SafetyNLP at EXIST 2026: Multimodal Sexism Detection in Memes Using LLM-Driven OCR Enhancement and LeWiDi Target Modeling

Notebook for the EXIST 2026 Lab at CLEF 2026

Bogdan Tamba[1,*], Alexandru Petrescu[1,2] and Ciprian-Octavian Truică[1,2]

[1]National University of Science and Technology Politehnica Bucharest, Splaiul Independenței 313, București 060042, Romania
[2]Academy of Romanian Scientists, 3 Ilfov, Bucharest, Romania

Abstract

In this work, we compare the text-only and multimodal architectures developed for the CLEF 2026 EXIST (sEXism Identification in Social neTworks) Laboratory. Our proposed methodology targets three tasks regarding social media memes: 1) sexism identification (Subtask 2.1), 2) source intention detection (Subtask 2.2), and 3) multi-label sexism categorization (Subtask 2.3). We build an OCR preprocessing pipeline that uses a Large Language Model and NLTK-based filtering to reconstruct distorted OCR (Optical Character Recognition) text while keeping emojis, hashtags, and mentions intact. We perform extensive experiments across three random seeds to test this text enhancement using text-only transformers (XLM-R, Twitter-XLM-R, mDeBERTa) and multimodal visual-language models (CLIP, SigLIP). For cross-modal integration, we test both late fusion and cross-attention mechanisms. We employ a meta-ensemble strategy to stabilize predictions for Subtasks 2.1 and 2.2, and design three soft-label variants (m_1, m_2, m_3) to account for label ambiguity within the Learning with Disagreements (LeWiDi) framework. OCR preprocessing quality proved surprisingly consequential across all three subtasks.

Keywords

Sexism Detection, Multimodal Classification, OCR Enhancement

1. Introduction

The rapid growth of harmful content on social media remains largely unsolved. While detection efforts have traditionally focused on text, the modern landscape of digital communication is multimodal. Memes spread ideological and objectifying content rapidly. Because their semantics rely on the complex interplay between visual elements and overlaid text, traditional text-only moderation systems frequently fail.

The CLEF 2026 EXIST Laboratory responds to this by challenging systems to identify and characterize sexism in multimedia formats under a Learning with Disagreements (LeWiDi) paradigm. Within this evaluation campaign, our work focuses on Task 2 (Memes), addressing three main objectives: 1) deciding whether a meme is sexist (Subtask 2.1), 2) categorizing the author's intention as direct or judgmental (Subtask 2.2), and 3) identifying the specific facets of sexism attacked in a multi-label setting (Subtask 2.3).

Extracting text from memes is a difficult problem due to poor OCR (Optical Character Recognition) quality. We try to solve this by introducing an LLM-driven OCR preprocessing pipeline that reconstructs noisy text without touching emojis, hashtags, or mentions. With cleaner text as input, we then compare a suite of text-only models against multimodal architectures (CLIP and SigLIP). To address the soft-label evaluation inherent to the LeWiDi framework, we implement a meta-ensemble approach for the single-label tasks, alongside three strategies for generating multi-label targets in Subtask 2.3.

CLEF 2026 Working Notes, September 21–24, 2026, Jena, Germany
*Corresponding author.
bogdantamba55@gmail.com (B. Tamba); alex.petrescu@upb.ro (A. Petrescu); ciprian.truica@upb.ro (C. Truică)
https://alexpetrescu.net/ (A. Petrescu); https://sites.google.com/view/ciprian-octavian-truica (C. Truică)
0009-0006-1750-766X (B. Tamba); 0000-0002-7731-2403 (A. Petrescu); 0000-0001-7292-4462 (C. Truică)


2. Related Work

Harmful content detection and mitigation have been tackled in the current literature by different communities. From a natural language point of view, methods range from word embeddings [1, 2, 3], to transformer-based embeddings [4, 5, 6, 7, 8, 9], a mixture of transformers [5, 10, 11], and Large Language Models [12]. Fusion models that integrate textual and network diffusion models have proved robustness in harmful content detection [13, 14]. Immunization strategies [15, 16, 17] have also been proposed to stop the spread of this content online. Real-time systems for harmful content detection and mitigation systems have also been developed to address this challenge [18, 19, 20].

2.1. Sexism Identification and the LeWiDi Paradigm

The task of sexism identification in social media has shifted considerably across CLEF iterations. Previous editions of the EXIST lab established the importance of the Learning with Disagreements (LeWiDi) paradigm [21]. This framework acknowledges that annotators disagree and that aggregating annotator variance into soft labels provides a richer, more realistic training signal than forced majority consensus. Recent efforts within the lab have moved toward multimodal analysis to tackle image-based and video-based content. For instance, Italiani et al. [22] explore inter-meme graph reasoning for sexism and misogyny detection, showing that text alone often can't resolve what a meme actually means. Similarly, Arcos et al. [23] investigate human-centered multimodal fusion, incorporating physiological signals to better understand how users process sexist memes.

2.2. Text-Only Representations for Social Media

Despite the multimodal nature of memes, the textual modality often carries the explicit ideological payload. Consequently, robust multilingual language models are foundational to our approach. XLM-RoBERTa (XLM-R) [24] pretrained on massive filtered CommonCrawl data, has demonstrated strong performance across a wide range of cross-lingual tasks. However, standard language models typically struggle with the unique language of social media. Twitter-XLM-R [25] is built by continuing pretraining on tweet corpora, making it better at hashtags and slang detection. In addition, to leverage the architectural advances in attention mechanisms, we incorporate mDeBERTa [26], which uses disentangled attention mechanisms and enhanced mask decoder training. This consistently tops benchmarks on tasks requiring nuanced understanding, such as hate speech and sexism detection.

2.3. Multimodal Vision-Language Architectures

Because the semantics of a meme rely heavily on the juxtaposition of text and image, text-only models frequently fail to flag content where the sexism is visually implied (e.g., objectification). While Contrastive Language-Image Pretraining (CLIP) [27] is highly effective, recent advancements like SigLIP (Sigmoid Loss for Language Image Pre-Training) [28] replace the standard softmax loss with a pairwise sigmoid loss. This allows SigLIP to handle noisy, densely packed image-text pairs better, which matters for memes, where image-text pairs are rarely clean. In our work, we contrast simple late-fusion strategies, which independently pool and concatenate these embeddings with deeper cross-attention mechanisms, where image patches attend directly to text tokens [29].

2.4. OCR Enhancement in Noisy Domains

OCR quality is the main bottleneck in any computer vision pipeline. OCR artifacts, e.g., missing spaces, random character substitutions, and structural noise, severely degrade the performance of downstream NLP models. Traditional rule-based spelling correctors often fail on social media text because they aggressively normalize intended slang or strip away emojis and mentions. Recently, instruction-tuned Large Language Models (LLMs) handle this better since they reconstruct garbled tokens using context

rather than dictionary lookup [30]. We use a quantized Qwen LLM, constrained by NLTK-based heuristics, to ensure that emojis, hashtags, and mentions are never touched.

3. Experimental Setup and Methodology

Our system tackles two problems: 1) noisy meme text and 2) subjective sexism annotation. It has three stages: 1) an initial OCR enhancement phase, 2) the extraction of features via text-only and multimodal base models, and 3) task-specific classification strategies (meta-ensembling for Subtasks 2.1 and 2.2, and varied soft-target generation for Subtask 2.3). The overall architecture is illustrated in Figure 1.

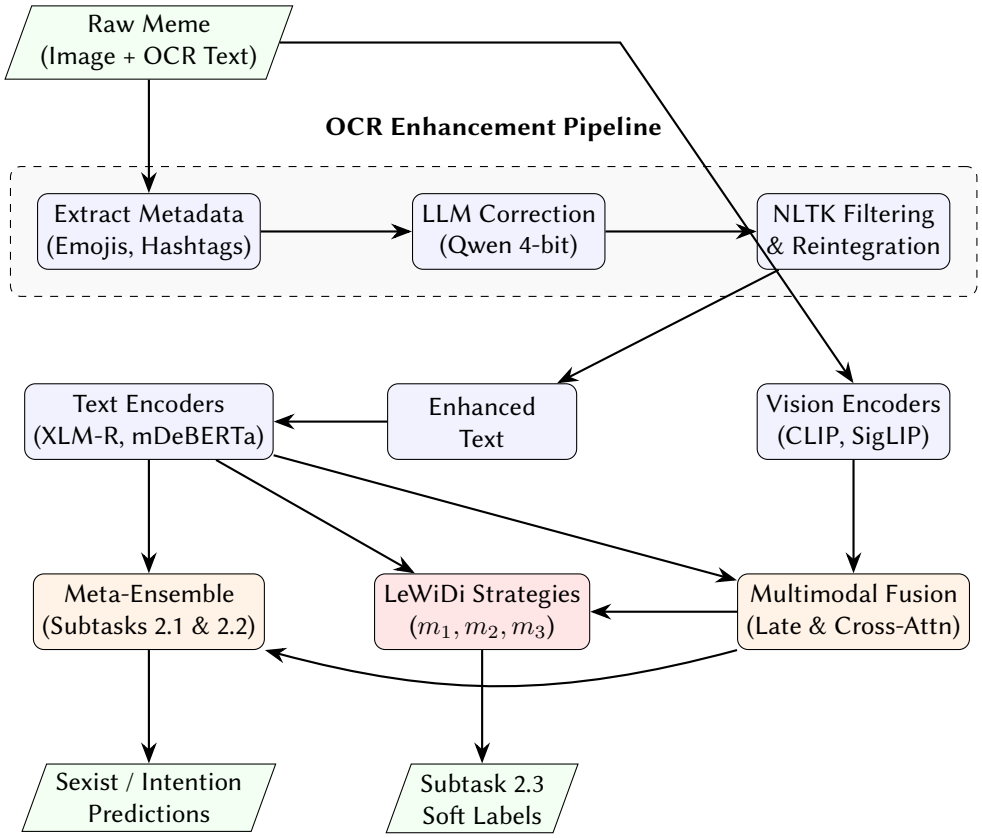


Figure 1: Architecture of our proposed system, illustrating the LLM-driven OCR enhancement pipeline, the extraction of textual and visual features, and the strategies applied for multimodal classification.

3.1. Dataset and OCR Preprocessing

We use the EXIST 2026 Memes Dataset, which contains over 5,000 labeled memes in English and Spanish [31, 32]. Because raw OCR often fails on memes due to varied typography and background clutter, we develop an enhancement pipeline. The pipeline first extracts and stores all URLs, @mentions, hashtags, and emojis via regular expressions, replacing them with sentinel placeholders. The remaining text is passed to a 4-bit quantized Qwen3.5-9B instruction-tuned model with a constrained system prompt that permits only character-level OCR repairs: restoring accented letters lost as "?", splitting run-together words, and reverting common glyph confusions (e.g., "rn"→"m"). The prompt explicitly forbids translation, rephrasing, or reordering, and includes few-shot examples covering English, Spanish, and mixed-language inputs. An NLTK-based consonant-ratio heuristic then acts as a safety filter, discarding tokens that remain unpronounceable after correction, before the preserved metadata is reintegrated at its original positions. The full system prompt is provided in Appendix A.

3.2. Base Models and Fusion Strategies

To capture the varied semantics of the dataset, we fine-tuned multiple models over three random seeds:

- **Text-Only:** XLM-R, Twitter-XLM-R, and mDeBERTa [26].
- **Multimodal:** Vision-language encoders using CLIP and SigLIP [28]. We experimented with two integration methods:
  - *Late Fusion:* Text and image modalities are encoded independently, then concatenated before classification.
  - *Cross-Attention:* Image patches attend to text tokens directly.

3.3. Meta-Ensemble (Subtasks 2.1 and 2.2)

For Subtask 2.1 (Sexism Identification) and Subtask 2.2 (Source Intention), we employ a meta-learning ensemble. A meta-classifier was trained over the probabilistic outputs of both the text-only and multimodal base models. In effect, it learns when to trust the text model and when to trust the visual one, stabilizing the final predictions.

3.4. Label Generation for Subtask 2.3

Subtask 2.3 is a multi-label problem that categorizes sexism into five non-mutually exclusive classes. Because the gold standard incorporates human disagreement, annotators may select multiple categories, indicate the absence of sexism ("-"), or mark the instance as "UNKNOWN". To evaluate our architectures under both traditional and LeWiDi paradigms, we generate both hard and soft training targets.

3.4.1. Hard-Label Derivation

For the hard-hard evaluation context, a category passes if it is selected by more than one annotator (> 1 vote), aligning with the official EXIST 2026 thresholding rules [31, 32]. The "UNKNOWN" label is discarded, and the "-" label explicitly counts toward the "NO" class.

3.4.2. Soft-Label Generation (LeWiDi)

For the soft-soft evaluation context, the system outputs a probability distribution. We devise three strategies to compute the target probability (y_c) for each category c given N total annotators, where n_c denotes the number of annotators who voted for category c, $n_{unknown}$ the number of "UNKNOWN" votes, and n_{no} the number of non-sexist votes.

Strategy m_1 (Dilution): "UNKNOWN" is kept in the denominator, treating it as a neutral vote that dilutes overall category certainty (Equation (1)).

$$y_c^{(m_1)} = \frac{n_c}{N} \tag{1}$$

Strategy m_2 (Explicit Votes Only): "UNKNOWN" is discarded from the denominator entirely, boosting the resulting probabilities of the explicit votes (Equation (2)).

$$y_c^{(m_2)} = \frac{n_c}{N - n_{unknown}} \tag{2}$$

Strategy m_3 (Conservative Scaling): Starting from m_2, we down-weight the probability mass for the five active sexism classes by half, while the non-sexist target (y_{NO}) remains unscaled (Equation (3)).

$$y_c^{(m_3)} = 0.5 \times \frac{n_c}{N - n_{unknown}} \quad \text{and} \quad y_{NO}^{(m_3)} = \frac{n_{no}}{N - n_{unknown}} \tag{3}$$

4. Results and Discussion

All metrics are computed using the official PyEvALL evaluation framework [31, 32]. Following the LeWiDi paradigm, we report both the traditional Hard-Hard metrics, i.e., primarily the Information Contrast Measure (ICM) and its normalized variant (ICMNorm), as well as the Soft-Soft metrics (ICM-Soft and ICMSoftNorm) to capture the models' ability to replicate human annotator distributions.

To ensure the statistical reliability of our findings, base models were trained and evaluated across three random seeds (42, 123, 7), with variance visualized in the accompanying figures.

4.1. Subtask 2.1: Sexism Identification

Tables 1 and 2 present the numerical results for binary sexism identification. We compare a strong text-only baseline trained on raw OCR text (`txt_twitterxlmr_raw`), a multimodal architecture trained on enhanced OCR text (`mm_siglip_late_clean`), and our proposed meta-ensemble approach.

Table 1
Performance on Subtask 2.1 (Sexism Identification) using Soft labels. Best results are in bold.

Model Configuration	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank
Text-Only (Twitter-XLM-R, Raw)	-0.3696	0.4406	0.9647	42/141
Multimodal (SigLIP Late, Clean)	-0.3292	0.4471	**0.9253**	38/141
Meta-Ensemble (Combined)	**-0.1295**	**0.4792**	0.9281	**18/141**

Table 2
Performance on Subtask 2.1 (Sexism Identification) using Hard labels. Best results are in bold.

Model Configuration	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Text-Only (Twitter-XLM-R, Raw)	-0.0179	0.4909	0.7003	109/214
Multimodal (SigLIP Late, Clean)	0.0879	0.5447	0.7122	57/214
Meta-Ensemble (Combined)	**0.1298**	**0.5660**	**0.7310**	**36/214**

The meta-ensemble has the overall best performance on every metric. While the text-only model performed strongly on its own, likely because explicit hate speech in memes is heavily driven by the textual modality, combining the probabilistic outputs of both text-only and multimodal models via the meta-classifier yielded the highest scores across all metrics. The Meta-Ensemble achieved an ICM-Soft score of -0.1295, suggesting the ensemble better tracks where annotators disagreed.

4.2. Subtask 2.2: Source Intention

Subtask 2.2 requires the system to classify the intention of the meme's author as either *Direct* or *Judgemental*. Because intention is heavily reliant on linguistic tone, irony, and phrasing, we evaluate our architectures using the OCR-enhanced (clean) text to minimize noise. Tables 3 and 4 show that the text-only Twitter-XLM-R model significantly outperforms the multimodal SigLIP architecture on this specific subtask. This supports the hypothesis that visual features, while useful for identifying the presence of sexism, are less informative when distinguishing between a direct attack and a judgmental condemnation. Once again, the Meta-Ensemble smooths the prediction variance, achieving the best overall ICM-Hard Norm (0.3861).

4.3. Overall Model Stability and the Impact of OCR Enhancement

To validate the stability of our models and the efficiency of our preprocessing pipeline, we visualize the performance variance across our multiple random seeds.

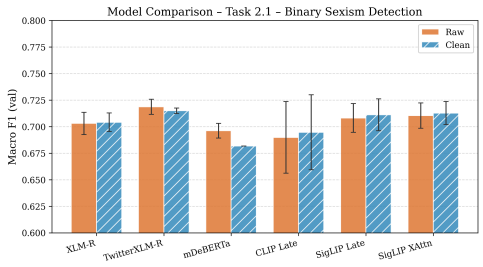
Table 3
Performance on Subtask 2.2 (Source Intention) using Soft labels. Best results are in bold.

Model Configuration	ICM-Soft	ICM-Soft Norm	Cross Entropy	Rank
Text-Only (Twitter-XLM-R, Clean)	-1.4973	0.3408	1.5530	28/114
Multimodal (SigLIP Late, Clean)	-1.7407	0.3149	**1.5148**	48/114
Meta-Ensemble (Combined)	**-1.4512**	**0.3457**	1.6665	**27/114**

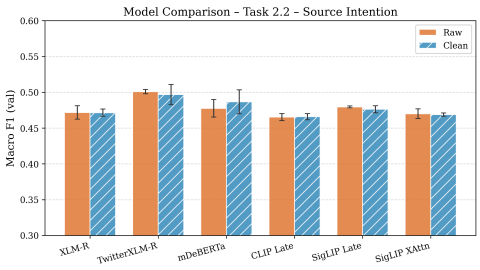
Table 4
Performance on Subtask 2.2 (Source Intention) using Hard labels. Best results are in bold.

Model Configuration	ICM-Hard	ICM-Hard Norm	Macro F1	Rank
Text-Only (Twitter-XLM-R, Raw)	-0.3634	0.3737	0.4508	55/183
Multimodal (SigLIP Late, Clean)	-0.4490	0.3439	0.4585	87/183
Meta-Ensemble (Combined)	**-0.3277**	**0.3861**	**0.4290**	**44/183**

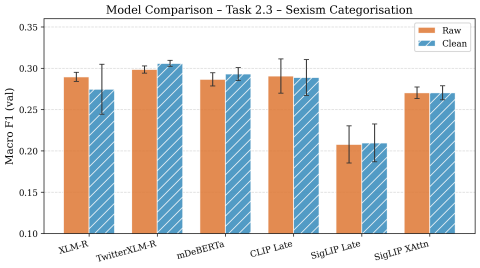
Figure 2 shows that the meta-ensemble works well for Subtasks 2.1 and 2.2, consistently reducing the variance (indicated by tighter error bars) compared to the standalone architectures.



(a) Task 2.1



(b) Task 2.2

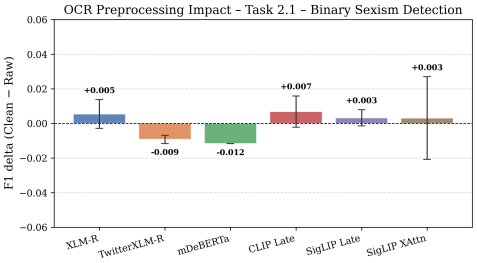


(c) Task 2.3

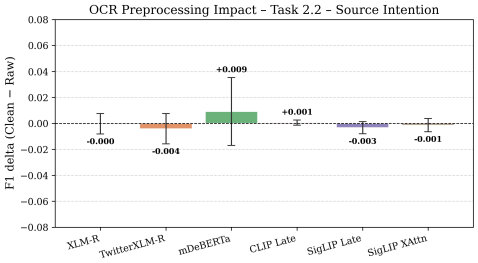
Figure 2: Performance comparison of text-only, multimodal, and meta-ensemble architectures across tasks. Error bars indicate variance across three random seeds.

Raw OCR noise introduces lexical artifacts that obfuscate the semantic content of the meme text. Figure 3 serves as a direct ablation of the enhancement pipeline: each bar reports the validation F1 delta between models trained on enhanced versus raw OCR text, with architecture, hyperparameters, and random seeds held fixed.

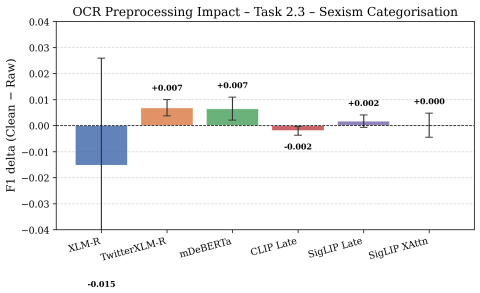
Across the board, models trained on the enhanced text yield higher mean scores while exhibiting reduced variance. The preservation of emojis and hashtags makes the biggest difference in Subtask 2.2, as these tokens frequently encode the sarcastic tone required to identify judgmental intent.



(a) Task 2.1



(b) Task 2.2



(c) Task 2.3

Figure 3: Comparison of model performance using Raw OCR versus Enhanced (Clean) OCR text. Enhanced text yields higher mean scores and lower inter-seed variance.

4.4. LeWiDi Label Generation Strategies (Task 2.3)

Subtask 2.3 presents a complex multi-label classification challenge evaluated under the Soft-Soft LeWiDi context. We contrast our three target generation strategies (m_1, m_2, and m_3) in Tables 5 and 6. The baseline results reveal the extreme difficulty of predicting exact multi-label probability distributions. Strategy m_2, which discards "UNKNOWN" annotations to boost the signal of explicit votes, obtained the best ICM-Soft score, although the margins between strategies are negligible, as we discuss next.

The near-identical ICM-Soft scores across all three strategies (differences below 10^{-2}) warrant a critical reading. Inspecting the validation predictions of the submitted runs reveals that the multimodal classifier suffers from severe output collapse: for every class, the per-example sigmoid outputs are confined to a narrow band around 0.5 (per-class means between 0.47 and 0.50, standard deviations below 0.04), regardless of the input and regardless of the true class frequency, which ranges from 0.09 (*Misogyny*) to 0.42 (*NO*). The mean pairwise L_1 distance between predicted vectors of random validation pairs is 0.18, compared to 1.36 between the corresponding gold vectors: the model's outputs vary roughly $7.5\times$ less across inputs than the annotations they are meant to approximate. Consistent with this, validation macro-F1 peaked at the first epoch in all three runs, after which the multi-label head failed to escape its near-uniform initialization, a failure we attribute to the severe class imbalance dominating the binary cross-entropy loss. The collapse also explains the strategy invariance directly: the three target formulations differ only in how they shape probability mass across categories, a signal a near-constant predictor cannot exploit, and indeed the maximum per-probability difference between the m_1, m_2, and m_3 predictions on identical validation examples is below 0.006.

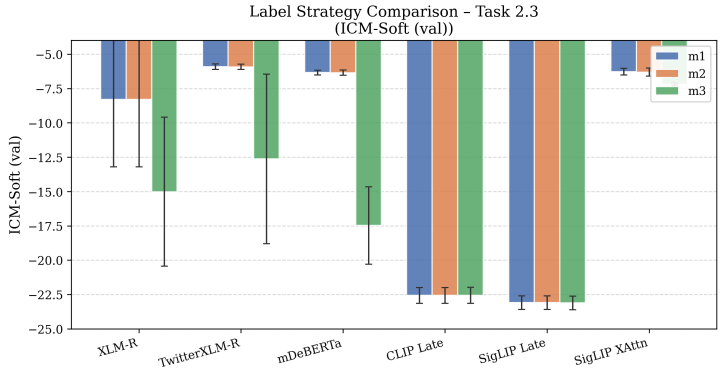
Table 5
Performance on Subtask 2.3 (Sexism Categorization) evaluating LeWiDi target strategies on Soft metrics.

Target Strategy	ICM-Soft	ICM-Soft Norm	Rank
Strategy m_1 (Dilution via UNKNOWN)	-7.5253	**0.1012**	58/115
Strategy m_2 (Explicit Votes Only)	**-7.5252**	**0.1012**	**57/115**
Strategy m_3 (Conservative Scaling)	-7.5337	0.1007	59/115

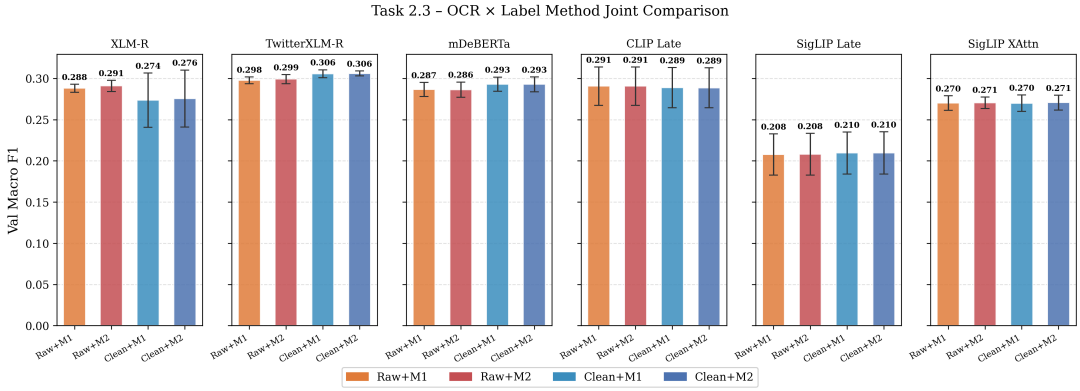
Table 6
Performance on Subtask 2.3 (Sexism Categorization) evaluating LeWiDi target strategies on Hard metrics.

Target Strategy	ICM-Hard	ICM-Hard Norm	F1	Rank
Strategy m_1 (Dilution via UNKNOWN)	**-2.2541**	**0.0323**	0.1745	**150/184**
Strategy m_2 (Explicit Votes Only)	-2.2549	0.0322	0.1744	151/184
Strategy m_3 (Conservative Scaling)	-2.2590	0.0313	**0.1760**	152/184

We note that the joint analysis demonstrates a compounding interaction between text quality and label strategy (Figure 4).



(a) ICM-Soft by Strategy



(b) Interaction: OCR vs. Target Strategy

Figure 4: Evaluation of target strategies (m_1, m_2, m_3) for Subtask 2.3. The joint analysis (b) reveals that the benefits of explicit-vote modeling (m_2) and conservative scaling (m_3) are amplified when applied to enhanced OCR text.

The performance gaps between m_1, m_2, and m_3 widen significantly when models are trained on clean OCR text compared to raw text. This suggests that when linguistic noise is removed, the model becomes much more sensitive to the nuances of the target probability distribution.

Error Analysis. Table 7 makes the failure mode concrete on individual validation examples. The consequences differ by gold structure. When annotator votes are diffuse (first row), the flat output happens to rank the correct category highest, yielding a deceptively low divergence. When a single visually grounded category dominates (second row, *Objectification* at 1.0 with text that does not carry the sexist content on its own), the model essentially ignores it. And when the gold spans several categories with a clear ordering (third row), the near-constant scores preserve almost none of that structure. Addressing this collapse, e.g., through per-class positive weighting, focal loss, or asymmetric losses tailored to imbalanced multi-label settings, is a prerequisite for the target-strategy differences

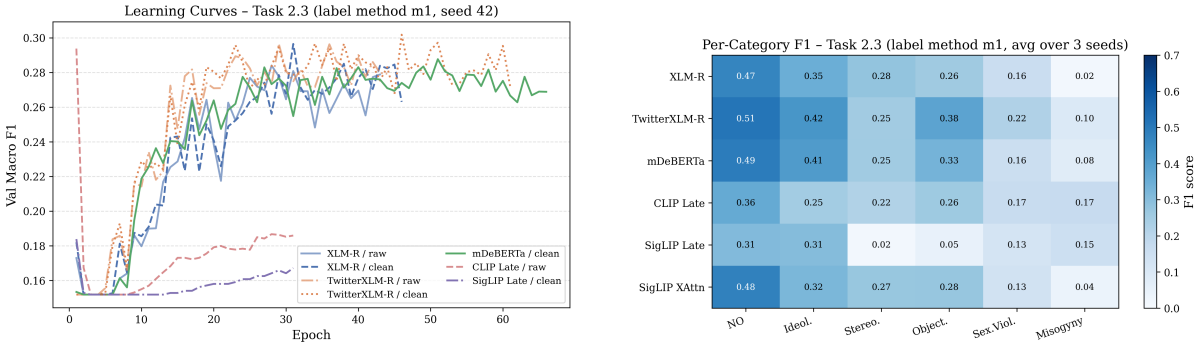
to become measurable, and we consider it the most important direction for improving Subtask 2.3 performance.

Table 7
Qualitative examples from the Subtask 2.3 validation set (CLIP late-fusion, m_2 targets, seed 123). Gold values are annotator vote proportions; Pred values are per-class sigmoid outputs. Regardless of input, the model produces near-uniform scores around 0.5, illustrating the output collapse discussed above.

OCR Text (excerpt)	Gold	Predicted	Failure Mode
ES LESBIANA Y SE QUEJA DE QUE VIVIMOS EN UNA SOCIEDAD MACHIS...	Ideol. 0.50; Object. 0.50; Stereo. 0.33; NO 0.17; Misog. 0.17	Ideol. 0.52; NO 0.50; Stereo. 0.49; Misog. 0.48; Sex.Viol. 0.47; Object. 0.45	Correct top label, no separation
La mujer perfecta según los hombres Cabellera de anuncio, si...	Object. 1.00; Stereo. 0.33; Sex.Viol. 0.33	NO 0.49; Object. 0.46; Misog. 0.46; Stereo. 0.45; Ideol. 0.44; Sex.Viol. 0.44	Missed dominant visual category
People Are Apparently Trying To Free A Murderer Because He's...	Stereo. 0.83; Ideol. 0.67; Misog. 0.50; Sex.Viol. 0.17	Ideol. 0.55; NO 0.53; Misog. 0.48; Stereo. 0.47; Object. 0.47; Sex.Viol. 0.44	Flat scores miss multi-label structure

4.5. Category-Level Analysis and Training Dynamics

To further understand model behavior in Subtask 2.3, we track learning dynamics and generate a per-category performance heatmap (Figure 5).



(a) Learning Curves (Subtask 2.3)

(b) Per-Category Heatmap

Figure 5: Detailed analysis of Subtask 2.3 training dynamics and category-specific classification performance.

The heatmap highlights significant class imbalances inherent to the dataset. Categories such as *Ideological and Inequality* are more readily identified by the text-only branches, whereas visually driven categories like *Objectification* require the cross-modal alignment provided by the SigLIP and CLIP architectures.

5. Conclusions and Future Directions

This work addresses sexism detection in memes through a pipeline that begins with LLM-driven OCR enhancement while reducing lexical noise. We then pass the cleaned text to a suite of text-only and multimodal classifiers that are afterwards aggregated via meta-ensemble. Annotator disagreement is handled through the three soft-label generation strategies evaluated under the LeWiDi paradigm [21]. OCR preprocessing reduces seed variance across all subtasks. The meta-ensemble outperforms every

standalone model. Modeling annotator uncertainty explicitly, however, remains an open challenge for the multi-label setting.

Our primary contribution is the development of an LLM-driven OCR enhancement pipeline. By pulling out emojis, hashtags, and mentions before correction and putting them back afterward, the noise in the extracted text dropped noticeably. The results show that this enhanced textual foundation reduces variance across random seeds and improves the downstream performance of both text-only and multimodal classifiers.

Through extensive comparative evaluations, we find that text-centric architectures (e.g., Twitter-XLM-R) excel at identifying explicit hate speech and judgmental intent (Subtask 2.2). However, for detecting visually implied sexism, such as objectification, the cross-modal alignment provided by multimodal vision-language models (CLIP and SigLIP) is necessary. By dynamically aggregating these models via a meta-ensemble, we achieve highly stable and competitive results for the single-label tasks.

Our exploration of the Learning with Disagreements (LeWiDi) paradigm in Subtask 2.3 shows that how annotator disagreement is modeled matters more than expected. By comparing three soft-label strategies (m_1, m_2, and m_3), we find that excluding "UNKNOWN" votes (m_2) yields marginally better ICM-Soft alignment, although the differences between strategies remain small in absolute terms, pointing to class imbalance rather than target design as the dominant bottleneck. On validation, the differences between strategies are more visible when models are trained on enhanced OCR text (Figure 4b), although on the submitted runs the absolute margins remain small.

5.1. Future Directions

A new addition to the EXIST 2026 campaign is the introduction of Human-Centered AI data, encompassing physiological signals (eye-tracking, heart rate, and EEG) recorded while subjects viewed the dataset [31, 32]. Due to computational and scope limitations, the present work relies exclusively on the visual and textual modalities.

We plan to integrate these physiological signals into the cross-attention layers, aligning with recent advancements in multimodal human-centric fusion [23]. By mapping gaze fixations to specific image patches or tracking arousal spikes elicited by the enhanced OCR text, future architectures could leverage unconscious cognitive processing as a latent indicator of a meme's underlying sexist intent.

Additionally, we aim to further refine our LeWiDi modeling by exploring dynamic thresholding mechanisms that weight explicit annotator votes based on historical demographic reliability.

Acknowledgments

The research presented in this paper is supported in part by The Academy of Romanian Scientists, through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV).

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to develop automation scripts for experiment orchestration (loading and sequentially executing model configurations from YAML files). After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

A. OCR Correction Prompt

Listing 1 shows the complete system prompt used for the Qwen-based OCR correction stage. The prompt restricts the model to character-level repairs, enforces language and metadata preservation,

and provides few-shot examples covering the most frequent failure modes observed in the dataset: accented-character loss, run-together words, and merged all-caps tokens.

Listing 1: System prompt for LLM-driven OCR correction.

```
You are an OCR error corrector. The text may be in English, Spanish, or a mix of both.
Fix ONLY characters that are clearly wrong due to OCR: '?' replacing accented letters, run-
    together words (missing spaces), obvious letter substitutions (e.g. 'rn'->'m', 'cl'->'d').
RULES - all mandatory:
1. NEVER translate. Keep every word in its original language.
2. Preserve every hashtag, @mention, and emoji exactly as-is.
3. Do NOT rephrase, summarise, reorder, or add words.
4. Output ONLY the corrected text, nothing else.

Examples:
Input:  El feminismo de hoy en d?a defiende la estupidez
Output: El feminismo de hoy en día defiende la estupidez

Input:  Lasrnujeres tambi?n acosan
Output: Las mujeres también acosan

Input:  Cuandonoobtiene suficienteslikes
Output: Cuando no obtiene suficientes likes

Input:  Women should notbe afraid
Output: Women should not be afraid

Input:  BECAUSESOMETIMESYOU'VE GOTTO TAKEONEFORTHETEAM
Output: BECAUSE SOMETIMES YOU'VE GOT TO TAKE ONE FOR THE TEAM
```

References

[1] V.-I. Ilie, C.-O. Truică, E.-S. Apostol, A. Paschke, Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings, IEEE Access 9 (2021) 162122–162146. doi:10.1109/ACCESS.2021.3132502.

[2] C.-O. Truică, E.-S. Apostol, It's all in the Embedding! Fake News Detection using Document Embeddings, Mathematics 11 (2023) 1–29(508). doi:10.3390/math11030508.

[3] C.-O. Truică, E.-S. Apostol, T. Ștefu, P. Karras, A Deep Learning Architecture for Audience Interest Prediction of News Topic on Social Media, in: International Conference on Extending Database Technology (EDBT2021), 2021, pp. 588–599. doi:10.5441/002/EDBT.2021.69.

[4] M.-D. Cotelin, E.-S. Apostol, C.-O. Truică, NetGuardAI at EXIST2025: Sexism Detection using mDeBERTa, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2025, pp. 1889–1897.

[5] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for EXIST2024, in: Working Notes of the Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 1157–1164.

[6] C.-O. Truică, E.-S. Apostol, MisRoBÆRTa: Transformers versus Misinformation, Mathematics 10 (2022) 1–25(569). doi:10.3390/math10040569.

[7] D.-E. Burghelea, C.-O. Truică, E.-S. Apostol, VERIT-ALBERT: A Finetuned LLM Approach for Verifying Information Credibility, in: The 2025 24th RoEduNet Conference: Networking in Education and Research (RoEduNet), 2025, pp. 1–6. doi:10.1109/RoEduNet68395.2025.11208267.

[8] N. C. J. Le Pollotec, E. S. Apostol, C.-O. Truică, Netguardai at multipride: Multilingual detection of reclaimed language, in: EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26– 27, Bari, IT, 2026, pp. 1–8.

[9] C.-O. Truică, E.-S. Apostol, A. Paschke, Awakened at CheckThat! 2022: Fake News Detection using BiLSTM and sentence transformer, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2022, pp. 749–757.

[10] A. Petrescu, E.-S. Apostol, C.-O. Truică, Awakened at EXIST2025: Adaptive Mixture of Transformers, in: Working Notes of the Conference and Labs of the Evaluation Forum, 2025, pp. 2104–2110.
[11] A. Petrescu, C.-O. Truică, E.-S. Apostol, Language-based Mixture of Transformers for Sexism Identification in Social Networks, in: Conference and Labs of the Evaluation Forum, 2025, pp. 142–155. doi:10.1007/978-3-032-04354-2_10.
[12] C.-O. Truică, E.-S. Apostol, A.-G. Ilie, A. Paschke, HarmLLaMA: Harmful Language Detection with Large Language Models, in: International Conference on Intelligent Computer Communication and Processing (ICCP 2025), 2025, pp. 1–8. doi:10.1109/ICCP68926.2025.11427180.
[13] C.-O. Truică, E.-S. Apostol, M. Marogel, A. Paschke, GETAE: Graph Information Enhanced Deep Neural NeTwork Ensemble ArchitecturE for fake news detection, Expert Systems with Applications 275 (2025) 126984. doi:10.1016/j.eswa.2025.126984.
[14] C.-O. Truică, E.-S. Apostol, P. Karras, DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection, Knowledge-Based Systems 294 (2024) 111715. doi:10.1016/j.knosys.2024.111715.
[15] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, P. Karras, MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media, IEEE Access 11 (2023) 125861–125873. doi:10.1109/ACCESS.2023.3331220.
[16] A. Petrescu, C.-O. Truică, E.-S. Apostol, P. Karras, Sparse Shield: Social Network Immunization vs. Harmful Speech, in: ACM International Conference on Information and Knowledge Management (CIKM2021), ACM, 2021, pp. 1426–1436. doi:10.1145/3459637.3482481.
[17] E.-S. Apostol, Özgur Coban, C.-O. Truică, CONTAIN: A community-based algorithm for network immunization, Engineering Science and Technology, an International Journal 55 (2024) 1–10(101728). doi:10.1016/j.jestch.2024.101728.
[18] M.-D. Cotelin, C.-O. Truică, E.-S. Apostol, Cleannews: a network-aware fake news mitigation architecture for social media, arXiv preprint arXiv:2509.04489 (2025).
[19] C.-O. Truică, A.-T. Constantinescu, E.-S. Apostol, StopHC: A Harmful Content Detection and Mitigation Architecture for Social Media Platforms, in: IEEE International Conference on Intelligent Computer Communication and Processing, 2024, pp. 1–5. doi:10.1109/ICCP63557.2024.10793051.
[20] E.-S. Apostol, C.-O. Truică, A. Paschke, ContCommRTD: A Distributed Content-Based Misinformation-Aware Community Detection System for Real-Time Disaster Reporting, IEEE Transactions on Knowledge and Data Engineering (2024) 1–12. doi:10.1109/tkde.2024.3417232.
[21] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023 - learning with disagreement for sexism identification and characterization, in: CLEF 2023, Springer, 2023.
[22] P. Italiani, D. Gimeno-Gomez, L. Ragazzi, G. Moro, P. Rosso, Memeweaver: Inter-meme graph reasoning for sexism and misogyny detection, 2026. URL: https://arxiv.org/abs/2601.08684. arXiv:2601.08684.
[23] I. Arcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals, 2026. URL: https://arxiv.org/abs/2602.23862. arXiv:2602.23862.
[24] 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. doi:10.18653/v1/2020.acl-main.747.
[25] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 258–266.
[26] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021.

[27] 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: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[28] X. Zhai, A. Kolesnikov, N. Houlsby, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11975–11986. doi:https://doi.org/10.1109/ICCV51070.2023.01100.

[29] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, Z. Liu, M. Zeng, An empirical study of training end-to-end vision-and-language transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18166–18176. doi:https://doi.org/10.1109/CVPR52688.2022.01763.

[30] T. Fang, S. Yang, K. Lan, D. F. Wong, J. Hu, L. S. Chao, Y. Zhang, Is ChatGPT a Good OCR Post-Processing System? A Comprehensive Evaluation, arXiv preprint arXiv:2303.00526 (2023).

[31] 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.

[32] 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.