Subjective Sexism Identification in Social Media Memes Using Multimodal Embedding Ensembles

Notebook for the EXIST Lab at CLEF 2026

Sudharshan PS

Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India

Abstract

Sexism identification in social media memes is a challenging multimodal task that requires understanding textual and visual information while accounting for the subjective nature of human perception. The EXIST 2026 Task 2.1 dataset further introduces physiological measurements collected during the annotation process, providing complementary information about annotator responses and disagreement. The proposed approach combines multilingual SentenceTransformer text embeddings, OpenAI CLIP ViT-B/32 image embeddings, and engineered demographic and physiological features derived from Eye Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) measurements. These multimodal representations are integrated through feature-level fusion and modeled using an ensemble of Logistic Regression, LightGBM, and Random Forest classifiers. To address moderate class imbalance and improve prediction quality, stratified cross-validation, class-aware learning strategies, median imputation for missing physiological values, and out-of-fold threshold optimization are employed. Experiments conducted on the official EXIST 2026 Task 2.1 dataset under a five-fold stratified cross-validation setting achieved an accuracy of 71.81%, a Macro F1-score of 0.6842, a ROC-AUC score of 0.7470, and an Average Precision score of 0.8322 during local validation. The official submissions ranked 33rd in the SOFT_ALL setting and 56th in the HARD_ALL setting, demonstrating competitive performance under the Learning with Disagreement evaluation framework. The results demonstrate that integrating multimodal semantic representations with demographic and physiological information provides an effective approach for modeling the subjective perception of sexism in social media memes.

Keywords

Sexism Identification, Multimodal Learning, Social Media Memes, Ensemble Learning, Sentence Transformers, CLIP, Subjective Classification, EXIST 2026

1. Introduction

Social media platforms have become an important medium for sharing opinions, experiences, and cultural content. Memes are particularly influential because they convey messages in a concise and engaging manner. While memes are often used for humor and entertainment, they can also communicate harmful stereotypes and discriminatory viewpoints, including sexist content. Automatically identifying sexism in memes remains a challenging task because the intended meaning frequently arises from the interaction between textual and visual information rather than either modality in isolation.

The sEXism Identification in Social neTworks (EXIST) shared task provides a benchmark for studying sexism detection in multilingual and multimodal online content while adopting the Learning with Disagreement (LeWiDi) paradigm, which preserves multiple annotator perspectives instead of relying solely on majority-vote labels. This edition further introduces physiological information collected during the annotation process, including Eye Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) measurements, to better capture the subjective nature of sexism perception. These additional modalities provide information about annotator responses and offer valuable insights into the variability that naturally arises during the annotation process.

This work presents a multimodal ensemble framework for EXIST 2026 Task 2.1 that combines textual, visual, demographic, and physiological information. Textual representations are extracted

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

sudhar0705@gmail.com (S. PS)

using multilingual SentenceTransformer embeddings, while visual representations are obtained using the pretrained OpenAI CLIP ViT-B/32 image encoder. Demographic characteristics and engineered physiological descriptors derived from ET, HR, and EEG signals are incorporated as structured numerical features to complement the multimodal semantic representations. These are integrated through feature-level fusion and modeled using an ensemble of Logistic Regression, LightGBM, and Random Forest classifiers. Rather than directly integrating heterogeneous physiological signals within an end-to-end multimodal architecture, the proposed embedding-based ensemble naturally accommodates structured numerical features while preserving the complementary information learned from textual and visual representations.

Experimental results demonstrate that integrating multimodal semantic representations with annotator-related physiological information provides an effective approach for modeling the subjective characteristics of sexism perception in social media memes. The proposed framework highlights the effectiveness of combining pretrained multilingual language representations, vision-language embeddings, and physiological descriptors for disagreement-aware sexism identification under the EXIST 2026 evaluation framework.

2. Related Work

Automatic sexism detection has attracted increasing attention in recent years as social media platforms continue to serve as major channels for the spread of harmful and discriminatory content. To support research in this area, the sEXism Identification in Social neTworks (EXIST) shared task was introduced as a benchmark for multilingual sexism identification and characterization across different forms of online content. The 2024 edition emphasized the concept of learning with disagreement by preserving multiple annotator perspectives instead of relying solely on majority-vote labels, highlighting the subjective nature of sexism perception [3]. Building on this foundation, the 2025 edition expanded the benchmark to include additional modalities such as memes and TikTok videos, encouraging the development of more sophisticated multimodal and annotator-aware systems [4].

Recent participant systems have demonstrated the effectiveness of transformer-based architectures for sexism detection. Models based on BERT, XLM-RoBERTa, and other multilingual transformers have achieved strong performance by leveraging contextual language representations and cross-lingual transfer capabilities. For meme classification tasks, several teams explored multimodal architectures that jointly process textual and visual information. Approaches such as PINK [5], UMUTeam [6], ArcosGPT [7], and I2C-UHU-Altair [8] showed that combining language models with visual encoders significantly improves the identification of sexist content in memes compared to text-only systems. These findings highlight the importance of capturing interactions between textual and visual cues when analyzing multimodal social media content.

Another important research direction within EXIST focuses on learning from annotator disagreement and modeling subjectivity. Since perceptions of sexism often vary across individuals, several studies have investigated soft-label learning and annotator-aware approaches rather than relying exclusively on hard labels. NYCU-NLP [9], DUTH[10], and the work of Chen et al. [11] demonstrated that incorporating annotator information and disagreement patterns can improve performance in subjective classification tasks. The introduction of physiological signals in EXIST 2026 further extends this idea by providing additional information about annotator responses through Eye-Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) measurements.

Ensemble learning has also proven effective for sexism detection. FraunhoferSIT [12], CLTL [14], and Abidi et al. [13] reported competitive results using combinations of multiple classifiers and representation models. By integrating diverse information from different feature spaces, ensemble approaches often achieve greater robustness and generalization than individual models. Motivated by these findings, this work integrates multimodal textual and visual representations with demographic and physiological descriptors within an ensemble learning framework. Unlike conventional multimodal

approaches that primarily focus on textual and visual information, the proposed framework additionally incorporates structured physiological features derived from Eye Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) measurements to capture annotator responses. The integration of semantic representations with demographic and physiological information enables the framework to better model the subjective nature of sexism perception while benefiting from the strengths of multiple machine learning classifiers.

3. Dataset and Task Description

3.1. Task Overview

The sEXism Identification in Social neTworks (EXIST) task focuses on the automatic identification and characterization of sexist content in social media. Task 2.1 addresses binary sexism identification in multilingual memes, where systems must determine whether a given meme should be classified as YES (sexist) or NO (non-sexist). Unlike conventional text classification tasks, meme understanding requires jointly analyzing textual and visual information, as the intended meaning often emerges from the interaction between both modalities. In addition, the task is inherently subjective, since perceptions of sexism may vary across annotators depending on their background, experiences, and interpretation of the content.

3.2. Dataset Structure

The EXIST 2026 dataset consists of multilingual social media memes annotated by multiple human annotators. Each sample contains the meme image, associated textual content, demographic information describing the annotators, and physiological measurements collected during the annotation process. These physiological signals include Eye-Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) recordings, which were introduced to capture aspects of human perception and disagreement during the labeling process.

Each dataset instance contains information including the meme identifier, language, textual content, image reference, annotator judgments, demographic metadata, and physiological measurements. The availability of multiple annotator judgments enables the computation of both majority-vote labels and disagreement-aware statistics, allowing the dataset to capture the inherent subjectivity of sexism perception. In this study, the binary labels are derived from the annotator decisions, where memes are categorized as sexist or non-sexist according to the task definition.

The dataset contains 3,984 training instances and 1,053 test instances spanning both English and Spanish. Analysis of the training data reveals a moderate class imbalance, with 2,617 samples labeled as YES and 1,367 samples labeled as NO. Beyond the final labels, the dataset also provides valuable information regarding annotator agreement and physiological responses. These additional modalities offer signals that may help capture the subjective characteristics of sexism perception. To leverage this information, the system extracts textual, visual, demographic, and physiological features and integrates them within a multimodal ensemble framework for classification.

3.3. Exploratory Data Analysis

Exploratory analysis was conducted to better understand the characteristics of the EXIST 2026 Task 2.1 dataset before model development. The dataset contains 3,984 training memes annotated for binary sexism identification across English and Spanish languages. In addition to binary labels, the dataset preserves annotator-level information, enabling the computation of soft-label scores that capture disagreement and uncertainty among annotators. These properties make the dataset particularly suitable for studying the subjective nature of sexism perception.

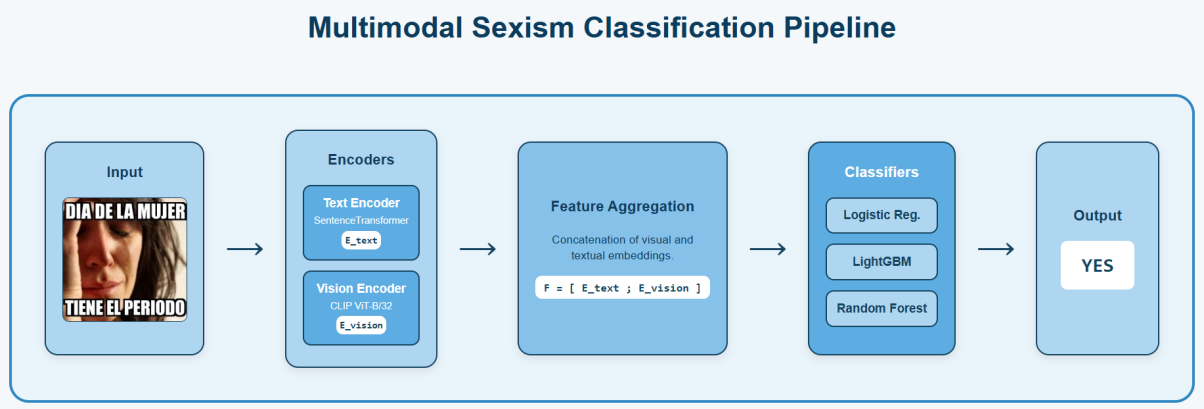

Figure 1: Example of Sexism Classification in Memes

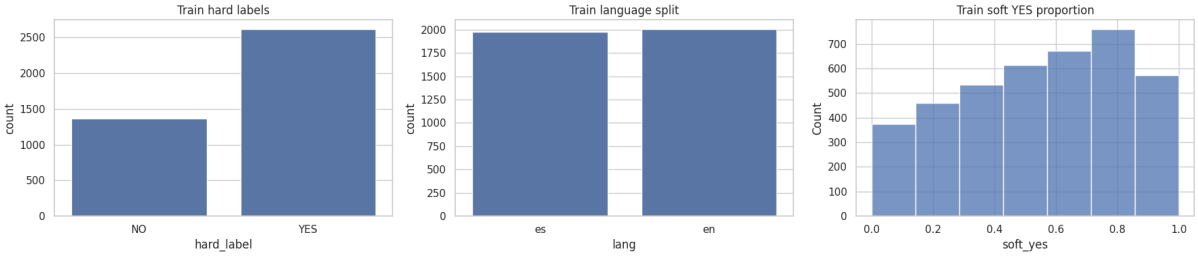

Figure 2: Dataset overview showing (left) binary label distribution, (center) language distribution, and (right) soft YES proportion distribution derived from annotator votes.

Figure 2 summarizes several important characteristics of the training data. The label distribution reveals a moderate class imbalance, with 2,617 memes labeled as YES (sexist) and 1,367 labeled as NO (non-sexist). The language distribution is nearly balanced, containing 2,005 English memes and 1,979 Spanish memes. This balanced multilingual composition reduces the likelihood of language-specific bias during training and allows the model to learn from comparable amounts of data across both languages.

The distribution of soft YES proportions further highlights the subjective nature of the task. Soft-label scores are computed from the proportion of annotators assigning a positive label to each meme. The distribution spans the entire range from 0 to 1, indicating substantial variation in annotator agreement. While many memes receive strong consensus, a considerable number of samples fall within intermediate ranges, reflecting uncertainty regarding whether the content should be considered sexist. This observation motivates the use of probabilistic learning strategies rather than relying solely on hard-label supervision.

Figure 3 illustrates the distribution of meme text lengths. The distribution is highly right-skewed, with most memes containing relatively short textual content and a small number of outliers extending to much larger lengths. The majority of samples are concentrated below 200 characters, suggesting that sexism detection often relies on short contextual statements, captions, or phrases. At the same time, the presence of longer textual instances requires models to effectively capture semantic information across varying sequence lengths.

Figure 4 presents the relationship between hard labels and soft-label proportions. Memes assigned a hard YES label generally exhibit higher positive proportions, while NO memes are associated with lower values. However, a noticeable overlap exists near the decision boundary, indicating that some memes receive mixed judgments from annotators. This overlap provides direct evidence of disagreement among annotators and demonstrates that sexism identification cannot always be treated as a purely objective classification problem.

Table 1 summarizes the main dataset characteristics. The mean soft-label score is 0.556, while

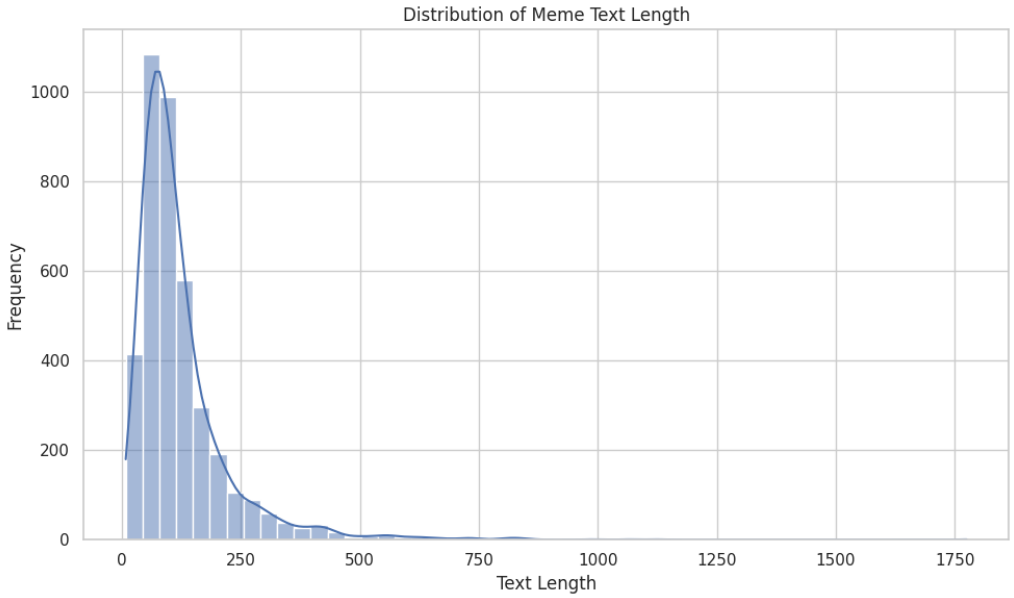

Figure 3: Distribution of meme text lengths in the training dataset. Most memes contain relatively short textual content, while a small number of samples exhibit substantially longer text lengths.

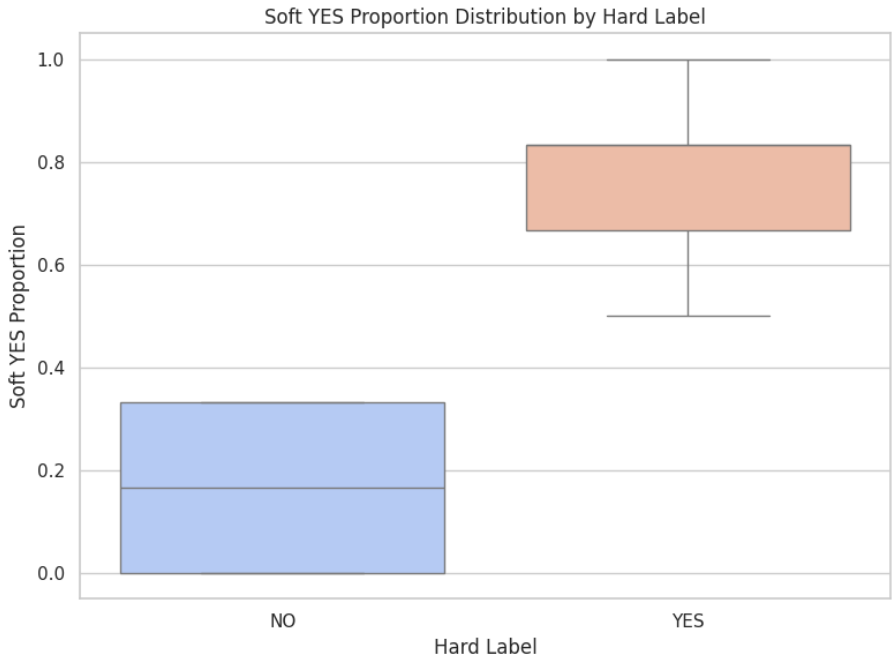

Figure 4: Distribution of soft YES proportions grouped by hard labels. The overlap near the decision boundary highlights the presence of annotator disagreement.

the median value is 0.667, indicating that positive judgments are slightly more prevalent overall. Furthermore, the average agreement strength of 0.542 suggests that complete consensus is not always achieved among annotators. These observations support the motivation behind EXIST 2026 and further justify the integration of multimodal semantic representations, physiological descriptors, and disagreement-aware learning strategies within the proposed framework.

Overall, the exploratory analysis reveals three important characteristics of the dataset: a moderate class imbalance, substantial annotator disagreement, and a balanced multilingual composition. These findings motivate the use of multimodal representations, soft-label information, and physiological signals to better capture the subjective nature of sexism perception in social media memes.

Table 1
Dataset Statistics Summary

Metric	Value
Training Samples	3,984
Test Samples	1,053
YES Labels	2,617
NO Labels	1,367
English Memes	2,005
Spanish Memes	1,979
Mean Soft YES Score	0.556
Median Soft YES Score	0.667
Mean Agreement Strength	0.542
Median Agreement Strength	0.667
Modalities	Text, Image, ET, EEG, HR

4. Proposed Methodology

The proposed system approaches sexism identification as a multimodal binary classification problem by combining textual, visual, demographic, and physiological information within a unified ensemble framework. Given a social media meme, the objective is to determine whether the content should be classified as sexist or non-sexist. Unlike conventional meme classification systems that rely solely on textual or visual content, the proposed approach additionally incorporates physiological signals collected during the annotation process, enabling the model to capture aspects of human perception and disagreement.

The overall framework consists of four main stages: multimodal feature extraction, physiological feature aggregation, feature fusion, and ensemble-based classification. Textual and visual representations are first extracted using pretrained foundation models. These representations are subsequently combined with statistical descriptors derived from Eye-Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) measurements. The resulting feature vector is then processed by an ensemble of machine learning classifiers to generate the final prediction.

4.1. Multimodal Feature Extraction

Textual information is represented using multilingual SentenceTransformer embeddings generated by the `paraphrase-multilingual-mpnet-base-v2` model. The model produces fixed-length 768-dimensional semantic representations that effectively capture contextual information across both English and Spanish while maintaining a shared multilingual embedding space. The multilingual embedding space facilitates semantic alignment across English and Spanish memes, enabling the framework to learn language-independent textual representations. Since the EXIST dataset contains both English and Spanish memes, multilingual sentence embeddings provide a unified representation space that facilitates cross-lingual generalization.

Visual information is extracted using the pretrained OpenAI CLIP ViT-B/32 image encoder, which produces 512-dimensional visual embeddings aligned with semantic textual representations. The pretrained vision-language model enables robust visual feature extraction without requiring task-specific fine-tuning. The pretrained CLIP image encoder captures high-level semantic information from meme images, complementing the contextual information obtained from textual embeddings, and learns aligned visual and textual representations through large-scale image-text pretraining and has demonstrated strong performance across a variety of multimodal understanding tasks. For each meme image, the pretrained image encoder produces a dense visual embedding that captures semantic information beyond the textual content alone.

The textual and visual embeddings provide complementary semantic information. While Sentence-

Transformer embeddings capture contextual linguistic cues from meme text, CLIP image embeddings encode high-level visual semantics present in the corresponding image. The combination of both modalities enables the framework to model interactions between textual and visual content that are essential for understanding the intended meaning of social media memes.

4.2. Physiological Feature Engineering

A distinguishing characteristic of the EXIST 2026 dataset is the inclusion of physiological measurements collected from annotators during the labeling process. These measurements include Eye Tracking (ET), Electroencephalography (EEG), and Heart Rate (HR) recordings, providing additional information regarding annotator responses beyond the semantic content of the meme.

Direct physiological recordings vary considerably in length and structure across annotators, making them unsuitable for direct integration into conventional machine learning models. To obtain fixed-length numerical representations, statistical descriptors are computed independently for each physiological variable. These engineered descriptors capture the overall characteristics of annotator responses while remaining robust to differences in recording duration.

In addition to physiological measurements, demographic characteristics derived from annotator metadata are incorporated as structured numerical descriptors. Together, the demographic and physiological features provide complementary information regarding annotator behaviour and disagreement that cannot be directly inferred from textual or visual content alone.

4.3. Feature Fusion

After feature extraction, all modalities are integrated through feature-level fusion. SentenceTransformer embeddings, CLIP image embeddings, demographic descriptors, and engineered physiological descriptors are concatenated to form a unified multimodal representation.

Prior to fusion, numerical features are standardized and missing physiological values are handled through median imputation. This preprocessing step ensures that all modalities contribute consistently during classification and prevents individual feature groups from dominating the learning process.

The final fused representation combines semantic, visual, demographic and physiological information within a unified feature space, allowing the downstream classifiers to learn interactions across modalities.

4.4. Ensemble-Based Classification

The fused multimodal features are processed using an ensemble of three classifiers: Logistic Regression, LightGBM, and Random Forest.

Logistic Regression serves as a strong linear classifier that captures global relationships across the multimodal feature space while producing well-calibrated probabilistic predictions. LightGBM models complex non-linear interactions among textual, visual, demographic, and physiological features through gradient-boosted decision trees. Random Forest further improves robustness by learning diverse decision boundaries from structured feature subsets, reducing sensitivity to noisy physiological measurements.

Each classifier independently estimates the probability of the positive class. The final prediction is obtained through weighted soft voting, where the individual probability estimates are combined using empirically determined ensemble weights. This strategy exploits the respective strengths of linear and non-linear classifiers, resulting in improved robustness and generalization compared with individual models.

4.5. Training Strategy and Threshold Optimization

Model development follows a five-fold stratified cross-validation strategy to preserve the class distribution across all validation folds. Since the dataset exhibits moderate class imbalance, model-specific

Table 2
Training and optimization configuration of the proposed ensemble framework

Parameter	Value
Task	Binary Sexism Classification
Training Samples	3,984
Test Samples	1,053
Languages	English, Spanish
Validation Strategy	5-Fold Stratified Cross Validation
Text Encoder	paraphrase-multilingual-mpnet-base-v2
Image Encoder	CLIP ViT-B/32
Physiological Features	EEG, ET, HR Statistics
Feature Fusion	Feature Concatenation
Missing Value Handling	Median Imputation
Feature Scaling	Standard Scaling
Logistic Regression Weight	0.35
LightGBM Weight	0.45
Random Forest Weight	0.20
Logistic Regression Class Weight	balanced
Random Forest Class Weight	balanced_subsample
LightGBM Scale Pos Weight	$\frac{N_{neg}}{N_{pos}}$
Oversampling	RandomOverSampler (ROS)
Threshold Search Range	0.20 – 0.80
Threshold Search Step	0.0025
Optimal Threshold	0.5675
Optimization Metric	Macro F1 Score
Feature Dimension	1732
Tabular Feature Dimension	452

class weighting strategies are employed during training. In addition, RandomOverSampler is applied for Logistic Regression to reduce the impact of class imbalance during model learning.

Rather than adopting the default classification threshold of 0.5, threshold optimization is performed using out-of-fold validation predictions. Candidate thresholds ranging from 0.20 to 0.80 with a step size of 0.0025 are evaluated, and the threshold that maximizes the Macro F1-score is selected for inference. This procedure improves the balance between precision and recall while accounting for the subjective and moderately imbalanced nature of the dataset.

Table 2 summarizes the principal training configuration of the proposed framework, including feature extraction models, preprocessing strategies, ensemble weights, validation protocol, and optimization settings.

4.6. Implementation Details

The proposed framework was implemented in Python using the SentenceTransformers library for multilingual text embeddings and OpenAI CLIP ViT-B/32 for visual feature extraction. Machine learning models were implemented using Scikit-learn and LightGBM. Missing physiological values were imputed using the median strategy, followed by feature standardization using StandardScaler. Model evaluation was performed using five-fold stratified cross-validation, and all experiments were conducted using the same preprocessing pipeline across every validation fold.

5. Results and Discussion

The proposed multimodal ensemble framework was evaluated using both local validation experiments and the official EXIST 2026 Task 2.1 evaluation benchmark. The experiments were designed to assess the

Figure 5: Overview of the proposed multimodal ensemble framework. Textual and visual representations are extracted from meme content and combined with physiological features derived from Eye-Tracking, EEG, and Heart Rate signals. The fused representation is processed by an ensemble of Logistic Regression, LightGBM, and Random Forest classifiers to produce the final sexism prediction.

effectiveness of integrating textual, visual, demographic, and physiological information for subjective sexism identification in multilingual social media memes.

Table 3
Validation Performance of the Proposed Ensemble Framework

Metric	Value
Accuracy	0.7181
Macro F1 Score	0.6842
ROC-AUC	0.7470
PR-AUC	0.8322
Optimal Threshold	0.5675

The validation results demonstrate that the proposed ensemble framework achieves balanced performance across both classes despite the moderate class imbalance present in the dataset. The Macro F1-score of 0.6842 indicates consistent performance across both sexist and non-sexist classes, while the ROC-AUC and PR-AUC scores demonstrate the ability of the fused multimodal representation to effectively discriminate between the two categories. The optimized decision threshold of 0.5675 further improves the balance between precision and recall compared with the default threshold of 0.5, resulting in more reliable predictions under the subjective nature of the task.

Table 4
Official EXIST 2026 Task 2.1 Test Set Results

Submission	Rank	ICM-Hard	ICM-Hard Norm	F1 YES
SOFT_ALL	33	-0.3042	0.4511	0.9356
HARD_ALL	56	0.0890	0.5453	0.7210

Table 4 presents the official evaluation results obtained on the hidden EXIST 2026 test set. The SOFT_ALL submission achieved the highest ranking among the submitted runs, obtaining Rank 33 with

an F1 score of 0.9356 for the positive class. The HARD_ALL submission achieved improved ICM-Hard and normalized ICM-Hard scores, indicating that the proposed framework effectively captures annotator disagreement under the Learning with Disagreement evaluation setting. These results demonstrate that the ensemble framework generalizes beyond local validation while remaining competitive under both soft and hard evaluation protocols.

Several factors contribute to the effectiveness of the proposed framework. First, the combination of multilingual SentenceTransformer and CLIP embeddings enables the model to jointly capture contextual linguistic information and high-level visual semantics. Since the intended meaning of memes frequently depends on the interaction between text and images, integrating both modalities provides a richer multimodal representation than either modality alone.

Second, the incorporation of demographic and physiological descriptors gives complementary information regarding annotator responses that is unavailable in conventional multimodal datasets. Although these engineered features do not directly describe meme content, they provide useful signals associated with subjective perceptions of sexism and disagreement among annotators. Their integration with semantic representations allows the ensemble to model both the content of the meme and aspects of human interpretation.

Finally, the ensemble strategy combines the complementary strengths of linear and non-linear classifiers. Logistic Regression provides stable probabilistic predictions across the multimodal feature space, LightGBM captures complex feature interactions, and Random Forest improves robustness through diverse decision boundaries. The weighted soft-voting strategy integrates these additional predictions, contributing to improved generalization across different validation folds and the official benchmark evaluation.

Despite these encouraging results, several challenges remain. Meme interpretation is often influenced by cultural context, language, and individual experiences, making sexism identification inherently subjective. Furthermore, physiological measurements are naturally noisy and exhibit considerable variability across annotators. Future work may investigate richer physiological representations, multimodal transformer architectures capable of incorporating heterogeneous sensor information, and more comprehensive analyses of the contribution of individual modalities through ablation studies.

6. Conclusion

This paper presented a multimodal ensemble framework for subjective sexism identification in social media memes as part of EXIST 2026 Task 2.1. The proposed approach combines textual, visual, demographic and physiological information to model the complex and subjective nature of sexism perception. Textual representations were extracted using multilingual sentence embeddings, visual information was captured through CLIP image features, and physiological signals including EEG, Eye-Tracking, and Heart Rate measurements were incorporated through statistical feature engineering. These heterogenous representations were fused and processed using an ensemble of Logistic Regression, Random Forest, and LightGBM classifiers.

The experimental results demonstrate that integrating multiple modalities provides valuable information for sexism identification beyond conventional text-only or image-only approaches. The ensemble framework was able to capture both semantic content and annotator-centered signals, allowing it to better address disagreement and uncertainty within the dataset. The official evaluation results further indicate that incorporating soft-label information and multimodal representations can improve the detection of subjective sexist content while maintaining competitive performance on the benchmark.

The findings also highlight the importance of modeling annotator disagreement when addressing socially sensitive classification tasks. Since perceptions of sexism can vary across individuals, relying solely on hard binary labels may overlook important nuances present in the data. By leveraging physiological signals and disagreement-aware learning, the proposed framework aligns with the core

objectives of the EXIST benchmark and contributes toward more human-centered approaches for content moderation and social media analysis.

Future work will investigate systematic ablation studies to quantify the contribution of individual modalities, richer representations of physiological measurements, and multimodal transformer architectures capable of jointly modeling textual, visual, and physiological information. These directions have the potential to further improve disagreement-aware sexism identification while providing greater insight into the role of human-centered signals in subjective content moderation. Additionally, larger vision-language foundation models and more sophisticated physiological signal representations may further improve the understanding of subjective social phenomena in online environments. The proposed framework demonstrates the potential of combining multimodal and human-centered information for robust sexism identification in social media memes.

References

[1] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, P. Rosso, E. García-Arias, and 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.

[2] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, P. Rosso, E. García-Arias, and D. Spina, "Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media (Extended Overview)," in *CLEF 2026 Working Notes*, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, and J. M. Struß (Eds.), 2026.

[3] L. Plaza-del-Arco, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina, "Overview of EXIST 2024: Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes (Extended Overview)," in *Working Notes of CLEF 2024*, pp. 908–941, 2024.

[4] L. Plaza, J. Carrillo-De-Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, and R. Morante, "Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos (Extended Overview)," in *Working Notes of CLEF 2025*, pp. 1690–1735, 2025.

[5] G. Rizzi, D. Gimeno-Gómez, E. Fersini, and C.-D. Martínez-Hinarejos, "PINK at EXIST 2024: A Cross-Lingual and Multi-Modal Transformer Approach for Sexism Detection in Memes," in *Working Notes of CLEF 2024*, pp. 1177–1186, 2024.

[6] R. Pan, T. Bernal-Beltrán, J. A. García-Díaz, and R. Valencia-Garcia, "UMUTeam at EXIST 2025: Multimodal Transformer Architectures and Soft-Label Learning for Sexism Detection," in *Working Notes of CLEF 2025*, pp. 2084–2099, 2025.

[7] I. Arcos, "ArcosGPT at EXIST 2025: Identifying Sexism in Memes with Multimodal Deep Learning: Fusing Text and Visual Cues," in *Working Notes of CLEF 2025*, pp. 1801–1808, 2025.

[8] M. Guerrero-García, F. Carrillo-García, J. Mata-Vázquez, and V. Pachón-Álvarez, "I2C-UHU-Altair at EXIST 2025: Multimodal Sexism Detection and Classification Using Advanced Vision-Language Models BLIP2 and Qwen, Large Language Models, and Learning with Disagreement Frameworks," in *Working Notes of CLEF 2025*, pp. 1978–1993, 2025.

[9] Y.-Z. Fang, L.-H. Lee, and J.-D. Huang, "NYCU-NLP at EXIST 2024: Leveraging Transformers with Diverse Annotations for Sexism Identification in Social Networks," in *Working Notes of CLEF 2024*, pp. 1003–1011, 2024.

[10] G. Arampatzis, V. Perifanis, S. Symeonidis, and A. Arampatzis, "DUTH at EXIST 2025: Multilingual Sexism Detection with Soft Labels and Transformers," in *Working Notes of CLEF 2025*, pp. 1793–1800, 2025.

[11] Q. Chen, L. Kong, Y. Chen, and C. Sun, "Modeling Annotator Subjectivity for Sexism Detection on Social Media," in *Working Notes of CLEF 2025*, pp. 1848–1857, 2025.

[12] S. Fan, R. A. Frick, and M. Steinebach, "FraunhoferSIT@EXIST2024: Leveraging Stacking Ensemble Learning for Sexism Detection," in *Working Notes of CLEF 2024*, pp. 993–1002, 2024.

[13] S. R. H. Abidi, M. S. Khursheed, S. F. Sikandar, S. Zahra, F. Alvi, and A. Samad, "Sexism Identification in Tweets Using Ensembles and Augmentation: A Multilingual Approach," in *Working Notes of CLEF 2025*, pp. 1736–1751, 2025.

[14] A. Britez and I. Markov, "CLTL at EXIST 2025: Identifying Sexist Memes Using an Ensemble of Shallow and Transformer Models," in *Working Notes of CLEF 2025*, pp. 1828–1839, 2025.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[16] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.