SPECTRA at EXIST2026: Sexism Prediction and Evaluation of Multimedia Content Through Physiological and Gaze Responses Analysis

EXIST at CLEF 2026

Stefano Mandelli[1], Giulia Rizzi[1,*], Aurora Saibene[1], Francesca Gasparini[1] and Elisabetta Fersini[1]

[1]DISCo, Università degli Studi di Milano-Bicocca, Viale Sarca, 336, 20126 Milan, Italy

Abstract

Warning: This paper contains examples of language and images which may be offensive.

The proliferation of online hate speech and gender-based bias demands automated moderation systems that are increasingly robust and context-aware. While current language models excel at text analysis, they often fail to capture the inherently complex and multimodal nature of human cognitive and emotional responses. In this paper, we describe the system developed by team SPECTRA for our participation in the EXIST shared task. We present a multimodal fusion framework that integrates textual, visual, and physiological derived information to tackle the sexism and misogyny classification challenge. Our architecture leverages pretrained foundation models, namely BERT and CLIP, for text and image encoding, paired with a multi-layer perceptron to process physiological derived information. To effectively combine such heterogeneous modalities, we introduce a cross-attention mechanism where the textual representation acts as the primary semantic anchor to look up the other modalities. Experimental results show that text remains the dominant source of information for sexism detection, while the inclusion of physiological signals provides complementary insights into human responses to harmful content, but does not consistently improve predictive performance. In particular, the best-performing configuration relies on textual, visual, and automatically generated caption features, highlighting the challenges associated with aligning noisy physiological measurements with high-level semantic representations. These findings contribute to a better understanding of the opportunities and limitations of incorporating human-centered physiological information into multimodal sexism detection systems.

Keywords

Sexism detection, Multimodal Cross-Attention, Memes

1. Introduction

The computational detection of toxic behavior and offensive content in social media streams represents a core open challenge within the Natural Language Processing and Computational Social Science communities. Among these digital threats, sexism and misogynistic behavior present significant detection difficulties due to their multi-faceted nature, ranging from explicit verbal abuse to subtle, implicit, and heavily contextualized expressions. This complexity is further amplified by the modern social media landscape, where harmful content increasingly exploits multimodal formats, such as internet memes and short-form videos. While traditional detection systems have largely relied on text-based sources, they inherently suffer from a systemic inability to model the cross-modal interaction between textual claims and visual semantics, where the different modalities can either complement or reverse the meaning conveyed by each source. To systematically overcome these limitations, the

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

*Corresponding author.

s.mandelli28@campus.unimib.it (S. Mandelli); giulia.rizzi@unimib.it (G. Rizzi); aurora.saibene@unimib.it (A. Saibene); francesca.gasparini@unimib.it (F. Gasparini); elisabetta.fersini@unimib.it (E. Fersini)

https://www.unimib.it/giulia-rizzi (G. Rizzi); https://www.unimib.it/aurora-saibene (A. Saibene); https://www.unimib.it/francesca-gasparini (F. Gasparini); https://www.unimib.it/elisabetta-fersini (E. Fersini)

0009-0004-5340-9376 (S. Mandelli); 0000-0002-0619-0760 (G. Rizzi); 0000-0002-4405-8234 (A. Saibene); 0000-0002-6279-6660 (F. Gasparini); 0000-0002-8987-100X (E. Fersini)

sEXism Identification in Social neTworks (EXIST) shared task series has continuously defined the state-of-the-art in sexism detection.

In this edition, acknowledging that the perception of misogyny is intrinsically subjective and mediated by individual cognitive profiles, EXIST 2026 [1, 2] enriches traditional textual and visual inputs of memes and short videos with physiological derived information (from now on called physiological data for the sake of brevity), including the ones extracted from electroencephalographic (EEG) signals, and from eye-tracking patterns and heart rate, captured from a diverse pool of annotators.

In this paper, we present the computational solution developed by team SPECTRA for the EXIST 2026 shared task. We designed a multimodal framework to jointly model textual, visual, and physiological data related to a meme. Our architecture adopts state-of-the-art pre-trained models (BERT and CLIP) to extract textual and visual embeddings, coupled with a specialized neural layer to map physiological data into a shared latent subspace. In the proposed architecture, rather than employing symmetric fusion paradigms that overlook all the possible dependencies between modalities, we propose an asymmetric cross-attention mechanism, where the textual representation acts as the primary semantic anchor to selectively probe the joint visual and physiological contextual space. To assess the contribution of the different information sources, we evaluate three experimental configurations that combine textual, visual, caption-based, and physiological features under a unified architecture. Through this analysis, we investigate the effectiveness of human-centered physiological information for multimodal sexism detection and its interaction with conventional semantic and visual representations.

2. Related Work

Early methodologies in computational social science formalized sexism and misogyny detection as standard text classification tasks, heavily relying on surface-level lexical features and shallow neural networks designed to capture explicit hate speech [3, 4]. The field underwent a significant paradigm shift with the introduction of pretrained Transformer architectures, such as BERT [5] and RoBERTa [6], which enabled a more granular modeling of deep contextual nuances. Within the framework of the EXIST shared tasks, competitive architectures have increasingly progressed toward capturing complex linguistic variations [7, 8]. Recent editions have highlighted the necessity of leveraging Large Language Models (LLMs) [9] and advanced embedding spaces to untangle intersectional and subtle forms of online sexism across heterogeneous social platforms [8]. The effectiveness of these approaches has been confirmed in multilingual sexism and misogyny detection settings, where fine-tuned transformer models and ensemble LLM-based systems achieve state-of-the-art performance while improving the modeling of subtle and culturally dependent forms of harmful content [10]. To overcome the semantic boundaries of language models, recent research has explored human-centered approaches that integrate physiological signals, such as eye-tracking, heart-rate, and EEG measurements, directly into the modeling of misogynistic and sexist multimodal content [11]. Rather than viewing text as an isolated linguistic artifact, these approaches treat text consumption as an interactive cognitive process that triggers measurable reactions in the reader. Recent studies have demonstrated that physiological signals such as galvanic skin response and electrocardiographic measurements are effective indicators of emotional arousal and autonomic activation [12]. When integrated with transformer-based textual representations, these signals could offer a promising avenue for modeling human reactions to emotionally charged and potentially offensive content beyond linguistic information alone. By decoding these physiological markers alongside textual tokens, architectures can capture a more holistic representation of toxicity, mapping not only the syntactic configuration of the text but also its impact on human subjects. Furthermore, advanced neuro-computational paradigms have integrated central nervous system monitoring, such as EEG and eye-tracking patterns, to enhance the interpretability of diverse linguistic and visual cues [13, 14]. In tasks like the detection of subtle sexism, structural indicators often remain neutral, leading to high classification error rates in standard language models. By alignment-tuning textual representations with information such as gaze duration, which captures cognitive load and processing delays, models can consider implicit human signals. This integration

fundamentally redefines sexism detection from a static text labeling task to a dynamic, reader-centric profiling paradigm. Consequently, purely textual approaches remain fundamentally constrained when compared with the pragmatic shifts typical of online discourse, emphasizing the need for architectures that bridge the gap between digital text and human experiences. In this direction, [11] demonstrated that the perception of sexist memes is accompanied by distinctive cognitive and physiological responses that can be exploited for automated sexism detection. Their experimental analysis showed that sexist content induces a higher cognitive load, reflected by increased fixation counts and longer reaction times, while simultaneously producing significant variations in EEG spectral power across the Alpha, Beta, and Gamma bands. These findings suggest that human reactions to harmful content encode valuable information that extends beyond the semantic information available in text and images. Motivated by this observation, the authors proposed a hierarchical attention-based multimodal architecture that fuses textual and visual features with eye-tracking, heart-rate, and EEG signals. Their results revealed consistent improvements in the detection of subtle and fine-grained forms of sexism, supporting the hypothesis that physiological cues provide a complementary perspective on how harmful content is perceived and interpreted by users.

3. Dataset and Task Description

3.1. Task Description

The *sEXism Identification in Social neTworks* task at CLEF 2026 [1, 2] aims to address the automatic identification and characterization of sexism across social networks. Building upon previous editions [7, 8], EXIST 2026 takes a major step forward by enriching the traditional annotation labels with physiological and behavioral responses. Within this framework, traditional text- or image-based labeling is enriched with physiological data (including eye tracking, heart rate, and EEG) to model both conscious and unconscious human responses to potentially sexist content.

The challenge continues to embrace the Learning with Disagreements (LeWiDi) paradigm, recognizing the inherent subjectivity and ambiguity involved in identifying sexist behavior. For both memes and videos, three distinct experimental subtasks are proposed, addressing (i) sexism identification, (ii) source intention detection, and (iii) sexism categorization. This paper focuses solely on **Subtask 2.1**, which is a binary classification task dedicated to sexism identification in memes. The objective is to determine whether a given meme describes a sexist situation or behavior, or does not.

3.2. Dataset

The multimedia corpus provided for EXIST 2026 integrates and extends data from previous iterations into a novel experimental setting. For the meme track utilized in Subtask 2.1, the collection features a balanced distribution of English and Spanish memes: 5,037 labeled memes, split into 3,984 samples for training and 1,053 samples for testing. Examples of memes from the dataset are reported in Figure 1.

(a) Non-Sexist (En) (b) Sexist (En) (c) Non-Sexist (Es) (d) Sexist (Es)

Figure 1: Examples of memes in the EXIST meme dataset, showcasing the inherent challenges of the task. En and Es refer to English and Spanish, respectively.

To accommodate the LeWiDi paradigm, the dataset includes soft annotations derived from a pool of human annotators, six for each meme, alongside detailed socio-demographic profiles (e.g., age, gender, ethnicity, level of study, and country). As previously reported, the dataset presents derived information from physiological signals collected continuously while participants were exposed to the stimuli. Therefore, these parameters, comprising 108 features per subject, were not provided as continuous instantaneous measurements, but as aggregated descriptors over the stimulus viewing interval. Specifically, they include 24 metrics for Eye-Tracking (ET), such as pupil diameter, blink counts, fixation counts, saccades counts, and reaction time; four Heart-Rate (HR) statistics covering mean, standard deviation, minimum, and maximum values; and 80 EEG features derived from bandpower measurements across 16 channels for the Delta, Theta, Alpha, Beta, and Gamma frequency bands. To maintain data consistency, two to four individuals were exposed to each meme, ensuring at least two recordings for each modality. Consequently, this representation was intended to enable models to exploit implicit cognitive responses for sexism detection.

4. Proposed Approach

4.1. Model Architecture

We propose a tri-modal architecture to combine textual, visual, and physiological data for binary classification. The architecture leverages pretrained foundation models for language and vision alongside a dedicated multi-layer perceptron (MLP) for physiological data. Integration across modalities is achieved via an asymmetric cross-attention mechanism.

We then evaluated this architecture on three distinct input configurations, each corresponding to one of our officially submitted runs to **Subtask 2.1**. The overall architecture is illustrated in Figure 2.

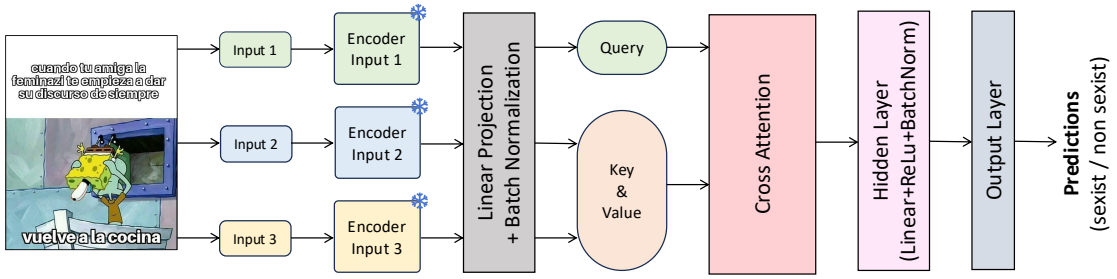

Figure 2: Overview of our architecture. Frozen pre-trained encoders extract modality-specific features, which are normalized and linearly projected into a common embedding space. A Multihead Cross-Attention module combines the modalities by using the first input as Query against a Key-Value context built from the second and the third features. The resulting representation feeds a non-linear dense layer followed by a linear classifier for binary prediction.

4.1.1. Inputs

The diagram omits explicit input labels since our three submitted runs each combined a different subset of four modality-specific features extracted or related to the memes.

- **Text Feature:** it denotes the raw textual content T of each meme, provided directly by the dataset.
- **Image Feature:** it represents the visual content I of the meme, converted to RGB format and preprocessed via the CLIP processor, which handles resizing and pixel-level normalization for compatibility with the pre-trained visual encoder.

- **Physiological Feature:** it denotes the 108-dimensional vector $\mathbf{x}_p \in \mathbb{R}^{108}$ aggregating the physiological data of the users who viewed the meme, including 24 ET features, 4 HR features, and 80 EEG features. During preprocessing, temporal metrics were converted into standard units (seconds) to ensure scale consistency. To address baseline variability among individuals, physiological features were standardized using user-specific z-score normalization; in cases where historical user data was insufficient, global dataset statistics were applied as a fallback. Finally, the normalized physiological vectors for each subject were then averaged, resulting in a unified 108-dimensional representation for each meme sample.
- **Caption Feature:** To provide the model with richer visual context, we augmented the input with image captions automatically generated by a Prompt-based Image Captioning model, namely Qwen2.5-V [15]. To generate the image caption C of each meme, we adopted the strategy performed by Italiani et al. [16] in the previous edition of the same shared task, and further explored and detailed as Prompt B in a recent follow-up work [17].

4.1.2. Encoders

Given an input sample consisting of text T, an image I, and a physiological data vector $\mathbf{x}_p \in \mathbb{R}^{108}$, we extract independent dense representations through modality-specific encoders. In particular:

- **Text Encoder:** we used a pretrained BERT-base model to encode the textual sequence within each meme. Given the input tokens, the sequence is passed through a BERT-based encoder, the final hidden state of the classification token (`[CLS]`) is used to represent the global semantic context of each meme:
$$\mathbf{h}_t = \text{BERT}(T) \in \mathbb{R}^{768} \tag{1}$$
In our experiment, we employed XLM-RoBERTa[1] fine-tuned on misogyny and sexism detection tasks.
- **Visual Encoder:** The visual features are extracted using the vision backbone of a pretrained CLIP model (ViT-B/32). The image is mapped to a joint embedding space, and we leverage the pooled representation output from the final layer:
$$\mathbf{h}_v = \text{CLIP}(I) \in \mathbb{R}^{512} \tag{2}$$
In our experiment, we used the CLIP vision backbone[2].
- **Physiological Encoder:** unlike text and vision, physiological data lack high-level spatial or sequential priors in this context. We therefore implement a lightweight MLP with a Rectified Linear Unit (ReLU) activation to project the raw physiological vectors $\mathbf{x}_p$ into a dense embedding space:
$$\mathbf{h}_p = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{x}_p + \mathbf{b}_1) + \mathbf{b}_2 \tag{3}$$
where $\mathbf{h}_p \in \mathbb{R}^d$, and d denotes the target common hidden dimension.
- **Caption Encoder:** the caption C of each meme has been encoded using the Multilingual BERT [3] to obtain its latent representation:
$$\mathbf{h}_c = \text{BERT}(C) \in \mathbb{R}^{768} \tag{4}$$

To prepare the feature representations for cross-attention, we project the obtained embeddings into the same hidden dimension $\mathbb{R}^k$. In particular, let $\mathbf{W}_t$, $\mathbf{W}_v$, $\mathbf{W}_p$, $\mathbf{W}_c$ be learnable linear projection matrices. We project each hidden state $\mathbf{h}_t$, $\mathbf{h}_v$, $\mathbf{h}_p$ and $\mathbf{h}_c$ into the common space as follows:

$$\mathbf{z}_t = \mathbf{W}_t \mathbf{h}_t, \quad \mathbf{z}_v = \mathbf{W}_v \mathbf{h}_v, \quad \mathbf{z}_p = \mathbf{W}_p \mathbf{h}_p, \quad \mathbf{z}_c = \mathbf{W}_c \mathbf{h}_c \tag{5}$$

[1] annahaz/xlm-roberta-base-misogyny-sexism-indomain-mix-bal
[2] openai/clip-vit-base-patch32
[3] google-bert/bert-base-multilingual-cased

where each projection block is then followed by 1D Batch Normalization, which helps to stabilize the distribution of each modality and facilitates faster convergence during the subsequent cross-modal training.

As further specified in Subsection 4.2, the target embedding dimensionality varies across runs: Run 1 and Run 3 project into a 256-dimensional space, while Run 2 uses 108 dimensions to match the physiological input feature space.

4.1.3. Multimodal Cross-Attention

To capture the interactions between the three different inputs, we adopted a Multihead Cross-Attention mechanism. Rather than employing a symmetric fusion strategy, such as concatenation, we adopted an asymmetric fusion paradigm where the primary semantic anchor (the text of the meme) queries the auxiliary contextual modalities, i.e. the visual features combined with either the generated captions or the physiological derived information.

We defined a unified context tensor $\mathbf{F}$ by concatenating the visual and physiological or caption representations along the sequence dimension:

$$\mathbf{F} = [\mathbf{z}_v \,;\, \mathbf{z}_{p/c}] \tag{6}$$

We then apply Multi-Head Attention with A attention heads. The textual embedding $\mathbf{z}_t$ serves as the **Query (Q)**, while the visio-physiological or visio-caption tensor $\mathbf{F}$ serves as both the **Key (K)** and **Value (V)**:

$$\text{Attn}(\mathbf{z}_t, \mathbf{F}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_c}}\right)\mathbf{V} \tag{7}$$

$$\mathbf{Q} = \mathbf{z}_t\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{F}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{F}\mathbf{W}^V \tag{8}$$

To stabilize training, a residual connection and Layer Normalization (LN) are applied to the attention output to yield the final multimodal representation $\mathbf{z}_f$:

$$\mathbf{z}_f = \text{LN}(z_t + \text{Attn}(\mathbf{z}_t, \mathbf{F})) \tag{9}$$

This asymmetric design allows the model to selectively attend to the most relevant visual and auxiliary signals, conditioned on the semantic content of the text.

4.1.4. Classification Head

Following the multimodal fusion, the attended representation is squeezed along the sequence dimension to derive a single context-aware vector. This resulting vector is subsequently input into a MLP classification head. The classification module incorporates a fully connected hidden layer that transforms the embedding into a 64-dimensional space, followed by Batch Normalization and a ReLU activation function to improve the training stability. A final linear layer then maps the 64-dimensional representation to a 2-dimensional output, providing the logits used for the binary classification task (sexist vs non sexist). More formally, the final multimodal representation $\mathbf{z}_f$ is passed through a final binary classification head, which consists of a linear layer, a ReLU non-linearity, and a projection to the action space logits:

$$\mathbf{y} = \mathbf{W}_4 \cdot \text{ReLU}(\mathbf{W}_3\mathbf{z}_f + \mathbf{b}_3) + \mathbf{b}_4 \tag{10}$$

where $\mathbf{y} \in \mathbb{R}^2$ represents the raw logits for the binary targets. The estimation of the parameter set is determined by minimizing a standard Cross-Entropy loss function.

4.2. Submitted Runs

During the training phase, text, visual, and, if used in the run, caption encoders were kept frozen to preserve their pre-trained representations and prevent forgetting. The modules were optimized for a maximum of *20 epochs* using the Adam optimizer with a learning rate of 1×10^{-4} and a *batch size of 8*, minimizing the standard Cross-Entropy Loss. To ensure robust evaluation and mitigate overfitting, we employed a 10-Fold Cross-Validation strategy paired with an Early Stopping mechanism *(patience 2)*. The data was shuffled and partitioned into *10 folds* using a deterministic seed, ensuring complete reproducibility across experiments.

Building on the unified architecture described in the previous section, we submitted three distinct experimental configurations to the challenge. While the core structure remains identical across all runs, they differ in their auxiliary input modality and embedding dimensionality. The three submitted runs are detailed below:

- **SPECTRA_1.** This configuration relies exclusively on the semantic and visual properties of the memes, without considering any physiological data. The system processes three primary inputs, which consist of the Text Feature, Image Feature, and Caption Feature. Each input is provided to the corresponding encoder, i.e., the Text Encoder, Visual Encoder, and Caption Encoder. The obtained embeddings are then projected into a 256-dimensional space. For the cross-attention setup, the Multihead Cross-Attention module is configured with 8 attention heads. Within this architecture, the textual embedding serves as the Query, dynamically attending over a Key-Value context formed strictly by concatenating the visual and caption features.
- **SPECTRA_2.** This configuration uses the Text Feature, Image Feature, and Physiological Feature as its main inputs. These are processed using a Text Encoder, a Visual Encoder, and a Physiological Encoder. The model maps these representations into a 108-dimensional shared embedding space. The Multihead Cross-Attention module is configured with 6 attention heads. In this setup, the textual embedding serves as the Query, dynamically attending over a Key-Value context formed strictly by concatenating the visual and physiological features.
- **SPECTRA_3.** This configuration builds upon Run 2 by employing the same physiological data, but mapping it into a wider representational space to allow the network a higher capacity for feature alignment. The input stream remains identical, using the Text Feature, Image Feature, and Physiological Feature, which are handled by the Text Encoder, Visual Encoder, and Physiological Encoder, respectively. However, the shared embedding space is expanded to 256 dimensions. Consequently, the Multihead Cross-Attention module is adjusted to use 8 attention heads. Like the previous configurations, the textual embedding serves as the Query, dynamically attending over a Key-Value context formed strictly by concatenating the visual and physiological features.

5. Results & Discussion

This section provides the quantitative evaluation of our proposed architecture across the three submitted runs. We begin by presenting internal performance metrics obtained through the 10-Fold Cross-Validation on the training set, followed by the official challenge results.

5.1. Results on the training dataset

Table 1 presents the aggregated validation metrics for Run 1, Run 2, and Run 3, on the training dataset using a 10-fold cross-validation strategy.

To conduct hyperparameter optimization and model selection, we used a validation set from the official training dataset. Upon identifying our optimal architectural configurations and hyperparameters, we retrained our final models on the complete official training set prior to submission. The results presented and discussed in Section 5.2 also evaluate these optimized models on the official, unreleased test set containing 1,053 instances.

Table 1
Aggregated 10-Fold Cross-Validation results on the training set.

System	Accuracy	F1-Score
Run 1	0.652	0.663
Run 2	0.647	0.666
Run 3	0.642	0.653

Among the three configurations, Run 1 seems to perform better than the others, suggesting that the generated captions provided a semantically coherent auxiliary signal that the model could effectively leverage. Run 2 obtained a marginally higher F1-Score, but the overall performance of the physiological runs remained slightly below expectations during internal validation. This gap may be attributed to the inherently noisy and high-variance nature of physiological signals, which likely introduced alignment difficulties during cross-modal training compared to the more structured textual representations.

5.2. Results on the test dataset

Besides the internal validation results, we present the official test set scores provided by the EXIST 2026 organizers in Tables 2 and 3.

The challenge evaluates participating systems under two distinct paradigms based on the LeWiDi framework: **Hard-Hard** and **Soft-Soft**. In this naming convention, the first term designates the format of the system output, while the second refers to the ground truth used for evaluation. Specifically, in the **Hard-Hard** setting, the system predicts a discrete class label, which is compared against a deterministic gold-standard derived from a majority threshold over the annotators. In this case the official evaluation measure used is the ICM metric. Conversely, in the **Soft-Soft** setting, the system predicts a probability distribution over the categories, which is evaluated against the full distribution of human annotator judgements to measure how well the model captures actual annotator variability. In this case the ICM-soft is the official metric used. Consistent with the setup of previous meme classification tracks in this challenge, no predefined validation set was officially provided.

To provide context for the presented results, the organizers provided three reference systems. The *Gold* baseline represents a model that achieves perfect predictions, establishing an upper bound for the ICM score. The *Majority Class* and *Minority Class* baselines are non-informative lower bounds that assign all instances to the majority or minority class, respectively. This allows us to perform sanity checks to confirm that systems submitted capture meaningful patterns rather than simply reflecting biases in class distribution.

Table 2
Official EXIST 2026 results of Task 2.1 under Hard evaluation.

System	ICM	ICM-Norm	F_1-Score
$EXIST_{2026}$ (Gold)	0.9832	1.0000	1.0000
SPECTRA_1	**-0.0737**	**0.4625**	**0.6462**
SPECTRA_2	-0.1488	0.4243	0.6057
SPECTRA_3	-0.2680	0.3637	0.5551
$EXIST_{2026}$ (Majority)	-0.4038	0.2947	0.6821
$EXIST_{2026}$ (Minority)	-0.6468	0.1711	0.0000

As detailed in Tables 2 and 3, SPECTRA_1 emerges as the most robust configuration compared to the other two runs, aligning with the trends identified during internal cross-validation. The gap can be attributed to the distinct characteristics of the modalities.

In our cases, generated captions offer clear, high-level semantic information that naturally integrates with the textual input space, enabling effective cross-attention fusion. In contrast, physiological data such as EEG, HR, and ET are intrinsically noisy, exhibit significant inter-individual variability, and

Table 3
Official EXIST 2026 results of Task 2.1 under Soft evaluation.

System	ICM-soft	ICM-soft-Norm	Cross Entropy
$EXIST_{2026}$ (Gold)	3.1107	1.0000	0.5852
SPECTRA_1	**-0.4185**	**0.4327**	**0.9702**
SPECTRA_2	-0.7218	0.3840	1.0303
SPECTRA_3	-0.8861	0.3576	1.2022
$EXIST_{2026}$ (Majority)	-2.3568	0.1212	4.4015
$EXIST_{2026}$ (Minority)	-3.5089	0.0000	5.5672

exist in a representational space far removed from the linguistic domain. This semantic-physiological disparity likely impeded effective multimodal alignment, thereby restricting the physiological runs capacity to generalize effectively to unseen test data.

6. Conclusion

In this paper, we presented the multimodal framework developed by team SPECTRA for the participation in the EXIST 2026 shared task. Our system is based on an asymmetric cross-attention designed to integrate text and images content with human physiological signals.

The empirical evaluation provided critical insights into the integration of implicit human cognitive traits for automated sexism classification. While the incorporation of physiological data, i.e., EEG, heart rate, and eye-tracking metrics, was intended to serve as a relevant contextual anchor to mitigate linguistic ambiguity, both our 10-fold cross-validation and the official challenge evaluations demonstrated that it did not boost classification performance. Instead, our baseline configuration, which relied exclusively on textual, visual, and generated caption features, consistently emerged as the most robust architecture. These findings point to a substantial semantic-physiological disparity. While high-level textual summaries and image captions naturally align within the linguistic embedding domain, physiological sensor data are intrinsically noisy, characterized by elevated inter-individual variance, and reside in a representational space far removed from explicit language semantics. This misalignment heavily restricts the capacity of physiological features to generalize effectively to unreleased test streams.

Acknowledgments

This work was partially supported by the MUR under the grant “Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and Gemini in order to: check grammar and spelling, paraphrase, and reword sentences for clarity. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] 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.

[2] 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.
[3] E. Shushkevich, J. Cardiff, Automatic misogyny detection in social media: A survey, Computación y Sistemas 23 (2019) 1159–1164.
[4] W. Lei, N. A. S. Abdullah, S. R. S. Aris, A systematic literature review on automatic sexism detection in social media, Engineering, Technology & Applied Science Research 14 (2024) 18178–18188.
[5] 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.
[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[7] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Proceedings of CLEF 2024, 2024.
[8] 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: Proceedings of CLEF 2025, 2025, pp. 266–289.
[9] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, AI-UPV at EXIST 2023: Sexism Characterization Using Large Language Models Under the Learning with Disagreements Regime, in: CLEF 2023 Working Notes, 2023.
[10] E. Hashmi, et al., Enhancing misogyny detection in bilingual texts using explainable ai and multilingual fine-tuned transformers, Complex & Intelligent Systems 11 (2025) 39.
[11] I. A. Gabaldón, P. Rosso, E. G. Vicent, Human-Centered Multimodal Fusion for Sexism Detection in Memes with Eye-Tracking, Heart Rate, and EEG Signals, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 9830–9840. doi:10.63317/2w2vaanu3pe7.
[12] A. F. Bulagang, N. G. Weng, J. Mountstephens, J. Teo, A review of recent approaches for emotion classification using electrocardiography and electrodermography signals, Informatics in Medicine Unlocked 20 (2020) 100363.
[13] N. Hollenstein, M. Troendle, C. Zhang, N. Langer, ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020), European Language Resources Association (ELRA), Marseille, France, 2020, pp. 138–146. URL: https://aclanthology.org/2020.lrec-1.18/.
[14] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, N. Langer, Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading, Scientific Data 5 (2018) 180291.
[15] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al., Qwen3-vl technical report, arXiv preprint arXiv:2511.21631 (2025).
[16] P. Italiani, F. Maqbool, D. Gimeno-Gómez, E. Fersini, C.-D. Martínez-Hinarejos, TrankilTwice at EXIST2025: detecting sexism in memes under multi-lingual settings, in: CLEF 2025 Working Notes, 2025.
[17] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, MemeWeaver: Inter-meme graph reasoning for sexism and misogyny detection, in: V. Demberg, K. Inui, L. Marquez (Eds.), Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 2120–2134. URL: https://aclanthology.org/2026.findings-eacl.111/. doi:10.18653/v1/2026.findings-eacl.111.