Human-Centered Multimodal Fusion for Sexism Detection in Memes with Eye-Tracking, Heart Rate, and EEG Signals

Iván Arcos[1], Paolo Rosso[1,2], Elena Gomis-Vicent[1]
[1]PRHLT Research Center, Universitat Politècnica de València (UPV), 46022 Valencia, Spain
iarcgab@etsinf.upv.es, prosso@dsic.upv.es, egomvic@upvnet.upv.es
[2]ValgrAI – Valencian Graduate School and Research Network of Artificial Intelligence, 46022 Valencia, Spain

Abstract

The automated detection of sexism in memes is a notoriously challenging task due to multimodal ambiguity, cultural nuance, and the use of humor to provide plausible deniability. As a result, content-only models often fail to capture the complexity of human perception. To address this fundamental limitation, we introduce and validate a human-centered paradigm that augments standard content features with rich physiological data. We created a novel resource by recording Eye-Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG) from 16 subjects (8 per experiment) while they viewed 3984 memes from the EXIST 2025 dataset. Our statistical analysis reveals significant physiological differences in how subjects process sexist versus non-sexist content. Sexist memes were associated with higher cognitive load (evidenced by increased fixation counts and longer reaction times), and with differences in EEG spectral power across the Alpha, Beta, and Gamma frequency bands. This pattern, commonly linked in previous research to increased attentional engagement and cognitive effort during visual processing, suggests that sexist memes may elicit more demanding neural activity compared to non-sexist ones. Building on these findings, we propose a novel multimodal fusion model that integrates these physiological signals with enriched textual-visual features derived from a Vision-Language Model (VLM). Our final model achieves an AUC of 0.794 in binary sexism detection, a statistically significant 3.4% improvement over a powerful VLM-based baseline. The fusion of physiological data proves particularly effective for nuanced and ambiguous cases, boosting the F1-score for the most challenging fine-grained category, *Misogyny and Non-Sexual Violence*, by an unprecedented 26.3%. Our work demonstrates that human physiological responses provide a robust, objective signal of perception that can significantly enhance the accuracy and human-awareness of automated systems for countering online sexism.

Keywords: sexism detection, memes, physiological signals, eye-tracking, electroencephalography, heart rate, multimodal fusion, vision–language and physiological resources

1. Introduction

The proliferation of social media has reconfigured communication, but it has also created a fertile environment for the dissemination of harmful content. Online sexism and misogyny have become pervasive phenomena that normalize discriminatory attitudes, silence the voices of women, and cause tangible psychological harm (United Nations Entity for Gender Equality and the Empowerment of Women (UN Women), 2025; Dehingia, 2020). The scale of the problem is staggering: a 2025 Amnesty UK report revealed that 73% of Gen Z users have witnessed misogynistic content online, with TikTok being the primary vector (70% of respondents) (Amnesty International UK, 2025). This abuse is not trivial and leads to lowered self-esteem and self-censorship (Swan, 2025), constituting a significant human rights issue in the digital age.

Within this ecosystem, memes have emerged as a key medium of communication, but also as a vehicle for propagating sexist ideologies (Issac, 2018). The challenge for their automatic detection lies in their unique semiotic structure: their meaning emerges from the multimodal interplay between text and image and is often shrouded in humor, irony, or satire (Argüello-Gutiérrez et al., 2022). This strategic ambiguity creates "plausible deniability," allowing creators and sharers of harmful content to evade accountability by claiming it was "just a joke." This feature makes sexist memes a particularly difficult problem for artificial intelligence systems (Singh et al., 2023).

Current computational models, even state-of-the-art ones, systematically fail when faced with this challenge. The "Hateful Memes Challenge" by Facebook AI demonstrated that cutting-edge multimodal models performed little better than random chance when classifying memes where the offensive nature arose from the combination of individually non-offensive text and images (Kiela et al., 2020). This failure highlights a fundamental gap: "sexism" is not an intrinsic property of the content but an emergent phenomenon shaped by human interpretation. Human perceptual and affective responses provide complementary information that can help models capture the nuanced and context-dependent nature of sexist content more effectively (Khan et al., 2025).

This work proposes a human-centered pivot: instead of training an AI model to replicate subjective labels, we propose teaching it to recognize unconscious physiological responses that manifest in a human who perceives a given content as

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9830–9840
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

sexist. Physiological and neural reactions—such as gaze patterns (Bradley et al., 2008), heart rate variability (García Martínez et al., 2024), or brain activity (Quan et al., 2021)—are objectively measurable and less ambiguous than conscious interpretations. This study leverages physiological data from eye-tracking (ET), heart rate (HR), and electroencephalography (EEG) as a more reliable proxy for the human experience, addressing the following research questions:

RQ1 *Are physiological responses—eye-tracking, heart rate, and electroencephalography—significantly different between sexist and non-sexist memes, and across distinct types of sexism?*

RQ2 *Are there statistically significant differences in performance among unimodal classifiers based solely on (a) textual, (b) visual, and (c) physiological-derived features when detecting and categorizing sexist memes?*

RQ3 *Can the incorporation of physiological data into a multimodal fusion model (text + image) yield a significant improvement in classification performance over a baseline text+image model?*

Contributions

1. **A new resource.** A multimodal dataset built from 3984 memes with rich physiological responses from 16 subjects (8 per experiment) across two large-scale studies: **Experiment 1** recorded **ET and HR** responses (**7782** recordings), while **Experiment 2** recorded **EEG and HR** responses (**7714** recordings), totaling **15496** instances.

2. **Physiological evidence.** An analysis showing statistically significant differences in neurophysiological responses across sexism types and categories, highlighting aspects that remain challenging for content-only systems and providing insights into their current limitations.

3. **A fusion model.** A hierarchical attention architecture that integrates physiological data with enriched content features, yielding significant, robust improvements over strong baselines on multiple sexism-detection tasks, with the largest benefits on nuanced, ambiguous cases.

2. Related Work

The automatic detection of harmful content is a well-established field of research. It is crucial to distinguish between **hate speech**, a broad concept (Nockleby, 2000); **misogyny**, which involves hatred or dislike of women; and **sexism**, which refers to prejudice or discrimination based on sex and can manifest subtly. Misogyny lies at the intersection of hate speech and sexism, sharing characteristics of both.

Sexism detection has evolved significantly, moving from textual analysis to multimodal and multilingual approaches, driven by shared tasks such as task 10 at SemEval-2023 (Kirk et al., 2023) and, more recently, the EXIST lab series (Plaza et al., 2025), which has expanded its focus to include memes and videos. Methodologies have progressed through several stages:

1. **Statistical Learning:** Early approaches used features like TF-IDF and N-grams with classical models such as SVM and Logistic Regression (Jha and Mamidi, 2017; Pelle and Moreira, 2017).

2. **Word Embedding-Based:** Later, neural networks (CNN, LSTM) fed with pre-trained embeddings (Word2Vec, GloVe) improved the capture of semantic context (Gasparini et al., 2018; Grosz and Conde-Cespedes, 2020).

3. **Pre-trained Language Models (PLMs):** The advent of Transformers (BERT, RoBERTa) marked a qualitative leap, enabling deep contextual understanding (Parikh et al., 2021). Multimodal models (e.g., Visual-BERT) began to more effectively integrate textual and visual information (Rizzi et al., 2023; Arcos and Rosso, 2024).

4. **Large Language Models (LLMs):** The most recent models, like ChatGPT and Llama, are used for their advanced reasoning capabilities, although their performance on specific classification tasks can vary compared to fine-tuned PLMs (Li et al., 2024; Abercrombie et al., 2024).

Despite these advances, key challenges persist. Models often lack generalizability to unseen data (Samory et al., 2021) and suffer from **biases** inherited from training data. Interpretability remains a problem, as most models function as "black boxes," making it difficult to understand their decisions.

In parallel, a line of research has emerged that integrates physiological data into Natural Language Processing (NLP) to capture human cognitive and affective responses.:

- **Eye-Tracking** has been used to understand cognitive load and attention, showing that gaze patterns shift when encountering complex, ambiguous, or emotionally salient content (Khurana et al., 2023; Kar et al., 2025).

- **Heart Rate** provides a general indicator of physiological arousal and stress. In our study, HR was continuously recorded to obtain descriptive measures of cardiac activity during meme-viewing, allowing temporal alignment with EEG and eye-tracking data (Azarbarzin et al., 2014).

- **Electroencephalography** offers a direct, high-temporal-resolution measurement of brain activity, allowing for the identification of neural correlates of semantic and emotional processing in real time (Lin and Yang, 2024).

While these sensors have been used in NLP, they have not addressed the specific ambiguity of sexist memes. We treat physiological responses as complementary content-level signals: eye-tracking, heart rate, and electroencephalography features are aggregated across viewers and aligned with the *majority* EXIST labels per meme, in order to make the model learn generalized patterns linked to consensually sexist content. These signals do not replace human annotations; they help disambiguate cases where text–image cues are insufficient, in line with recent work aligning models with human perceptual and cognitive signals (Zermiani et al., 2024; Hollenstein et al., 2020). However, this approach also entails important limitations: models may encounter conflicting information when physiological reactions from a small number of subjects diverge from the majority labels, introducing potential noise and reducing the reliability of aggregated physiological patterns. For instance, Zermiani et al. (2024) introduced *InteRead*, a 50-subject eye-tracking dataset of real-world texts with systematically triggered interruptions, providing fine-grained annotations of gaze behavior and resumption lags. Their analyses confirmed that interruptions, as well as lexical variables such as word length and frequency, significantly shape reading dynamics, highlighting gaze as a robust indicator of cognitive load. Similarly, Hollenstein et al. (2020) released *ZuCo 2.0*, which records simultaneous EEG and eye-tracking during both natural reading and annotation tasks, thereby capturing neural and ocular correlates of comprehension and task-specific processing. Together, these resources demonstrate the feasibility and scientific value of integrating physiological signals into NLP pipelines, underscoring the timeliness of our proposal to incorporate ET, HR, and EEG into sexism detection models, moving beyond purely textual or visual features.

3. A Multimodal Resource of Sexist Meme Perception

To build and test our human-centered models, we created a new multimodal resource by collecting physiological responses to existing, expertly annotated sexist and non-sexist memes.

Table 1: EXIST 2025 Meme Dataset Statistics (TRAIN). Percentages for T2 and T3 are calculated over the sexist subset of each language.

Task	Label	Spanish (N, %)	English (N, %)
T1: Sexism ID	Sexist	1040 (52.6)	963 (48.0)
	Non-Sexist	626 (31.6)	741 (37.0)
	Ties	313 (15.8)	301 (15.0)
T2: Intention	Direct	731 (70.3)	575 (59.7)
	Judgmental	179 (17.2)	226 (23.5)
	Ties	130 (12.5)	162 (16.8)
T3: Category	Ideological Ineq.	290 (27.9)	296 (30.7)
	Stereotyping	249 (23.9)	202 (21.0)
	Objectification	228 (21.9)	212 (22.0)
	Sexual Violence	77 (7.4)	60 (6.2)
	Misogyny NSV	32 (3.1)	19 (2.0)
	Ties	164 (15.8)	174 (18.1)
Total Memes		**1979**	**2005**

3.1. EXIST 2025 Memes

We used the official training split of the EXIST 2025 shared task (Plaza et al., 2025), a publicly available dataset containing 3984 memes (1979 in Spanish, and 2005 in English). In EXIST, sexist memes are further categorized by communicative intent: **Direct sexism** refers to content that itself expresses or promotes sexist ideas, whereas **Judgmental sexism** describes content that portrays or condemns sexist situations or behaviors. The memes were annotated in EXIST for three hierarchical tasks, with statistics summarized in Table 1.

Each meme was annotated by six annotators, revealing a clear trend in subjectivity. While **Task 1** (binary sexism detection) achieved fair agreement (Fleiss' $\kappa = 0.282$), the consensus dropped significantly for more nuanced judgments. For **Task 2** (source intention), agreement was merely slight ($\kappa \approx 0.19$). This subjectivity was most pronounced in **Task 3** (fine-grained categorization), where the *Misogyny and Non-Sexual Violence* category showed the lowest agreement of all ($\kappa = 0.103$). This high degree of human disagreement on the content's meaning strongly motivates our physiological approach, which seeks a more objective signal of perception. For our models, we used the majority-vote labels provided by the task organizers, and all experiments reported in this paper were conducted on the training partition.

3.2. Physiological Data Collection

We conducted two experiments with a total of 16 subjects of different nationalities, with a gender-balanced distribution, in their 20s and 30s. This age range was considered particularly relevant, as younger populations have been reported to exhibit greater tolerance toward sexist content (Plaza et al., 2024), enabling the investigation of their unconscious cognitive and emotional reactions to such content. Each of the 3984 memes was viewed by at least two subjects.

Procedure. Subjects were seated comfortably while memes were displayed on a screen until they provided a response, followed by a 3-second pause before the next stimulus to prevent overlap. In each session ($\approx$ 45 minutes), subjects viewed 100-170 memes. After each meme, they answered control questions about its content to ensure engagement and comprehension. All subjects gave their consent to use the data anonymously for research purposes.

We used the following devices to collect physiological data during the experiments:

- **Eye-Tracking:** Pupil Labs Neon glasses recorded binocular gaze data at 200 Hz.
- **Heart-Rate:** A Garmin Venu 3 watch continuously measured inter-beat intervals.
- **Electroencephalography:** A 16-channel OpenBCI Cyton Ultracortex Mark IV headset recorded neural activity from 16 scalp locations (10–20 system) at 250 Hz.

3.3. Physiological Feature Extraction

From the raw physiological data, we extracted a comprehensive feature set for each meme-viewing instance, structured by experimental setup.

Experiment 1 (ET)

- **ET:** For each oculomotor event (fixations, blinks, pupil diameter), we computed summary statistics (mean, standard deviation, minimum, maximum, count) capturing detailed patterns of visual attention.

Experiment 2 (EEG)

- **EEG:** Signals were acquired concurrently with HR. Preprocessing included: (1) conversion to μV; (2) zero-phase 4th-order Butterworth band-pass filter (0.5–40 Hz); and (3) baseline correction using a 2-second pre-stimulus interval. For each of the 16 channels (10–20 system), we extracted time-domain statistics and frequency-domain features by estimating PSDs with Welch's method (256-sample windows) and integrating power over Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–40 Hz). Features were harmonized via Box–Cox transformation, ComBat correction (Fortin et al., 2018), Winsorization, and robust Z-scoring.

Common measures (HR and RT)

- **HR:** Recorded in both experiments to provide a shared physiological reference and enable temporal alignment across modalities. For each trial, we computed summary statistics (mean, standard deviation, minimum, maximum) as indicators of overall arousal.
- **Reaction Time (RT):** Recorded in both experiments, defined as the elapsed time between the initial presentation of a meme and the subject's response to proceed to the next stimulus. Stimuli remained on screen until response, followed by a 3-second pause to prevent overlap.

This process yielded a rich, multimodal dataset linking meme content to synchronized physiological responses, forming our new resource.

4. Physiological Correlates of Sexism Perception

Our analysis (**RQ1**) revealed significant physiological markers, showing that the human brain and body appear to respond with specific patterns to sexist content.

4.1. Cognitive Load and Autonomic Arousal

Eye-tracking data established a clear cognitive load gradient. As shown in Table 2, both **reaction time** and **fixation count** increased significantly and progressively ($p < .001$) from non-sexist memes, to direct sexist memes, and finally to judgmental sexist memes, which required the most cognitive effort. Blink duration was also significantly shorter for direct sexist content ($p = .005$), consistent with heightened visual attention, as shorter blink durations have been shown to occur under increased visual workload and sustained attentional engagement (Benedetto et al., 2011).

4.2. Category-Specific and Neural Signatures

A granular analysis revealed unique neurophysiological signatures for specific types of sexism. Most notably, memes featuring **Objectification** triggered a significant **constriction of the right pupil** ($p = .036$), a known physiological marker of processing aversive visual stimuli, as pupillary constriction has been shown to occur in response to unpleasant or emotionally negative content independent of luminance or arousal effects (Ayzenberg et al., 2018; Blini and Zorzi, 2023).

EEG analysis offered insights into brain activity. For binary sexism detection (**Task 1**), we observed a significant **power reduction (desynchronization) in the Alpha, Beta, and Gamma bands** over

Table 2: Key physiological metrics across sexism levels (Mean ± SD). p-values from ANOVA.

Metric	Non-Sexist	Direct Sexist	Judgmental	p-value
Reaction Time (s)	13.68±9.10	15.84±11.40	17.58±12.08	0.000***
Fixation Count	40.31±28.37	44.42±33.67	50.34±37.61	0.000***
Blink Duration (ms)	267.36±57.65	261.13±55.27	263.05±50.58	0.007**

right frontal (anterior) channels when viewing sexist compared to non-sexist memes. This pattern is consistent with prior work on affective picture processing reporting alpha–beta desynchronization and gamma-band modulation in both posterior and anterior (including frontal) regions (Schubring and Schupp, 2019; Strube et al., 2021). The right-frontal localization in our data may reflect higher-order evaluative or conflict-monitoring processes engaged by sexist content. (Figure 1a). When distinguishing *direct* from *judgmental* sexism (**Task 2**), viewing direct memes triggered a **widespread power increase (synchronization) across multiple frequency bands** in the right parietal region, consistent with prior evidence showing enhanced gamma-band synchronization in parietal and visual cortices during the processing of emotionally salient or hostile visual stimuli (Luo et al., 2009; Headley and Paré, 2013), suggesting more intense sensory and affective processing for overtly hostile content (Figure 1b). The analysis of fine-grained categories and emotions revealed further unique signatures. For instance, memes categorized as **Objectification** were associated with significant power increases in Alpha/Theta at central channel C4 and Gamma at frontal Fp2 (Figure 1c). Finally, when the meme's OCR text was automatically classified as inducing **Fear**[1], the EEG registered widespread, bilateral frontal activation across all bands, a pattern consistent with prior findings linking frontal–parietal oscillatory dynamics to threat detection and heightened vigilance during emotionally salient or fear-inducing contexts (Grimshaw et al., 2014; Bodala et al., 2016) (Figure 1d).

5. Multimodal Fusion Model

To leverage these rich signals, we developed a hierarchical attention-based fusion model (Figure 2) designed to integrate enriched content features with sequences of physiological responses.

Enriched Content Representation. Instead of directly fusing raw image features with text, which can be ineffective, we first use a Vision-Language Model (`Qwen2.5-VL-7B`[2]) to generate a detailed textual description and preliminary sexism analysis of each meme. This generated text is concatenated with the meme's original OCR text, separated by a special token.

Physiological Fusion Core. The EEG and ET/HR feature vectors are treated as sequences (where each token represents one subject's reaction). Each physiological sequence is independently passed through a multi-head cross-attention layer. In this crucial step, the physiological sequence acts as the **Query**, while the text token embeddings from the XLM-RoBERTa (XLM-R) encoder serve as the **Key** and **Value**. This allows the model to learn which textual elements are most correlated with the observed neural and attentional reactions. The resulting text-aware physiological representations are aggregated via a learned weighted sum and concatenated with the main text embedding (`[CLS]` token) for final classification through a small Multilayer Perceptron (MLP) head.

Training Strategy. The model is trained in two phases for stability: first, only the fusion head and attention layers are trained for 5 epochs with the XLM-R backbone frozen (LR = 5e-5). Then, the entire model is fine-tuned for 10 epochs using discriminative learning rates (2e-6 for lower layers, 1e-5 for upper layers, 5e-5 for the head) with the AdamW optimizer and a weighted Binary Cross-Entropy (BCE) loss function to handle class imbalance.

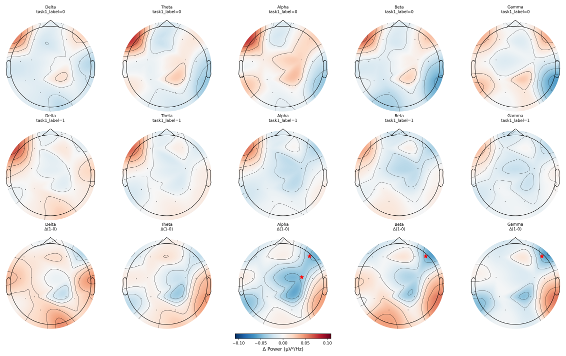
6. Results and Analysis

We evaluated our models using 5-fold cross-validation, reporting macro F1-score and AUC. Our unimodal analysis (**RQ2**) showed that EEG-based classifiers performed surprisingly well, achieving an AUC of 0.717 for Task 1, on par with the text-only XLM-R (AUC 0.704). This confirms that physiological data alone is a powerful predictor of perceived sexism.

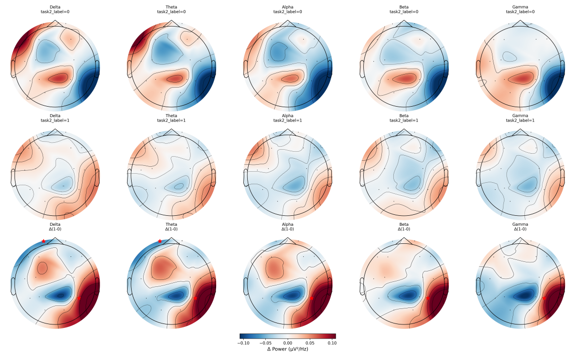
Our multimodal results (**RQ3**), summarized in Table 3, demonstrate the synergistic power of our fusion strategy. Our content-only baseline (`Qwen-VL + XLM-R`) is already very strong, significantly outperforming a simple ViT+XLM-R concatenation.

[1]Emotion labels were derived from the meme's OCR text using language-specific models: `daveni/twitter-xlm-roberta-emotion-es` (Spanish) and `j-hartmann/emotion-english-distilroberta-base` (English).

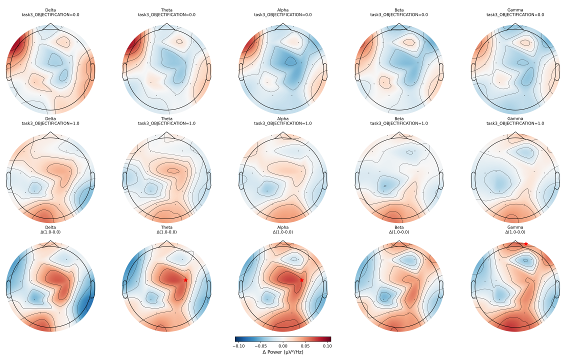
[2]https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct



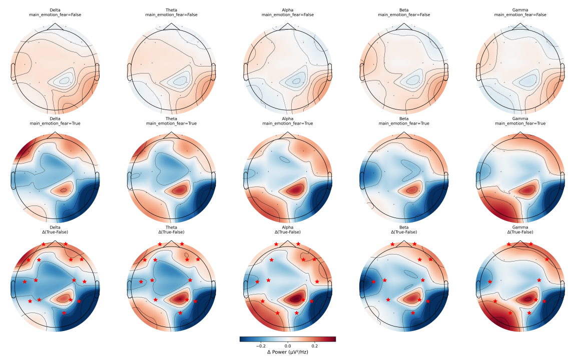
(a) Task 1: Sexist vs. Non-Sexist



(b) Task 2: Direct vs. Judgmental



(c) Task 3: Objectification



(d) Emotion: Fear

Figure 1: EEG topographic maps of band-power differences across key experimental contrasts. For each subfigure, the **top row** shows the mean power for the first condition and the **middle row** for the second; the **bottom row** shows their difference (*Condition 2 – Condition 1*). Specifically: (a) Non-Sexist vs. Sexist; (b) Judgmental vs. Direct Sexism; (c) Non-Objectification vs. Objectification; (d) Neutral vs. Fear (OCR-based emotion). Columns correspond to Delta, Theta, Alpha, Beta, and Gamma bands. Red indicates power increase, blue indicates power decrease; red stars mark channels with statistically significant differences ($p < 0.05$).

Table 3: Multimodal model performance (AUC) across all tasks. Improvements from the baseline are statistically significant ($p < .05$).

Model / Fusion	**T1: Binary**	**T2: Intention**	**T3: Category**
(Simple Fusion) XLM-R + ViT	$0.699 \pm .012$	$0.575 \pm .028$	$0.703 \pm .031$
(Our Models) Qwen-VL + XLM-R	$0.768 \pm .016$	$0.628 \pm .017$	$0.782 \pm .023$
+ EEG	$0.783 \pm .013$	$0.634 \pm .019$	$0.779 \pm .020$
+ EEG + ET/HR	$\mathbf{0.794} \pm \mathbf{.019}$	$\mathbf{0.655} \pm \mathbf{.025}$	$\mathbf{0.784} \pm \mathbf{.015}$

Table 4: F1-score for Task 3 categories. The full model significantly improves classification accuracy on the hardest classes.

Category (Task 3)	**Baseline F1**	**Full Model F1**
Ideological Inequality	$0.566 \pm .020$	$\mathbf{0.585} \pm \mathbf{.007}$
Stereotyping Dominance	$0.479 \pm .051$	$\mathbf{0.498} \pm \mathbf{.020}$
Objectification	$0.541 \pm .037$	$\mathbf{0.561} \pm \mathbf{.026}$
Sexual Violence	$0.396 \pm .045$	$\mathbf{0.428} \pm \mathbf{.041}$
Misogyny NSV	$0.259 \pm .074$	$\mathbf{0.327} \pm \mathbf{.033}$
Macro Average	$0.448 \pm .045$	$\mathbf{0.480} \pm \mathbf{.025}$

However, incrementally adding physiological signals yields statistically significant improvements across all tasks, confirmed by non-overlapping 95% confidence intervals (Figure 3).

For binary sexism detection (**Task 1**), our full model achieves an AUC of **0.794**, a 3.4% relative improvement over the strong baseline. For the highly nuanced intention detection (**Task 2**), where human agreement is low, the physiological signals provide a crucial disambiguating signal, boosting AUC by 4.3% from 0.628 to **0.655**.

The most dramatic impact is in fine-grained categorization (**Task 3**). As shown in Table 4, the F1-score for the most challenging class, *Misogyny & Non-Sexual Violence*, jumps by an unprecedented **26.3%** (from 0.259 to 0.327) with the full model. This shows that when content features are weak, the objective signal of the human response becomes indispensable for more accurate classification.

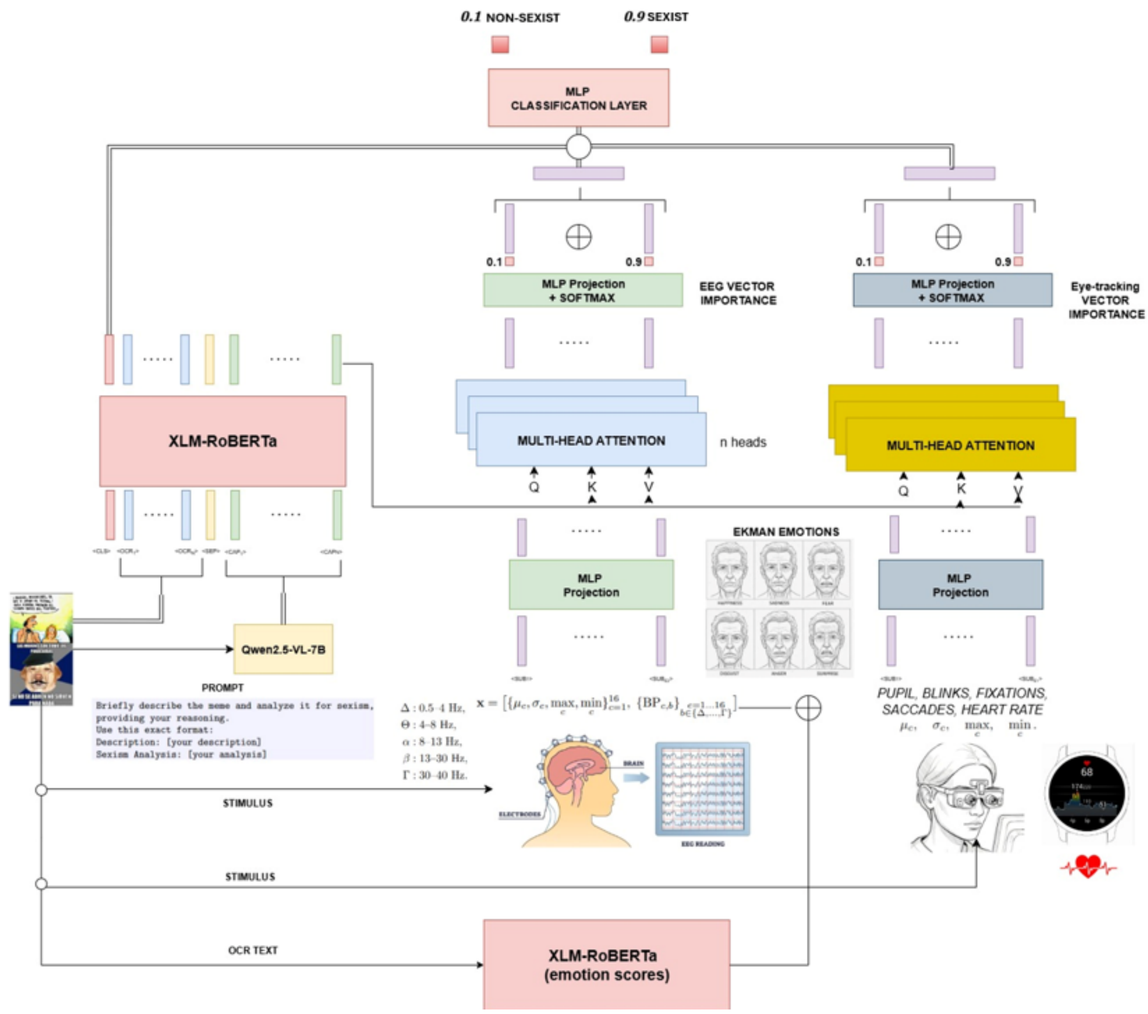


Figure 2: Architecture of the final hierarchical attention-based fusion model. It integrates enriched text (OCR + VLM caption) with sequences of physiological reactions from several subjects (EEG and Eye-Tracking/HR). Cross-attention mechanisms allow the model to learn correlations between specific textual tokens and physiological responses.

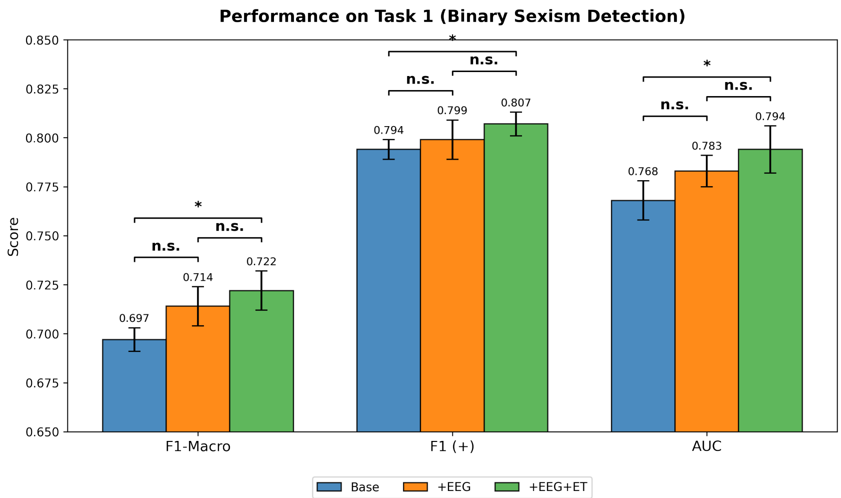


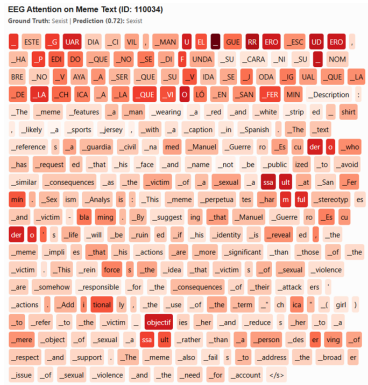
Figure 3: Performance on Task 1 (Binary Sexism Detection) with 95% confidence intervals. Bars represent model performance scores (Macro F1, F1+, and AUC), showing a progressive and statistically significant improvement ($p < 0.05$) as all physiological signals are added.

6.1. Attention-Based Interpretability

To explore how the multimodal fusion model integrates physiological and textual information, we visualized its cross-attention mechanisms. Figure 4 presents a case study of a subtle sexist meme that the content-only baseline misclassified but our full model correctly identified as sexist. The meme (Figure 4a) shows a well-known female political leader with the Spanish caption “¿MUJERES AL PODER? ¿PARA HACER COMO LA TATCHER? FFFFUUUU!!!!” (“WOMEN IN POWER? TO DO LIKE THATCHER? FFFFUUUU!!!!”). The meme uses sarcasm, taking an extreme example of female leadership to ridicule the idea of women in power. It expresses frustration toward women’s political participation, reflecting ideological or stereotypical rather than overt misogynistic sexism.

The attention map (Figure 4b) illustrates how the model integrates multimodal cues. High attention weights align with the meme’s literal text, particularly “¿MUJERES AL PODER?”, and with analytical phrases generated by the VLM, such as “critique of women’s ability to lead.” These associations suggest that the model learns to connect implicit neural patterns with explicit semantic representations of sexism. Note that this visualization does not represent a direct physiological mapping between EEG activity and specific words. It reflects correlations learned by the model, not temporally aligned neural responses. This should be viewed as an AI-level inference rather than a physiological interpretation, which we acknowledge as a limitation that future work should address by integrating EEG with ET for word-level alignment.

(a) Original meme (ID 110051).



(b) Cross-modal attention map between EEG signals and textual tokens.

Figure 4: Attention-based interpretation of a correctly classified sexist meme.

7. Conclusion and Future Work

This work validates a novel, human-centered paradigm for sexism detection. The results seem to indicate that physiological responses to memes provide a rich, objective signal of perception that complements content analysis. We have collected and will release a first-of-its-kind multimodal resource pairing memes with ET, HR, and EEG data, opening new avenues for research on affective computing, focused on systems that sense and interpret human cognitive and emotional reactions. Our multimodal fusion model, which learns to associate implicit physiological patterns with explicit content features, delivers consistent gains over strong content-only baselines, indicating that aligning multimodal systems with human psychological responses is an effective strategy for building more accurate and robust classifiers. Notably, no significant effects were observed for heart rate, likely because static and brief meme stimuli may not elicit sufficient autonomic activation (Brouwer et al., 2013). A further limitation is that memes were not controlled for visual or textual complexity, which could also influence physiological responses independently of sexist content.

Future work should focus on modeling the rich subjectivity in our data, moving beyond majority-vote labels to predict user-specific or demographic-aware perceptions of harm using frameworks like Learning with Disagreements (Uma et al., 2021). Extending this paradigm to dynamic video content (e.g., from TikTok), which introduces temporal complexities, is another critical next step. By continuing to "listen" to the nuanced signals of the human brain and body, we can build more accurate, robust, and ultimately more human-aware systems to foster safer digital environments.

8. Ethical Considerations

The physiological data was gathered from 16 subjects of different nationalities that gave their consent to use the data anonymously for research purposes. The meme stimuli, while sometimes offensive, were part of a pre-existing dataset used for the EXIST shared tasks. For the analysis, the demographic and physiological data were anonymized to protect subject privacy. The goal of this research is to develop human-centered AI models, and we advocate for the use of the dataset and models introduced in this work in systems that prioritize transparency, user agency, and robust appeals processes.

Acknowledgments

This work was done in the framework of the research project *Multimodal and Multisensor data-centric AI models (ANNOTATE-MULTI2)* funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU (Grant PID2024-156022OB-C32). The work of Elena Gomis-Vicent was carried out while she was at the PRHLT Research Center.

9. References

Gavin Abercrombie, Nikolas Vitsakis, Aiqi Jiang, and Ioannis Konstas. 2024. Revisiting annotation of online gender-based violence. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 31–41, Torino, Italia. ELRA and ICCL.

Amnesty International UK. 2025. Toxic tech: New polling exposes widespread online misogyny driving gen z away from social media. Press release, Amnesty International UK.

Iván Arcos and Paolo Rosso. 2024. Sexism identification on TikTok: A multimodal AI approach with text, audio, and video. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 61–73, Cham. Springer Nature Switzerland.

Catalina Argüello-Gutiérrez, Ana Cubero, Fabiola Fumero, Diana Montealegre, Pía Sandoval, and Vanessa Smith-Castro. 2022. I'm just joking! Perceptions of sexist humor and sexist beliefs in a latin american context. *International Journal of Psychology*, Online first.

Vladislav Ayzenberg, Meghan R. Hickey, and Stella F. Lourenco. 2018. Pupillometry reveals the physiological underpinnings of the aversion to holes. *PeerJ*, 6:e4185.

Ali Azarbarzin, Michele Ostrowski, Patrick Hanly, and Magdy Younes. 2014. Relationship between arousal intensity and heart rate response to arousal. *Sleep*, 37:645–653.

Simone Benedetto, Marco Pedrotti, Luca Minin, Thierry Baccino, Alessandra Re, and Roberto Montanari. 2011. Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(3):199–208.

Elvio Blini and Marco Zorzi. 2023. Pupil size as a robust marker of attentional bias toward nicotine-related stimuli in smokers. *Psychonomic Bulletin & Review*, 30(2):596–607.

Indu P. Bodala, Junhua Li, Nitish V. Thakor, and Hasan Al-Nashash. 2016. Eeg and eye tracking demonstrate vigilance enhancement with challenge integration. *Frontiers in Human Neuroscience*, 10:273.

Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

Anne-Marie Brouwer, Nelleke van Wouwe, Christian Mühl, Jan B. F. van Erp, and Alexander Toet. 2013. Perceiving blocks of emotional pictures and sounds: effects on physiological variables. *Frontiers in Human Neuroscience*, 7:295.

Nabamallika Dehingia. 2020. When social media is sexist: A call to action against online gender-based violence. `https://geh.ucsd.edu/when-social-media-is-sexist-a-call-to-action-against-online-gender-based-violence/`.

Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. 2018. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120.

Constantino García Martínez, Sofía Bardají, Pablo Pérez-Tirador, and Abraham Otero. 2024. Rhrveasy: Heart rate variability made easy. *PLOS ONE*, 19.

Francesca Gasparini, Elisabetta Fersini, Zsolt Perjési, Marianna Anzovino, Emanuela Messina, and Manuela Pievani. 2018. Multimodal classification of sexist advertisements. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE)*, volume 1, pages 399–406. SciTePress.

Gina M. Grimshaw, Joshua J. Foster, and Paul M. Corballis. 2014. Frontal and parietal eeg asymmetries interact to predict attentional bias to threat. *Brain and Cognition*, 90:76–86.

Dylan Grosz and Patricia Conde-Cespedes. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2020 Workshops, DSFN, GII,* *BDM, LDRC and LBD, Singapore, May 11–14, 2020, Revised Selected Papers*, page 104–115, Berlin, Heidelberg. Springer-Verlag.

Drew B. Headley and Denis Paré. 2013. In sync: gamma oscillations and emotional memory. *Frontiers in Behavioral Neuroscience*, 7:170.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France. European Language Resources Association.

Aalen Chacko Issac. 2018. Reinforcement of sexism through memes. Report at National Dialogue on Gender-based Cyber Violence. Retrieved from IT for Change website: `https://projects.itforchange.net/e-vaw/wp-content/uploads/2018/01/Aalen_Issac-2.pdf`.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada. Association for Computational Linguistics.

Avishake Kar, Ipsita Pathak, Shibani Mukherjee, and Siddhartha Chatterjee. 2025. Decoding complexity and emotion: A computational linguistic approach to sentence comprehensibility and writer affect. *International Journal of Innovative Science and Research Technology*, pages 423–428.

Sahrish Khan, Arshad Jhumka, and Gabriele Pergola. 2025. Explaining matters: Leveraging definitions and semantic expansion for sexism detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16553–16571, Vienna, Austria. Association for Computational Linguistics.

Varun Khurana, Yaman Kumar, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Trans. Web*, 18(2).

Peiran Lin and Liwen Yang. 2024. Building a virtual psychological counselor by integrating EEG emotion detection with large-scale NLP models. In *Third International Conference on Biological Engineering and Medical Science (ICBioMed2023)*, volume 12924, page 129241X. International Society for Optics and Photonics, SPIE.

Qian Luo, Derek Mitchell, Xi Cheng, Krystal Mondillo, Daniel McCaffrey, Tom Holroyd, Frederick Carver, Richard Coppola, and James Blair. 2009. Visual awareness, emotion, and gamma band synchronization. *Cerebral Cortex*, 19(8):1896–1904.

John T. Nockleby. 2000. *Hate Speech*, 2 edition. Macmillan, New York.

Pulkit Parikh, Harika Abburi, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Trans. Web*, 15(4).

Rogers Prates Pelle and Viviane P. Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Proceedings of the 6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 510–519, São Paulo, Brazil. Sociedade Brasileira de Computação.

Laura Plaza, Jorge Carrillo-de Albornoz, Iván Arcos, Paolo Rosso, Damiano Spina, Enrique Amigó, Julio Gonzalo, and Roser Morante. 2025. EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos. In *Advances in Information Retrieval*, pages 442–449, Cham. Springer Nature Switzerland.

Laura Plaza, Jorge Carrillo-de Albornoz, Víctor Ruiz, Alba Maeso, Berta Chulvi, Paolo Rosso, Enrique Amigó, Julio Gonzalo, Roser Morante, and Damiano Spina. 2024. Overview of exist 2024 — learning with disagreement for sexism identification and characterization in tweets and memes. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 93–117, Cham. Springer Nature Switzerland.

X.-L Quan, Z.-G Zeng, J.-H Jiang, Y.-Q Zhang, B.-L Lv, and D.-R Wu. 2021. Physiological signals based affective computing: A systematic review. *Zidonghua Xuebao/Acta Automatica Sinica*, 47:1769–1784.

Giulia Rizzi, Francesca Gasparini, Aurora Saibene, Paolo Rosso, and Elisabetta Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):103474.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. Call me sexist, but... : Revisiting sexism detection using psychological scales and adversarial samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584.

David Schubring and Harald T. Schupp. 2019. Affective picture processing: Alpha- and lower beta-band desynchronization reflects emotional arousal. *Psychophysiology*, 56(8):e13386.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. Female Astronaut: Because sandwiches won't make themselves up there: Towards multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada. Association for Computational Linguistics.

Andreas Strube, Michael Rose, Sepideh Fazeli, and Christian Büchel. 2021. Alpha-to-beta- and gamma-band activity reflect predictive coding in affective visual processing. *Scientific Reports*, 11:24059.

Teia Swan. 2025. Online misogyny: A digital threat. The KNOW MORE Blog, NO MORE Project. Accessed: 11 Sept. 2025.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

United Nations Entity for Gender Equality and the Empowerment of Women (UN Women). 2025. Repository of UN women's work on technology-facilitated violence against women and girls. Catalogue/directory, Ending Violence Against Women Section, UN Women. Accessed: July 11, 2025.

Francesca Zermiani, Prajit Dhar, Ekta Sood, Fabian Kögel, Andreas Bulling, and Maria Wirzberger. 2024. InteRead: An eye tracking dataset of interrupted reading. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9154–9169, Torino, Italia. ELRA and ICCL.