BioSentinel at EXIST 2026: Soft-Label Optimization with XLM-RoBERTa for Sexism Intent Classification in Memes

Chandru Munisamy[1,*], Karthikeya Raguveer[1] and Alapan Kuila[1]

[1]*Indian Institute of Information Technology, Design and Manufacturing, Kurnool, India*

Abstract

This paper describes the BioSentinel team's participation in EXIST 2026 Task 2.2: Source Intention in Memes, part of the CLEF 2026 evaluation campaign. The task requires classifying the communicative intent behind memes as DIRECT, JUDGEMENTAL, or NO (non-sexist), under a Learning with Disagreement (Le-Wi-Di) paradigm that mandates both hard-label and soft-label (probability distribution) predictions. We present a text-centric approach built on `xlm-roberta-base` (270M parameters) trained with a composite loss function combining KL divergence on soft annotator distributions and weighted cross-entropy on hard labels. On the official test set, the system achieved an ICM-Soft-Norm of 0.3229 and ICM-Norm of 0.3778, with a hard F1-score of 0.4236, ranking 40th (out of 118 submissions) in the soft-soft evaluation and 49th (out of 187 submissions) in the hard-hard evaluation. We provide an analysis of the dataset characteristics, exploratory larger-architecture runs, and the role of annotator disagreement in shaping model design for subjective NLP tasks. Ablation results show that KL loss improves soft-label metrics, while CE loss improves hard-label accuracy. We also report a separate validation-set temperature analysis.

Keywords

Sexism Detection, Learning with Disagreement, Soft Labels, XLM-RoBERTa, Memes, CLEF 2026

1. Introduction

Online sexism [1] manifests in diverse forms across social media, ranging from overtly hostile statements to subtly coded memes that require cultural and visual context for interpretation [2]. The EXIST (sEXism Identification in Social neTworks) shared task series has established itself as a key benchmark for evaluating automated sexism detection systems [3, 4]. EXIST 2026 Task 2.2, *Source Intention in Memes*, focuses on classifying the communicative intent of meme creators into three categories: NO (not sexist), DIRECT (explicit sexist intent), and JUDGEMENTAL (implicit or ironic sexist intent), as described in the official overview papers [5, 6].

A defining feature of this task is the adoption of the Learning with Disagreement (Le-Wi-Di) paradigm [7]. The complexity lies not just in identifying sexism, but in modeling the inherent subjectivity of human perception [8]. Annotators frequently disagree on what constitutes implicit sexism [9] (e.g., the JUDGEMENTAL class). Consequently, models are required to output not just a single hard label, but a soft probability distribution representing the anticipated variance across human annotators [10]. This setting requires objectives that preserve soft annotator distributions rather than collapsing them to one-hot targets [11].

Standard approaches to sexism detection typically collapse annotator disagreement [12] into a majority-vote hard label and employ standard cross-entropy loss over pre-trained transformer models [13]. While effective for objective tasks, this approach discards the valuable uncertainty signal present in subjective tasks and often leads to overconfident, poorly calibrated models.

We propose a text-centric approach built on `xlm-roberta-base` (270M parameters). Instead of relying purely on cross-entropy, we optimize a composite objective that jointly uses Kullback–Leibler (KL) divergence [14] on the soft annotator distributions and weighted cross-entropy on the hard labels.

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

*Corresponding author.

chandrumunisamy5@gmail.com (C. Munisamy); 123ad0026@iiitk.ac.in (K. Raguveer); alapan.cse@iiitk.ac.in (A. Kuila)

0009-0008-3625-210X (C. Munisamy)

We deliberately focused on OCR text rather than full multimodal fusion for two reasons. First, OCR text was the most consistently aligned modality available for our training pipeline, whereas image and physiological signals would have required additional missing-feature and synchronization strategies. Second, our main objective was to isolate whether soft-label optimization could improve disagreement-aware prediction under a lightweight, reproducible bilingual setup. We therefore treat the absence of image modeling as a known limitation and examine its implications in the error analysis (Section 4.4).

Under our shared experimental configuration, larger architectures were less stable than `xlm-roberta-base`. These exploratory runs reused settings selected for the base model, so they do not establish a general conclusion about model scale. Prioritizing KL divergence in the loss is motivated by the goal of aligning predictions with annotator distributions, as examined in the ablation study (Section 4.2).

Our system[1] produced soft-label predictions achieving an ICM-Soft-Norm of 0.3229, an ICM-Norm of 0.3778, and an F1 score of 0.4236 on the official test set, ranking 40th in the soft-soft evaluation and 49th in the hard-hard evaluation. The system outperformed the majority-class hard baseline (ICM-Norm = 0.2375, F1 score = 0.1676) and the prior-distribution soft baseline (ICM-Soft-Norm = 0.2835).

2. Problem Definition

2.1. Task Formulation

The core problem is to determine the intent behind a meme that may contain potentially sexist content. Because memes heavily rely on irony, juxtaposition, and cultural subtext, different people will interpret the same meme differently. The problem is twofold: predicting the most likely interpretation (hard classification) and predicting the spread of interpretations across a population (soft classification).

2.2. Formal Problem Definition

Although each benchmark instance contains an image I and extracted OCR text T in either English or Spanish, our system consumes OCR text only and learns a mapping function $f : T \to (y, P)$ where:

- $y \in \{\text{NO}, \text{DIRECT}, \text{JUDGEMENTAL}\}$ is the hard predicted class.
- $P = [p_{\text{NO}}, p_{\text{DIRECT}}, p_{\text{JUDGEMENTAL}}]$ is a valid probability distribution such that $\sum p_i = 1.0$, representing the soft label prediction.

The task enforces a label hierarchy: $\mathcal{H} = \{\text{YES} \to [\text{DIRECT}, \text{JUDGEMENTAL}],\ \text{NO} \to []\}$. Confusing NO with a sexist category incurs a higher penalty than confusing DIRECT with JUDGEMENTAL.

2.3. Dataset Description

The dataset comprises 3,984 training instances and 1,053 test instances, balanced across English and Spanish. Each instance was evaluated by 6 human annotators. The raw annotator votes were mapped to a soft gold probability distribution $p(c)$ for training. For example, an instance receiving 4 DIRECT, 1 JUDGEMENTAL, and 1 NO votes yields a soft label of $[0.167, 0.667, 0.167]$.

The official hard gold labels (assigned only if a class received strictly more than 2 votes out of 6) demonstrate severe class imbalance across the 3,149 resolvable instances: NO accounts for 1,801 instances (57.2%), DIRECT for 902 instances (28.6%), and JUDGEMENTAL for only 446 instances (14.2%). The remaining 835 instances were highly ambiguous, receiving no majority class and thus excluded from the hard evaluation. While the organizers also provided physiological sensor data (EEG, Eye Tracking, Heart Rate) from annotators, we ultimately excluded these due to a train-test feature mismatch (64 additional EEG features present in the test set but absent from training) and focused entirely on the OCR text modality.

[1]Implementation code is available at https://github.com/kirito-meta/EXIST2026_PROJECT.

3. Proposed Method

3.1. Architecture Overview

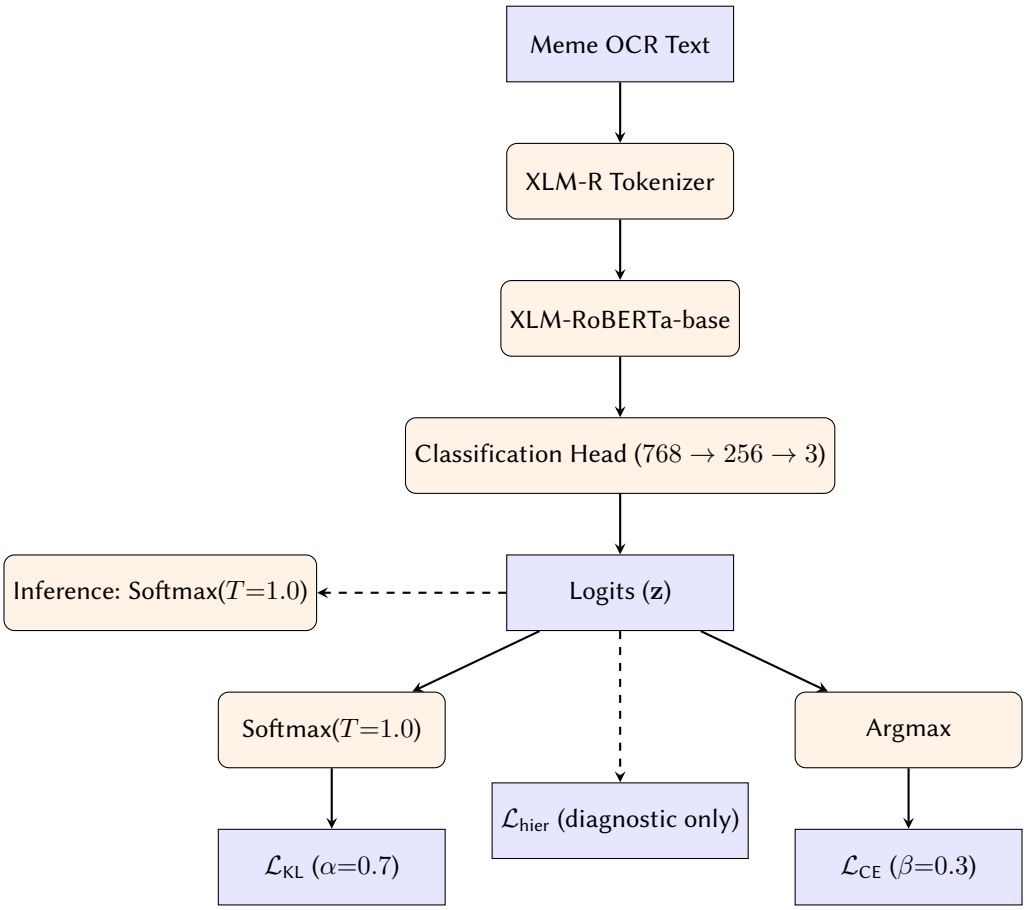

Figure 1: Architecture diagram. During training, $\mathcal{L}_{\text{train}} = \alpha \cdot \mathcal{L}_{\text{KL}} + \beta \cdot \mathcal{L}_{\text{CE}}$ updates weights. $\mathcal{L}_{\text{hier}}$ is logged as a non-differentiable diagnostic. At inference, probabilities are obtained by raw softmax (T=1.0).

Our proposed methodology, illustrated in Figure 1, relies on a text-only pipeline using the `xlm-roberta-base` transformer model. We chose this model for its strong cross-lingual transfer capabilities, crucial for the bilingual nature of the dataset.

3.2. Methodology Details

Classification Head. The first special-token representation (768-dimensional) from XLM-RoBERTa is passed through a multi-layer perceptron with dropout to produce 3-dimensional logits corresponding to the target classes.

Composite Loss Function. To address the Le-Wi-Di paradigm, the gradient-based training objective is:

$$\mathcal{L}_{\text{train}} = \alpha \cdot \mathcal{L}_{\text{KL}} + \beta \cdot \mathcal{L}_{\text{CE}}, \tag{1}$$

where $\alpha = 0.7$ and $\beta = 0.3$.

1. **KL Divergence ($\mathcal{L}_{\text{KL}}$).** The primary objective minimizes the Kullback–Leibler divergence between the predicted distribution and the soft annotator distribution. This directly aligns the gradient-based training signal with the soft-label objective.

2. **Weighted Cross-Entropy ($\mathcal{L}_{\text{CE}}$).** For instances with official hard labels, we use the fixed normalized class weights $[0.482, 0.572, 1.946]$ for NO, DIRECT, and JUDGEMENTAL, respectively, as implemented in the training script. These fixed weights increase the contribution of the under-represented JUDGEMENTAL class during training.
3. **Hierarchy Boundary Diagnostic ($\mathcal{L}_{\text{hier}}$).** We also record a discrete boundary-error quantity when the hard prediction crosses the NO/sexist boundary. It is computed from argmax predictions in a no-gradient block, so it does not update model parameters and was not used to select the reported checkpoint. The reported Epoch 4 checkpoint was selected solely by minimum validation KL divergence.

Temperature Analysis. Official test predictions use raw softmax probabilities ($T = 1.0$). We separately evaluated post-hoc temperature scaling on the held-out validation split. As shown in Section 4.3, $T = 0.7$ improved validation ICM-Soft-Norm and reduced the reported validation ECE; this analysis was not used for the official test run.

3.3. Model Training Parameters

The system was implemented in PyTorch and trained on NVIDIA A100 GPUs via Google Colab Pro with the following hyperparameters:

- **Base Model**: `FacebookAI/xlm-roberta-base` (270M parameters)
- **Hidden Layers (Head)**: $768 \to 256 \to 3$ (with GELU activation and 0.1 Dropout)
- **Epochs**: 6 (Epoch 4 checkpoint selected by minimum validation KL divergence)
- **Optimizer**: AdamW (weight decay = 0.01)
- **Learning Rate**: 2×10^{-5} with linear warmup over 10% of steps
- **Batch Size**: 32
- **Max Sequence Length**: 128 tokens

4. Experiments and Results

4.1. Validation and Official Test Results

Our model was validated on a stratified 15% hold-out validation set (598 instances), split by hard label and language to preserve the joint distribution of class proportions and bilingual balance. Table 1 reports the validation analysis, including the post-hoc T=0.7 setting discussed in Section 4.3.

Table 1
Validation set evaluation using PyEvALL.

Hard-Hard Evaluation		Soft-Soft Evaluation	
ICM	−0.482	ICM-Soft	−1.670
ICM-Norm	0.333	ICM-Soft-Norm	0.326
F1 (macro)	0.356	Cross-Entropy	1.468
F1-NO	0.579	Val KL Divergence	0.352
F1-DIRECT	0.418		
F1-JUDGEMENTAL	0.071		

The system's performance on the official EXIST 2026 test set is presented in Table 2.

4.2. Ablation Study

To understand the contribution of the gradient-based objectives, we conducted a single-split ablation study on the validation set. Table 3 compares the loss configurations with the post-hoc temperature analysis.

Table 2
Official test set results.

System	ICM-Soft-Norm	ICM-Norm	F1 (macro)
Majority-class baseline	—	0.2375	0.1676
Prior-distribution baseline	0.2835	—	—
BioSentinel (Ours)	**0.3229**	**0.3778**	**0.4236**

Table 3
Single-split ablation study on the validation set. All configurations use the same `xlm-roberta-base` architecture and hyperparameters. The final row reports a validation-only post-hoc analysis.

System configuration	ICM-Soft-Norm	ICM-Norm	F1 (macro)
CE only (β=1.0)	0.142	0.285	0.348
KL only (α=1.0)	0.312	0.310	0.294
KL + CE	0.315	0.321	0.354
Selected KL + CE checkpoint, raw softmax (T=1.0)	0.317	0.332	0.356
Selected checkpoint + post-hoc T=0.7 (validation only)	0.326	0.333	0.356

The ablation results support several observations. First, KL loss leads to better soft-label metrics (ICM-Soft-Norm of 0.312 vs. 0.142 under CE-only), whereas CE loss improves hard-label macro-F1 (0.348 vs. 0.294 under KL-only). Combining the objectives gives stronger joint development performance. For the selected checkpoint, post-hoc T=0.7 raises validation ICM-Soft-Norm from 0.317 to 0.326 while leaving the hard prediction unchanged.

4.3. Validation Temperature Analysis

Table 4 presents the temperature sweep on the validation set, showing ICM-Soft-Norm and Expected Calibration Error (ECE) at each temperature value.

Table 4
Validation-set post-hoc temperature sweep.

Temperature	ICM-Soft-Norm	ECE
$T = 1.0$ (no scaling)	0.317	0.185
$T = 0.9$	0.321	0.159
$T = 0.8$	0.324	0.138
$T = 0.7$ (selected)	0.326	0.118

Across the tested validation values, ICM-Soft-Norm increases from 0.317 at T=1.0 to 0.326 at T=0.7, while the reported validation ECE decreases from 0.185 to 0.118. This shows that sharpening improved the held-out validation measurements under this setup. These figures are from the validation split; official predictions follow the raw-softmax procedure described in Section 3.2. Temperature scaling remains a standard post-hoc adjustment [15].

4.4. Error Analysis

At the selected Epoch 4 checkpoint, 221 of the 598 validation instances had a raw hard prediction that differed from the official hard label. Table 5 summarizes recurring error mechanisms qualitatively; Table 6 provides three verified examples drawn from these errors.

The qualitative taxonomy highlights irony and omitted visual context as recurring challenges for an OCR-only system. Because no controlled multimodal comparison was conducted, these observations are descriptive rather than causal. OCR noise and cultural references are additional plausible sources of error.

Table 5
Qualitative taxonomy used to interpret validation-set errors. These are descriptive mechanisms, not frequency estimates.

Error Mechanism	Description
Sarcasm/Irony missed	OCR text may appear neutral or literal while framing changes the intended meaning.
Visual context missing	OCR text alone may be insufficient when a caricature, image, or visual juxtaposition carries the relevant signal.
OCR noise	Corrupted or incomplete OCR can omit words that determine the intended interpretation.
Cultural reference	Region-specific slang or cultural tropes may not be represented adequately by the text encoder.

To make the error analysis concrete, Table 6 reports three genuine misclassified validation instances exported from the selected Epoch 4 checkpoint: one for each gold class. The soft distributions are ordered as [NO, DIRECT, JUDGEMENTAL] and use raw softmax probabilities from the selected checkpoint. We provide OCR excerpts rather than reproducing the source images.

Table 6
Concrete validation-set misclassifications. Each row is a genuine held-out instance; OCR excerpts are shortened only for readability. Soft distributions are ordered as [NO, DIRECT, JUDGEMENTAL].

ID / lang.	Gold / soft dist.	Pred. / soft dist.	OCR excerpt and observed error
110627 (es)	NO; [1.000, 0, 0]	DIRECT; [0.221, 0.442, 0.336]	OCR (English translation): "When you do not want to get up but remember that you are an empowered and independent woman…" The model assigns the highest probability to DIRECT despite an all-NO gold distribution.
210599 (en)	DIRECT; [0.167, 0.667, 0.167]	NO; [0.594, 0.259, 0.147]	OCR: "I crave a relationship with traditional gender roles." The model predicts NO, although the majority of annotator votes label its source intention DIRECT.
210534 (en)	JUDGEMENTAL; [0.167, 0.167, 0.667]	NO; [0.628, 0.197, 0.175]	OCR: "Sitting at your family table when the biggest fish fry goes to your father and the smallest one sits on your mother's plate." The text-only system predicts NO, while the majority gold distribution is JUDGEMENTAL.

4.5. Analysis and Discussion

Soft-Label Alignment and Hard-Label Accuracy. The model's ICM-Soft-Norm of 0.3229 improves on the prior-distribution soft baseline (0.2835) by 13.9% relative, whereas its hard macro-F1 is 0.4236. This pattern suggests that the model captures distributional signal not fully reflected by a single hard decision. The ablation study (Table 3) supports this interpretation: removing the KL term (CE-only) reduces ICM-Soft-Norm to 0.142, while CE remains important for hard-label accuracy.

The Challenge of the JUDGEMENTAL Class. The system performed poorly on the JUDGEMENTAL class (F1 = 0.071). Ironic and judgemental sexism can rely on the juxtaposition of text and image, while OCR text alone lacks contextual cues such as facial expressions and visual tropes. The qualitative taxonomy and the verified examples are consistent with missing visual context as a plausible contributor, but a controlled multimodal comparison would be required for causal attribution.

JUDGEMENTAL Cold-Start Behavior. Notably, the model produced zero JUDGEMENTAL predictions during the first two training epochs, with only 22 such predictions emerging at Epoch 3. At the selected Epoch 4 checkpoint, 43 out of 598 validation instances received a JUDGEMENTAL prediction. This cold-start behavior suggests that the model initially learns to distinguish only the NO/DIRECT boundary before gradually recognizing the more subtle JUDGEMENTAL patterns, indicating that longer training or dedicated JUDGEMENTAL augmentation strategies may be beneficial.

Architecture Scale and Stability. During development, we experimented with larger models (Gemma 2 2B, DeBERTa-v3-large, XLM-RoBERTa-large). Under our shared experimental configuration, these architectures were less stable than `xlm-roberta-base`: Gemma 2 2B exhibited divergent losses beyond epoch 2, DeBERTa-v3-large converged to majority-class predictions, and XLM-RoBERTa-large produced increasingly extreme probability distributions. Because these runs reused hyperparameters selected for the base model, they are exploratory rather than controlled architecture comparisons. A dedicated hyperparameter search would be required before drawing a general conclusion about model scale.

5. Related Work

Sexism Detection in Social Media. Early approaches to online sexism detection relied on feature-engineered classifiers [16], while recent methods leverage pre-trained transformer models fine-tuned on domain-specific data [17, 18]. Multimodal meme classification has received increasing attention following the Hateful Memes Challenge [19], which underscored the need to reason jointly about text and images. Our work instead focuses on optimizing soft-label predictions from text alone and shows improvement over the reported soft baseline.

Learning with Disagreement. The Le-Wi-Di paradigm [7] advocates preserving annotator disagreement rather than collapsing it into majority labels. Soft-label training can preserve information about uncertainty in human labels [20]. The EXIST task series has progressively adopted this paradigm, making it a core evaluation criterion. Our composite loss design is directly informed by this paradigm, using KL divergence as the primary training signal to preserve annotator distribution information.

Model Calibration. Neural networks are known to be poorly calibrated [15]. Temperature scaling, where logits are divided by a scalar T before softmax, is a standard post-hoc adjustment. Our validation analysis found lower reported ECE and higher ICM-Soft-Norm at $T{=}0.7$ (Table 4).

Evaluation with ICM. The Information Contrast Model [21] provides a hierarchy-aware evaluation metric that penalizes cross-level confusions more severely than within-level errors. It remains useful for interpreting cross-boundary mistakes in the hard-label evaluation.

6. Limitations and Ethics

Our study has several limitations. First, the text-only approach inherently cannot capture visual-textual interactions that can be central to meme interpretation, which is especially relevant to the low JUDGEMENTAL-class F1. The qualitative error taxonomy identifies irony and omitted visual context as recurring patterns, but does not establish causal contribution without a controlled multimodal comparison. Second, with only 3,984 training instances, the generalizability of our findings to broader meme datasets remains uncertain. Third, the ablation study was conducted on a single validation split; multi-seed experiments would provide more robust estimates of each component's contribution. Fourth, sexism detection systems should not be deployed for automated content moderation without human oversight, as false positives could lead to unjust censorship and false negatives could allow harmful

content to propagate. Finally, the training data may contain cultural and linguistic biases that are reflected in the model's predictions, particularly given the bilingual but limited geographic scope of the annotator pool.

7. Conclusion

In this shared-task setting, the BioSentinel `xlm-roberta-base` system combines KL divergence with weighted cross-entropy to model annotator disagreement from OCR text. On the official test set, it achieves an ICM-Soft-Norm of 0.3229.

Our ablation study shows that KL loss improves the soft-label metric, while CE loss improves hard-label accuracy; combining them (α=0.7, β=0.3) balances both objectives. The post-hoc validation analysis at T=0.7 improves ICM-Soft-Norm by 0.009 and lowers the reported validation ECE. Temperature scaling was assessed only on validation data; official predictions use raw softmax. The hierarchy boundary error is used only for diagnostic analysis.

Our exploratory runs suggest that stability under noisy soft labels warrants further study. Future work should integrate visual features to address the weakness on JUDGEMENTAL sexism, evaluate components across multiple random seeds, and investigate ways to align physiological signals with textual and visual inputs.

Acknowledgments

We thank the EXIST 2026 organizers for providing the dataset, evaluation framework, and the opportunity to participate. We also acknowledge Google Colab for providing the computational infrastructure used in our experiments.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (GPT-4o) and Gemini for grammar and spelling checks, text restructuring, and code debugging. The authors reviewed and edited all generated suggestions and take full responsibility for the final content.

References

[1] 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.

[2] H. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2193–2210.

[3] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[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. 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.

[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 (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.

[7] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.
[8] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110.
[9] L. A. Rudman, S. E. Kilianski, Implicit and explicit attitudes toward female authority, Personality and social psychology bulletin 26 (2000) 1315–1328.
[10] X. Geng, Label distribution learning, IEEE Transactions on Knowledge and Data Engineering 28 (2016) 1734–1748.
[11] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning from noisy labels with deep neural networks: A survey, IEEE transactions on neural networks and learning systems 34 (2022) 8135–8153.
[12] A. Jiang, N. Vitsakis, T. Dinkar, G. Abercrombie, I. Konstas, Re-examining sexism and misogyny classification with annotator attitudes, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 15103–15125.
[13] M. Mozafari, R. Farahbakhsh, N. Crespi, Hate speech detection and racial bias mitigation in social media based on bert model, PloS one 15 (2020) e0237861.
[14] S. Kullback, R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics 22 (1951) 79–86. doi:10.1214/aoms/1177729694.
[15] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: International conference on machine learning, PMLR, 2017, pp. 1321–1330.
[16] Z. Talat, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: 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), 2019, pp. 4171–4186.
[18] 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: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 8440–8451.
[19] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in neural information processing systems 33 (2020) 2611–2624.
[20] J. C. Peterson, R. M. Battleday, T. L. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9617–9626.
[21] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers), 2022, pp. 5809–5819.