar, Hate-CLIPper: Multimodal hateful meme classification based on cross-modal interaction of CLIP features, in: L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao, J. Zhao (Eds.), Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 171–183. URL: https://aclanthology.org/2022.nlp4pi-1.20/. doi:10.18653/v1/2022.nlp4pi-1.20.

[16] L. Tian, J. R. Trippas, M.-A. Rizoiu, Mario at exist 2025: A simple gateway to effective multilingual sexism detection, 2025. URL: https://arxiv.org/abs/2507.10996. arXiv:2507.10996.

[17] K. R. White, J. Crites, Stephen L., J. H. Taylor, G. Corral, Wait, what? assessing stereotype incongruities using the n400 erp component, Social Cognitive and Affective Neuroscience 4 (2009) 191–198. URL: https://doi.org/10.1093/scan/nsp004. doi:10.1093/scan/nsp004. arXiv:https://academic.oup.com/scan/article-pdf/4/2/191/27105683/nsp004.pdf.

[18] M. Shanbhogue, Z. Li, S. Zhang, G. H. Ábrego, S.-C. Huang, A. Jain, D. Salz, S. Goenka, C. Hegde, J. Ma, F. Chen, J. Wu, T. Dabral, B. Samari, K. Poulet, D. Cer, K. Chen, P. Suganathan, H. Hui, J. Andonov, P. Schlattner, J. Han, I. Naim, W. Lowe, V. Pchelin, A. Yang, Y.-T. Chen, Z. Ding, G. Zhang, G. Heigold, Y. Chen, A. Reveillon, B. Mccloskey, W. Zhou, D. Kim, R. Meng, E. Wang, J. Zheng, H. Fede, Z. Yang, K. Mosley, B. Potetz, S. Dua, H. S. Vera, S. Gao, H. Zhang, A. Hess, H. Ying, A. Montes, K. Gill, M. Choi, S. Russo, A. Hauth, J. Lee, M. Boratko, M. Barnes, V. Rao, C. Musat, C. Allauzen, E. Variani, S. Kumar, T. Bagby, J. Jiao, Y. Gu, T. Li, A. Agrawal, R. Santana, D. Nath, S. Karukas, S. Han, L. Loher, A. Twu, N. Vyas, S. Bhai, F. P. Gomez, W. Zhang, C. Liu,

J. Yang, S. Qiu, S. Zhang, S. Kulkarni, S. Rothe, S. Nakamoto, R. Hoffmann, Z. Gleicher, Y. Sung, Q. Yin, T. Duerig, M. Seyedhosseini, Gemini embedding 2: A native multimodal embedding model from gemini, 2026. URL: https://arxiv.org/abs/2605.27295. arXiv:2605.27295.

[19] N. Shazeer, Glu variants improve transformer, 2020. URL: https://arxiv.org/abs/2002.05202. arXiv:2002.05202.

[20] S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics 22 (1951) 79–86.

[21] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7482–7491.

[22] 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.

[23] I. A. Gabaldón, P. Rosso, E. G. Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 9830–9840. doi:10.63317/2w2vaanu3pe7.

[24] Ö. Alacam, S. Hoeken, S. Zarrieß, Eyes don't lie: Subjective hate annotation and detection with gaze, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 187–205. URL: https://aclanthology.org/2024.emnlp-main.11/. doi:10.18653/v1/2024.emnlp-main.11.

[25] F. Oguz, A. Alkan, T. Schöler, Emotion detection from ecg signals with different learning algorithms and automated feature engineering, Signal, Image and Video Processing 17 (2023) 1–9. doi:10.1007/s11760-023-02606-y.

[26] Y. Zhang, Q. Li, S. Nahata, T. Jamal, S.-K. Cheng, G. Cauwenberghs, T.-P. Jung, Integrating large language model, eeg, and eye-tracking for word-level neural state classification in reading comprehension, IEEE Transactions on Neural Systems and Rehabilitation Engineering 32 (2024) 3465–3475. doi:10.1109/TNSRE.2024.3435460.