CLIP-Based Multimodal Approach for Sexism Identification in Memes: Farabi University at EXIST 2026 Task 2.1*

Moldakhan Yermukhamed[1*] and Bolatbek Milana[1]

[1] *Department of Cryptology and Cybersecurity, Farabi University, Almaty, Kazakhstan*

Abstract

This paper describes our participation in the EXIST 2026 shared task, specifically Task 2.1 — Sexism Identification (hard label) in multimodal meme content. We propose a multimodal classification system based on the CLIP (Contrastive Language–Image Pre-Training) model, fine-tuned to detect sexist content in memes containing both text and image modalities. Text and visual embeddings are concatenated and passed through a binary classification head. Our system achieved an accuracy of 0.6173 and a macro F1-score of 0.65 on the validation set, with an AUC of 0.65. We analyze the strengths and limitations of our approach and outline directions for future improvement.

Keywords

Sexism identification, Multimodal learning, CLIP, Meme classification, EXIST 2026

1. Introduction

Online sexism in social media has become an increasingly serious concern, with memes representing a particularly challenging form of harmful content due to their multimodal nature. The EXIST (sEXism Identification in Social neTworks) shared task at CLEF 2026 provides a benchmark for systems that automatically detect sexist content across text, images, and their combination.

Task 2.1 specifically addresses Sexism Identification under a hard-label scheme: given a meme (image + text), the system must output a binary decision — whether the meme is sexist (YES) or not (NO). The ground truth labels are derived from multiple annotators via majority vote, making this a soft-to-hard label mapping problem.

Memes pose unique challenges. They often rely on implicit cultural references, irony, or visual metaphors that are difficult to interpret without understanding both modalities jointly. This motivates the use of cross-modal models such as CLIP, which are pre-trained to align text and visual representations in a shared embedding space.

In this work, we describe our approach: a fine-tuned CLIP classifier that concatenates text and image embeddings and applies a multilayer perceptron (MLP) head for binary classification. We present our experimental results, visualizations, and an analysis of model behavior.

CLEF 2026 Working Notes, 21 — 24 September 2026, Jena, Germany
*Corresponding author.
ermosh112233@gmail.com (M. Yermukhamed)

2. Task Description

EXIST 2026 Task 2.1 is a binary classification task. Each data instance consists of:

- A meme image (JPEG/PNG)
- OCR-extracted or manually annotated text embedded in the meme
- A set of annotator labels (YES/NO), from which the hard gold label is derived by majority vote

The dataset covers multilingual content including English and Spanish memes. Systems are evaluated on accuracy and macro-averaged F1-score.

3. System Description

3.1. Preprocessing

Text extracted from memes is normalized using a custom cleaning pipeline: URLs and web links are removed, text is lowercased, and non-alphabetic characters (except spaces) are stripped. Whitespace is normalized. Images that cannot be loaded are replaced with a black 224×224 placeholder to ensure pipeline robustness.

Ground truth labels are obtained from the annotator label lists via majority vote: if the number of YES annotations exceeds NO annotations, the instance is labeled positive (1); otherwise negative (0).

3.2. Model Architecture

We use **CLIP ViT-B/32** (`openai/clip-vit-base-patch32`) as the backbone. CLIP provides pre-aligned text and image encoders trained on 400 million image-text pairs via contrastive learning, making it naturally suited for multimodal tasks.

Our classification head receives the concatenation of the 512-dimensional text embedding and the 512-dimensional image embedding (total: 1024 dimensions), and maps them to a single logit through:

- Linear($1024 \longrightarrow 512$)
- ReLU activation
- Dropout(0.3)
- Linear($512 \longrightarrow 1$)

Binary cross-entropy with logits loss (BCEWithLogitsLoss) is used as the training objective.

3.3. Training Setup

The model was trained for 10 epochs with the Adam optimizer (learning rate 2×10^{-5}) and a batch size of 32. The training set was split 80/20 into train and validation subsets (stratified split with random seed 42). No class weighting was applied.

4. Experimental Results

4.1. Quantitative Results

Table 1 summarizes the evaluation results on the validation set.

Table 1
Classification Results on Validation Set

Metric	Value
Accuracy	0.6173
Macro F1-score	0.65
AUC (ROC)	0.65
Precision (class 0 — not sexist)	0.58
Recall (class 0)	0.85
F1-score (class 0)	0.68
Precision (class 1 — sexist)	0.71
Recall (class 1)	0.38
F1-score (class 1)	0.49

The model achieves moderate accuracy but shows a notable asymmetry: it identifies non-sexist content with much higher recall (0.85) than sexist content (0.38), indicating a bias toward the negative class.

4.2. Confusion Matrix

The confusion matrix (Figure 1) shows 231 true negatives and 260 true positives, but also 173 false negatives and 133 false positives. The high number of false negatives (sexist memes classified as not sexist) is the primary bottleneck for the model's F1-score on the positive class.

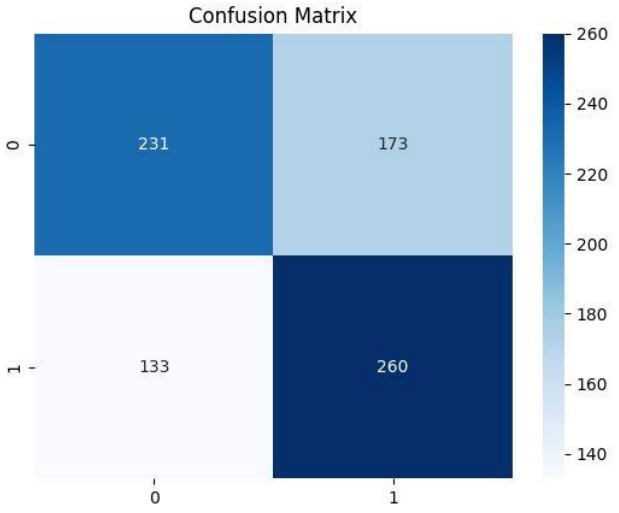

Figure 1: Confusion matrix on the validation set (0 = not sexist, 1 = sexist).

4.3. ROC Curve

The ROC curve (Figure 2) shows an AUC of 0.65, indicating the model performs better than a random baseline (AUC = 0.50) but has substantial room for improvement, particularly in the low false-positive-rate regime.

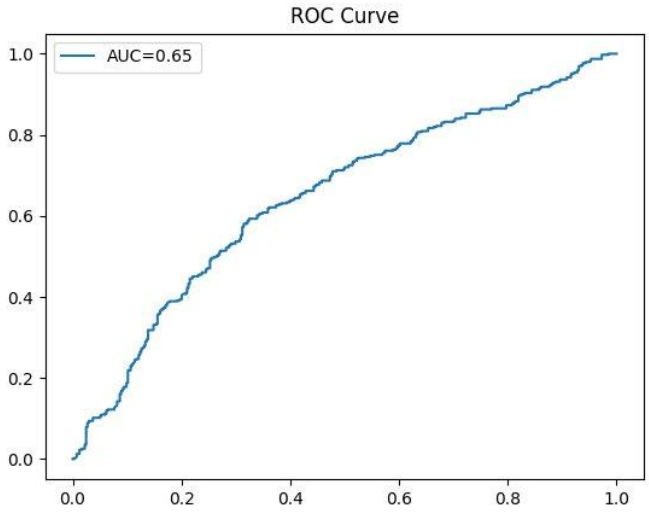

Figure 2: ROC curve with AUC = 0.65.

4.4. Training Dynamics

Figure 3 shows the training loss and validation accuracy over 10 epochs. Both curves exhibit high variance and do not converge stably, suggesting the model is sensitive to the learning rate and batch-level randomness. Validation accuracy fluctuates between approximately 0.50 and 0.63, with no consistent upward trend, pointing to optimization instability.

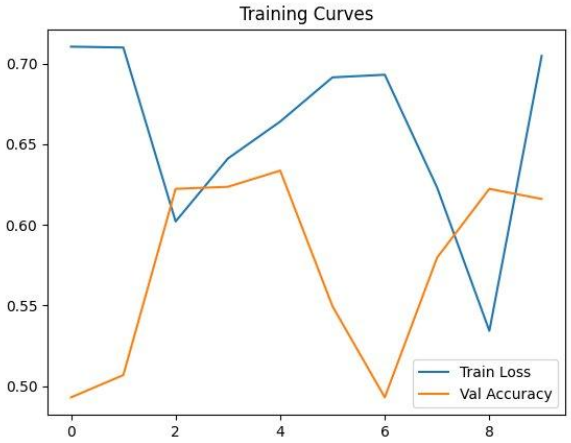

Figure 3: Training loss and validation accuracy over 10 epochs.

4.5. t-SNE Embedding Visualization

Figure 4 shows t-SNE projections of the concatenated CLIP embeddings from the validation set, colored by ground truth label (yellow = not sexist, dark purple = sexist). The classes show significant overlap in the embedding space, indicating that the CLIP features alone do not produce well-separated representations for this task.

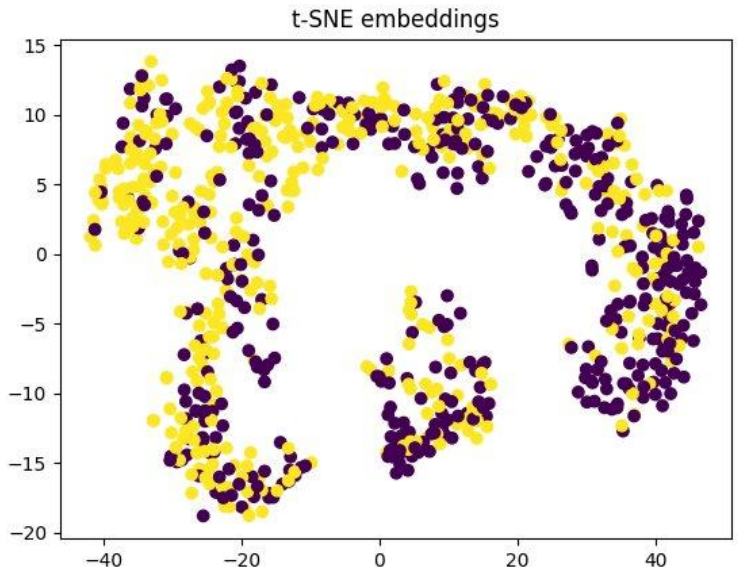

Figure 4: t-SNE visualization of CLIP embeddings. Yellow: not sexist; dark purple: sexist.

5. Analysis

Several factors contribute to the observed results:

Class imbalance and annotation noise. The majority-vote label aggregation introduces label noise, particularly for borderline cases where annotators disagreed. Our model struggles on these ambiguous examples.

Insufficient fine-tuning depth. We fine-tune the entire CLIP backbone end-to-end, but the low learning rate (2×10^{-5}) and relatively small number of epochs may be insufficient for the model to adapt to this domain. A frozen backbone with a larger classification head, or layer-wise learning rate decay, could improve results.

Lack of augmentation and regularization. No data augmentation was applied to images, and only a single dropout layer is used. More aggressive regularization may improve generalization.

Multilingual content. The dataset includes both English and Spanish memes. CLIP's text encoder has limited multilingual capacity, which may reduce performance on Spanish-language content. A multilingual text encoder (e.g., mCLIP or XLM-RoBERTa) could address this.

Implicit and visual sexism. Many sexist memes rely on visual metaphors or cultural references that are difficult for a fixed CLIP backbone to capture without task-specific adaptation.

6. Conclusion and Future Work

We presented a CLIP-based multimodal classifier for EXIST 2026 Task 2.1 (Sexism Identification). Our system achieves accuracy of 0.6173 and macro F1 of 0.4925, with an AUC of 0.65. While the approach demonstrates the feasibility of joint text-image modeling for meme classification, results indicate that the task requires more sophisticated strategies.

Future directions include: (1) incorporating a multilingual text encoder to better handle Spanish content; (2) applying class-balanced sampling or loss weighting to address the recall gap on sexist instances; (3) exploring contrastive fine-tuning strategies tailored to the hateful content domain; (4) leveraging larger multimodal models such as LLaVA or BLIP-2; and (5) applying ensemble methods combining CLIP with text-only classifiers.

Acknowledgements

The authors thank the EXIST 2026 organizers for providing the dataset and evaluation framework.

References

[1] Radford, A., Kim, J. W., Hallacy, C., et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML 2021*.

[2] Fersini, E., Nozza, D., Rosso, P., et al. 2022. EXIST 2022: sEXism Identification in Social neTworks. In *Working Notes of CLEF 2022*.

[3] Plaza-del-Arco, F. M., et al. 2023. EXIST 2023 — Learning with Disagreement for Sexism Identification. In *Working Notes of CLEF 2023*.

[4] Wolf, T., et al. 2020. HuggingFace's Transformers: State-of-the-Art Natural Language Processing. In *EMNLP 2020*.

[5] Van der Maaten, L., and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.

[6] Plaza, L., Carrillo-de-Albornoz, J., Arcos, I., Aloy-Mayo, M., Rosso, P., García-Arias, E., and Spina, D. 2026. 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)*.

[7] Plaza, L., Carrillo-de-Albornoz, J., Arcos, I., Aloy-Mayo, M., Rosso, P., García-Arias, E., and Spina, D. 2026. Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media (Extended Overview). In *CLEF 2026 Working Notes*. Eva Sánchez Salido, Alberto Barrón-Cedeño, Alba García Seco de Herrera, Sean MacAvaney, and Julia Maria Struß (Eds.).