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

Fusión Multimodal Centrada en el Humano para la Detección de Sexismo en Videos de TikTok con Seguimiento Ocular, Ritmo Cardíaco y EEG

Iván Arcos[1,*], Paolo Rosso[1,2]

[1]PRHLT Research Center, Universitat Politècnica de València (UPV), Valencia, Spain
[2]ValgrAI – Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain

Abstract

Sexism ranges from explicit misogyny to subtle forms of prejudice disguised as humor, irony, or seemingly harmless remarks. This challenge is particularly acute on short-form video platforms such as TikTok, where meaning emerges from the interaction of spoken audio, on-screen text, gestures, editing, and social context. To address this limitation, we propose a human-centered approach that augments content features with physiological data. We rely on a multimodal resource built on 2524 TikTok videos from the EXIST 2026 dataset, evaluated by 10 subjects across 15 sessions while Eye-Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG) signals were recorded. Our analysis reveals significant physiological differences between sexist and non-sexist content, including longer reaction times, increased fixation patterns, and differential EEG signatures. Building on these findings, we propose a hierarchical multimodal fusion architecture that integrates physiological sequences with enriched textual-visual features derived from `Qwen3-VL-4B-Instruct` using four sampled frames per video. The complete model achieves an AUC of 0.790 in binary sexism detection and improves over the text+frames baseline by 3.9% in Task 1 (binary sexism detection), 6.5% in Task 2 (direct vs. judgmental sexism), and 6.8% in Task 3 (multiclass and multilabel sexism categories). These results show that human physiological responses provide a robust complementary signal for improving automated sexism moderation in social media videos.

Keywords

sexism detection, TikTok videos, eye-tracking, heart rate, EEG

1. Introduction

Sexism refers to a multifaceted phenomenon encompassing explicit misogyny as well as subtle forms of prejudice expressed through stereotypes, objectification, irony, humor, or the reinforcement of traditional gender roles. Whether presented as jokes, apparently positive remarks, or overtly hostile comments, sexism shapes domestic expectations, professional opportunities, body image, and broader life prospects. Recognizing its diverse manifestations is therefore crucial to understanding its social impact.

Social media platforms have become conduits for the dissemination of sexist content, perpetuating and normalizing gendered inequalities and biased attitudes. The internet reflects and amplifies existing social asymmetries, while also accelerating the visibility of harmful narratives. This issue is especially pressing for younger audiences, for whom social media constitute a major space of communication, identity construction, and cultural consumption.

Among current platforms, TikTok is particularly relevant. With around 1.9 billion monthly active users worldwide in 2026, it has become one of the largest social platforms globally[1]. Its recommendation dynamics and highly immersive audiovisual format make it especially influential for younger users. Recent public reporting and civil-society analyses have raised concerns that TikTok can expose users to misogynistic and divisive content, and Amnesty UK polling found that 70% of surveyed Gen Z respondents reported encountering misogynistic content on TikTok[2]. These concerns reinforce the need for more robust computational approaches to sexist content moderation.

Previous work has progressively moved from textual sexism detection toward multimodal approaches [1, 2, 3]. However, even strong multimodal systems still struggle when harmfulness depends on implicit interpretation rather than explicit words or visuals. In such cases, sexism is not only a property of the content itself, it is also a phenomenon shaped by how humans perceive and interpret that content. A purely content-driven classifier observes the stimulus but not the cognitive and affective process it triggers in a viewer; yet it is precisely that process—the extra attention, the hesitation, the emotional salience—that often disambiguates whether a borderline clip is read as sexist, ironic, or innocuous.

In this paper, we propose a human-centered approach. Instead of learning only from content and majority labels, we investigate whether physiological reactions can provide a complementary and more objective signal of perception. Neurophysiological and attentional responses such as gaze behavior, heart-rate fluctuations, and EEG activity can reveal cognitive load, emotional salience, and interpretative effort in ways that may help disambiguate subtle sexist content. Unlike self-reported annotations, these signals are recorded continuously and are largely involuntary, so they capture aspects of the perception process that an annotator may not be able, or willing, to verbalize. We do not treat them as a replacement for human judgment but as an additional, content-level source of evidence that is naturally aligned in time with the stimulus.

To this end, we rely on a multimodal resource based on 2524 TikTok videos from EXIST 2026, evaluated by 10 subjects across 15 sessions while recording Eye-Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG). We combine these signals with enriched visual-semantic representations generated from four video frames using

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22–25 September 2026.
*Corresponding author.
iarcgab@etsinf.upv.es (I. Arcos); prosso@dsic.upv.es (P. Rosso)
0009-0005-5290-1753 (I. Arcos); 0000-0002-8922-1242 (P. Rosso)

[1]https://www.charleagency.com/articles/tiktok-statistics/
[2]https://www.amnesty.org.uk/press-releases/toxic-tech-new-polling-exposes-widespread-online-misogyny-driving-gen-z-away-social

`Qwen3-VL-4B-Instruct`[3], and we fuse all modalities through a hierarchical attention mechanism. Although audio is also a meaningful channel in short-form videos, in this work we focus on textual, visual, and physiological signals and leave the explicit integration of the audio modality for future work.

This work addresses the following research questions:

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

RQ2 *Do multimodal classifiers significantly benefit from the integration of visual-semantic and physiological signals beyond textual input alone for sexism detection and categorization in TikTok videos?*

RQ3 *Can the incorporation of physiological data into a multimodal fusion model yield a significant improvement over a strong text+video baseline?*

Contributions. The main contributions of this paper are threefold:

1. **A new multimodal setting.** A human-centered experimental setting for sexism perception in TikTok videos, built on 2524 videos from EXIST 2026 with synchronized physiological responses from 10 subjects across 15 sessions.
2. **Physiological evidence.** An analysis showing statistically significant differences in ocular, autonomic, and neural responses across sexism tasks and categories, highlighting aspects that remain difficult for content-only systems.
3. **A fusion model.** A hierarchical attention architecture that integrates physiological sequences with enriched textual-visual features derived from `Qwen3-VL-4B-Instruct`, yielding consistent gains over strong baselines across multiple sexism-detection tasks.

The remainder of this paper is organized as follows. Section 2 reviews prior work on sexism detection and on the use of physiological signals in language and affective computing. Section 3 describes the multimodal resource, the EXIST 2026 video set, and the data-collection protocol. Section 4 presents the physiological correlates of sexism perception (RQ1). Section 5 introduces the fusion model, and Section 6 reports the experimental results (RQ2–RQ3). Sections 7 and 8 discuss the findings and their limitations, and Section 9 draws some conclusions.

2. Related Work

The automatic detection of harmful content is a well-established research area. It is important to distinguish between **hate speech**, a broad concept [4]; **misogyny**, which involves hatred or dislike of women; and **sexism**, which refers to prejudice, stereotyping, or discrimination based on sex and can manifest subtly. Misogyny lies at the intersection of hate speech and sexism, sharing characteristics of both. These distinctions matter for modeling: hate speech is frequently signaled by explicit lexical markers, whereas sexism, and especially benevolent or ironic sexism, may contain no overtly offensive token at all, which is exactly the regime in which content-only systems degrade.

From textual to multimodal sexism detection. Sexism detection has evolved considerably, moving from textual analysis to multimodal and multilingual approaches, driven by shared tasks such as SemEval [5] and the EXIST lab series [6]. Early approaches relied on classical models such as SVM and Logistic Regression with TF-IDF and n-grams [7, 8]. Later, neural models with pre-trained embeddings improved the capture of semantic context [9, 10]. The emergence of transformer-based pre-trained language models such as BERT and RoBERTa enabled deeper contextual understanding [11], while multimodal systems began to combine textual and visual information more effectively [12, 1]. More recently, large language and vision-language models have been adopted for their reasoning capabilities, although their effectiveness depends strongly on the task and the way they are integrated [13, 14]. A recurring lesson across this literature is that performance is consistently bounded by the cases where harmfulness is implicit: irony, reclaimed language, in-group humor, and context-dependent references all remain difficult regardless of model scale.

Physiological and cognitive signals in NLP. In parallel, a line of research has emerged that integrates physiological data into NLP and multimodal analysis to capture human cognitive and affective responses. **Eye-Tracking** has been used to study cognitive load and attention, showing that gaze patterns shift when people process complex, ambiguous, or emotionally salient material [15, 16]. **Heart Rate** provides a general indicator of physiological arousal and stress and can support temporal alignment across modalities [17]. **Electroencephalography** offers direct, high-temporal-resolution measurements of neural activity, making it possible to observe oscillatory correlates of semantic and emotional processing in real time [18]. A substantial body of psychophysiological work establishes the interpretive vocabulary we rely on later: alpha-band dynamics are linked to attentional gating, theta to cognitive control and memory demands, and beta/gamma to top-down maintenance and feature binding.

Human-centered signals as complementary evidence. Although these sensors have been explored in NLP and affective computing, they have not been systematically integrated into sexism detection in short social media videos. As in previous human-centered multimodal work [19, 20], physiological responses are treated as complementary content-level signals. These responses do not replace human annotations; instead, they provide extra evidence in cases where textual and visual cues are not enough. This is particularly valuable for TikTok, where meaning often emerges from complex audiovisual interplay and where subtle sexism may be difficult even for human annotators to agree upon. Rather than replacing human judgments, physiological signals provide complementary information about how content is perceived, helping the model capture cognitive and attentional patterns that are not directly observable from the video itself.

[3] https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct

Recent trends: fine-grained, temporal, and relational modeling. More recently, research in multimodal sexism and hate detection has shifted towards modeling fine-grained and context-dependent phenomena in short-form videos. New datasets such as FineMuSe [2] extend previous resources by incorporating hierarchical annotations that capture subtle communicative strategies, including irony and humor, which are particularly prevalent in short-form platforms and often hinder content-only approaches. In parallel, recent advances in video-capable vision-language models (VLMs), such as Qwen3-VL [21], have enabled richer multimodal representations by integrating temporal, visual, and textual signals into unified embeddings. Complementary work has also emphasized the importance of temporal localization, proposing models such as MultiHateLoc [22], which move beyond video-level classification to identify when harmful content emerges within a sequence. These approaches highlight a growing trend towards temporally-aware and hierarchically structured multimodal architectures.

At the same time, a novel research direction explores the integration of human-centered signals and relational reasoning mechanisms into multimodal moderation systems. Physiological feedback is increasingly considered as an implicit supervision signal, as shown in recent work on cognitive feedback from EEG [23], which demonstrates that neural signals can encode user preferences and interpretative effort during language processing. In the context of sexism detection in memes, recent multimodal approaches incorporating EEG, eye-tracking, and heart rate signals [24] show that physiological responses improve the identification of sexist content, leading to consistent gains in both binary detection and challenging fine-grained cases. Beyond sensor-based supervision, relational reasoning has also emerged as a key paradigm, with models such as MemeWeaver [3] leveraging inter-instance graph structures to capture dependencies between memes and improve the detection of implicit misogyny. Together, these advances point towards a new generation of multimodal systems that combine temporal reasoning, human-centered signals, and relational modeling to address the complexity of harmful content in modern social media. Our work follows this direction by applying human-centered physiological modeling to short-form TikTok videos, moving beyond static content and explicitly addressing the temporal and audiovisual nature of the platform.

3. A Multimodal Resource for Sexism Detection in Videos

To build and evaluate the proposed human-centered models, a multimodal setting is used in which physiological responses are recorded while subjects watch existing sexist and non-sexist videos annotated as part of the EXIST 2026 shared task[4].

3.1. EXIST 2026 TikTok Videos

We used the official training split of the EXIST 2026 shared task, a publicly available dataset containing 2524 short-form videos (**1524** in Spanish and **1000** in English). In EXIST, sexist videos are categorized by communicative intent: **Direct sexism** refers to content that itself expresses or promotes sexist ideas, whereas **Judgmental sexism** refers to content that portrays or condemns sexist situations or behaviors. This distinction is relevant because judgmental videos often require contextual interpretation of the speaker's stance. Videos are also annotated for fine-grained categories. Table 1 summarizes the dataset statistics.

Table 1
EXIST 2026 Video Dataset Statistics (TRAIN). Percentages for T2 and T3 are calculated over the sexist subset of each language.

Task	Label	Spanish	English
T1: Sexism ID	Sexist	747 (49.0%)	455 (45.5%)
	Non-Sexist	768 (50.4%)	538 (53.8%)
T2: Intention	Direct	495 (66.3%)	241 (53.0%)
	Judgmental	230 (30.8%)	194 (42.6%)
	Ties	22 (2.9%)	20 (4.4%)
T3: Category	Ideolog. Ineq.	141 (18.9%)	53 (11.6%)
	Stereotyping	320 (42.8%)	219 (48.1%)
	Objectification	50 (6.7%)	62 (13.6%)
	Sexual Violence	36 (4.8%)	12 (2.6%)
	Misogyny NSV	24 (3.2%)	7 (1.5%)
	Ties	176 (23.6%)	102 (22.4%)
Total Videos		**1524**	**1000**

The training dataset exhibits a moderate class imbalance, particularly in the fine-grained categorization task. While *Stereotyping and Dominance* is by far the most frequent category (42.8% in Spanish and 48.1% in English), other categories such as *Misogyny and Non-Sexual Violence* (3.2% and 1.5%) and *Sexual Violence* (4.8% and 2.6%) are considerably underrepresented. This long-tail distribution reflects the inherent variability of sexist content, where subtle and context-dependent phenomena occur less frequently but are often more challenging to identify. In addition, the presence of ties—especially in Task 3—further highlights the ambiguity and subjectivity of the annotation process. The bilingual composition is also relevant for our analysis: because the physiological subjects were of different nationalities, language-dependent effects in agreement and performance must be interpreted with care, and we report Spanish and English separately wherever the sample size permits.

To quantify annotation reliability, inter-annotator agreement is computed using Fleiss' κ, allowing a variable number of annotators per instance. Results are shown in Table 2. For Task 1 (binary sexism detection), agreement is substantial ($\kappa \approx 0.69$) and highly consistent across languages, indicating that annotators largely agree on whether content is sexist or not. For Task 2 (intention classification), agreement is even higher ($\kappa = 0.799$ overall), reaching $\kappa = 0.837$ in Spanish and $\kappa = 0.7325$ in English, suggesting that once content is identified as sexist, annotators can more reliably distinguish between direct and judgmental forms.

In contrast, Task 3 (fine-grained categorization) presents significantly lower agreement. Given its multilabel nature, agreement is computed independently for each category and a macro-average κ of 0.430 overall (0.498 for Spanish and 0.305 for English) is reported. As detailed in Table 3, agreement varies notably across categories: *Objectification* and *Stereotyping* achieve the highest agreement levels (above 0.48), while *Misogyny and Non-Sexual Violence* shows sub-

[4] https://nlp.uned.es/exist2026/

stantially lower agreement ($\kappa = 0.249$ overall), reflecting its ambiguity and rarity. These results confirm that fine-grained sexism classification is inherently more subjective and challenging, particularly for less frequent and more implicit categories.

Table 2
Inter-annotator agreement (Fleiss' κ) for Tasks 1–3.

Task	All	ES	EN
T1	0.6898	0.6899	0.6896
T2	0.7990	0.8370	0.7325
T3	0.4297	0.4982	0.3049

Table 3
Inter-annotator agreement for Task 3 (Fleiss' κ, multilabel). Agreement computed independently per category as a binary decision among annotators who marked the item as sexist in Task 1.

Category	All	Spanish	English
Ideological Inequality	0.4336	0.5628	0.2351
Stereotyping Dominance	0.4874	0.5641	0.3184
Objectification	0.5049	0.5355	0.4464
Sexual Violence	0.4736	0.5152	0.3970
Misogyny NSV	0.2487	0.3136	0.1276

Overall, these results confirm that while coarse-grained sexism detection is relatively reliable, fine-grained categorization remains a challenging task characterized by class imbalance, ambiguity, and lower inter-annotator agreement. This motivates the need for additional signals—such as physiological responses—to better capture how users perceive and interpret complex multimodal content. For the experiments in this work, a majority-voting strategy is adopted to derive a single ground-truth label per instance from the annotations. We are aware that majority voting discards potentially informative disagreement; we return to this point in Section 8, where we argue that the physiological signals partially recover the information lost by aggregation.

3.2. Physiological Data Collection

Physiological data were collected from a total of **10 subjects** of different nationalities, with a gender-balanced distribution, across **15 recording sessions**. Subjects were seated comfortably while videos were displayed on a screen. Before the main experiment, a short resting baseline was recorded. During each trial, the subject watched the full video and then responded to proceed to the next stimulus. A 3-second neutral inter-stimulus interval was used to reduce carry-over effects. All data were collected under informed consent, and the resulting records were anonymized for research use.

The fixed presentation order within a session, combined with the resting baseline and the inter-stimulus interval, allows each trial to be segmented and aligned to its stimulus onset, which is essential for the band-power and reaction-time computations described below. Because several subjects could view the same video across sessions, each video can be associated with a *set* of physiological responses rather than a single trace; this many-to-one relationship is exploited directly by the fusion model, which treats the per-subject responses to a given video as a sequence.

The following devices were used:

- **Eye-Tracking:** Pupil Labs Neon glasses recorded binocular gaze data at 200 Hz.
- **Heart Rate:** A Garmin Venu 3 watch continuously recorded heart-rate signals.
- **Electroencephalography:** A 16-channel OpenBCI Cyton headset recorded neural activity at 250 Hz using the 10–20 system.

3.3. Physiological Feature Extraction

From the raw physiological data, a comprehensive feature set was extracted for each video-viewing instance.

Eye-Tracking (ET). For each oculomotor event (fixations, saccades, blinks, pupil dynamics), summary statistics such as mean, standard deviation, minimum, maximum, and count were computed, capturing patterns of visual attention and exploration. These event-level aggregates summarize how intensively and how erratically a subject scanned the stimulus, which prior work links to cognitive load and visual-search difficulty [25, 26].

Electroencephalography (EEG). Signals were preprocessed through: (1) conversion to μV; (2) zero-phase 4th-order Butterworth band-pass filtering (0.5–40 Hz); and (3) baseline correction using a pre-stimulus interval. For each of the 16 channels, time-domain statistics and frequency-domain features were extracted by estimating PSDs with Welch's method and integrating power over the Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–40 Hz) bands. Features were harmonized via Box–Cox transformation, ComBat correction [27], Winsorization, and robust z-scoring. The harmonization step is important in a multi-session setting: it mitigates session- and subject-level shifts in signal scale that would otherwise dominate the genuine, condition-related variance we wish to model.

Common Measures (HR and RT). Heart rate was recorded in all sessions to provide a shared physiological reference and a measure of arousal. For each trial, summary statistics such as mean and standard deviation were computed. **Reaction Time (RT)** was also recorded, defined as the elapsed time between video onset and the subject's response to proceed. Because RT integrates viewing time and decision time, it functions as a coarse but robust proxy for the overall interpretative effort demanded by a clip.

This process yielded a multimodal resource linking video content to synchronized physiological responses. To support reproducibility, the preprocessing code for ET/HR/EEG, the exact prompts used for the vision–language model, and the random seeds used in all experiments are publicly available.[5]

4. Physiological Correlates of Sexism Perception

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

[5] https://github.com/ivanarcos02/exist2026-human-centered-sexism-tiktok

Table 4
Key physiological features across sexism levels (Mean ± SD). p-values from one-way ANOVA.

Feature	Non-Sex.	Direct	Judg.	p
React. Time (s)	23.64±20.30	25.81±16.18	34.19±21.74	0.00000***
Fixation Count	58.13±37.09	62.53±34.10	74.86±49.61	0.00000***
Saccade Count	57.26±37.27	61.34±34.59	73.88±49.81	0.00000***
Blink Count	7.31±7.04	7.79±7.31	9.58±9.56	0.00000***
HR Std (bpm)	1.94±1.36	2.12±1.52	2.09±1.36	0.00011***
HR Mean (bpm)	66.92±10.76	66.36±9.99	67.47±11.37	0.04601*
Blink Dur. (ms)	252.59±34.37	249.94±33.83	250.59±33.14	0.05001

4.1. Cognitive Load and Autonomic Arousal

Eye-tracking and HR data revealed a clear cognitive-load gradient. Both oculomotor activity and cardiovascular measures are widely used as indices of momentary changes in mental effort and workload [28, 29, 25]. As shown in Table 4, **reaction time**, **fixation count**, **saccade count**, and **blink count** all increased significantly from non-sexist videos to direct sexist videos, and further to judgmental sexist videos. Longer reaction times are consistent with increased task demands, as cognitive and visual workload manipulations reliably slow responses in detection-response paradigms [30]. Similarly, increased fixation-related activity has been reported under higher mental workload and task difficulty [25], while higher saccade frequency reflects more demanding visual search processes [26, 31]. Taken together, the increase in reaction time and visual scanning behavior in the judgmental condition indicates the highest level of visual search and interpretative effort [29, 28].

The ordering of the three conditions is itself informative. The monotonic non-sexist → direct → judgmental progression mirrors the intuitive difficulty ordering of the task: judgmental content requires the viewer to recognize that a sexist situation is being depicted *and* to infer the author's critical stance, a second-order inference that direct content does not demand. The physiological gradient therefore offers an external, behavior-level corroboration of why Task 2 is harder than Task 1 for models as well.

HR measures also differed across conditions, suggesting differences in autonomic modulation. Although standard heart rate variability (HRV) metrics based on inter-beat intervals are not computed here, variations in mean heart rate and short-term fluctuations have been shown to reflect changes in arousal, cognitive workload, and stress-related autonomic activity [32, 33, 34]. Finally, blink duration was marginally shorter for sexist stimuli, consistent with heightened visual attention and increased visual workload. Shorter blink durations have been reported as a sensitive indicator of increased visual workload, and workload increases have been associated with reductions in blink duration [35, 29].

4.2. Category-Specific and Neural Signatures

A more detailed analysis revealed distinct EEG signatures across tasks and categories. Only electrodes showing statistically significant differences are interpreted in the difference maps below. Within the ERD/ERS framework, decreases in band power are interpreted as event-related desynchronization rather than as reduced processing [36].

For **Task 1** (*Sexist – Non-Sexist*), electrodes with statistically significant differences in the Delta, Theta, and Alpha bands showed **negative differences**, indicating lower power for sexist videos. These effects were mainly located over central and parietal scalp regions. Under the ERD/ERS framework, this pattern is consistent with increased neural engagement. In particular, alpha-band modulations have been linked to attentional gating, where changes in alpha power reflect the allocation of processing resources and the suppression of irrelevant information [37]. More broadly, the observed pattern suggests modulation of low-frequency rhythms related to cognitive demands (theta) and integrative processes (delta), rather than a global increase in oscillatory activity [38, 39, 36].

For **Task 2** (*Judgmental – Direct*), all electrodes with statistically significant differences were **positive** and located in the **Beta** and **Gamma** bands over frontal and right-lateral regions, indicating higher power for judgmental content. This is consistent with stronger top-down processing and contextual disambiguation demands. Beta activity has been linked to maintenance of the current cognitive state, while gamma-band synchrony has been associated with effective neuronal communication. In addition, feedforward and feedback signaling can rely on different frequency channels involving gamma and alpha/beta rhythms [40, 41, 42, 43].

For **Task 3**, the *Sexual Violence – Rest* contrast showed a consistent pattern of **negative differences** across all electrodes with statistically significant differences. This is consistent with event-related desynchronization and greater engagement under emotionally salient or aversive stimuli [36, 44, 45]. The concentration of significant sites over right-lateral scalp regions is also in line with reported hemispheric asymmetries in affective processing, especially in theta and upper-alpha bands [46].

Finally, in the *Joy – Rest* contrast, all electrodes with statistically significant differences showed **positive differences**, indicating higher power for joy-related stimuli. This suggests that oscillatory activity is modulated by emotional salience more generally, and not only by negative affect [44, 45]. These effects were mainly located over frontal and fronto-lateral scalp regions, consistent with prior EEG evidence linking frontal activity to emotional processing. In particular, previous work has reported differential frontal involvement in positive versus negative emotions [47].

Read together, the four contrasts describe a coherent

Table 5
Agreement between GOLD_EXIST (external annotators) and GOLD_SUBJECT (physiological subjects) for the EXIST 2026 TikTok video training set. Cohen's κ is computed on videos with valid, non-tied labels from both annotation groups.

Task	Lang.	N	Accuracy	κ
T1	ALL	1927	0.815	0.619
	EN	711	0.802	0.543
	ES	1216	0.822	0.644
T2	ALL	988	0.917	0.817
	EN	364	0.852	0.703
	ES	624	0.955	0.886

picture: low-frequency desynchronization tracks the engagement elicited by sexist and aversive material, whereas higher-frequency synchronization tracks the contextual and evaluative processing that distinguishes judgmental from direct content. This complementary behavior across frequency bands is precisely the kind of structured signal that a learned attention mechanism can exploit, and it motivates the band-resolved EEG features used by our model.

The corresponding topoplots are shown in Figure 1. Since these are scalp-level maps, they should be interpreted as distributed electrophysiological patterns rather than as precise source localization results.

4.3. Agreement Between Physiological Subjects and External Annotators

To connect the perception of our physiological subjects with standard human annotation, we measured the agreement between the majority labels derived from the physiological subjects (GOLD_SUBJECT) and the official EXIST 2026 labels provided by the external annotators (GOLD_EXIST) for the TikTok video subset. Table 5 reports Cohen's κ computed on the agreement set, i.e., videos with valid, non-tied labels in both sources. For binary sexism detection (T1), agreement is substantial overall ($\kappa = 0.619$), with higher agreement in Spanish ($\kappa = 0.644$) than in English ($\kappa = 0.543$). For intention classification (T2), agreement is very high overall ($\kappa = 0.817$), especially in Spanish ($\kappa = 0.886$), confirming that the physiological subjects who provided the sensor signals were strongly aligned with the external EXIST annotations on the distinction between direct and judgmental sexism.

5. Multimodal Fusion Model

To leverage these signals, we developed a hierarchical attention-based fusion model (Figure 2) designed to integrate enriched content features with sequences of physiological responses. The design follows two principles. First, content and physiological signals should be combined *asymmetrically*: the text remains the backbone of the representation, and physiology is used to reweight and enrich it rather than to overwrite it. Second, because a single video may be associated with responses from several subjects, the model must accept a variable-length set of physiological observations and remain invariant to their order.

Enriched Content Representation. Instead of directly fusing raw frame features with text, we first sample **four frames per video** and process them jointly with the vision-language model `Qwen3-VL-4B-Instruct`. The model is prompted to produce a structured description of the video and a brief sexism-oriented interpretation using the following instruction:

Prompt used for `Qwen3-VL-4B-Instruct`

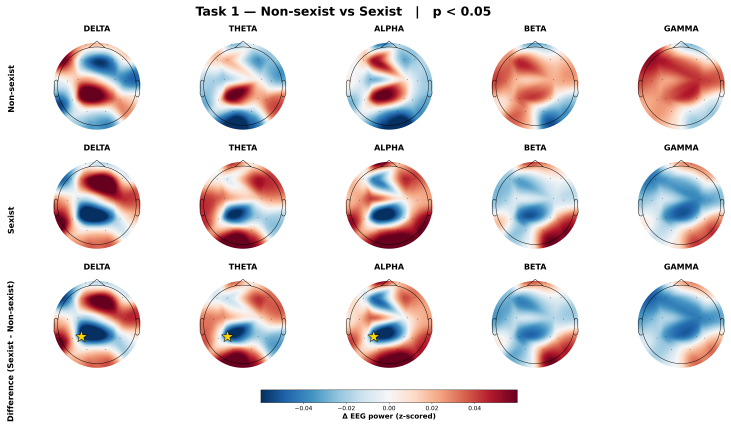
```
You are analyzing MULTIPLE frames extracted from
a short social media video.
Use ONLY what is visible. Do not invent.
Task:
1) Summarize what happens in the video (1-2
sentences) and include on-screen text if present.
2) State whether there are sexist elements
(stereotypes, objectification, slurs, etc.) and
why (very brief).
Follow EXACTLY this format:
Description: <1-2 sentences>
Sexism Analysis: <sexism or not + brief reason>
```

The resulting generated text is concatenated with the original video text (e.g., transcript and OCR) and then encoded. Generating a compact, format-constrained description rather than passing raw frame embeddings has two practical advantages: it grounds the visual evidence in language that the text encoder can directly contextualize, and it keeps the multimodal signal interpretable, since the generated description can be reviewed when analyzing model decisions.

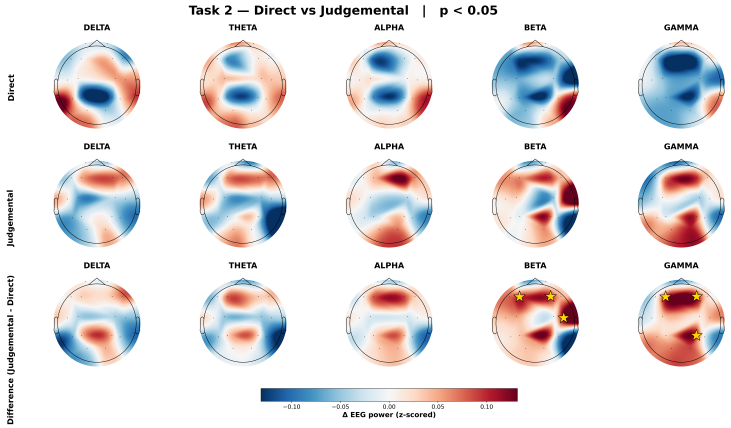
Text Encoder. The concatenated text is processed by XLM-ROBERTa (XLM-R), which provides a multilingual contextual representation of both the original textual signal and the VLM-generated description. A multilingual encoder is essential given the bilingual (Spanish/English) composition of the dataset, as it allows a single model to be trained jointly across both languages.

Physiological Fusion Core. The EEG and ET/HR feature vectors are treated as sequences, where each token corresponds to one subject's physiological reaction to the video. Each physiological sequence is independently passed through a multi-head cross-attention layer. In this step, the physiological sequence acts as the **Query**, while the XLM-R text token embeddings serve as the **Key** and **Value**. This allows the model to learn which textual or visually grounded elements are most correlated with the observed neural and attentional reactions. Casting physiological signals as the query is a deliberate choice: it lets each subject's response "ask" which parts of the content best explain it, so the fusion is driven by the perceptual trace rather than by a fixed, content-only attention map.

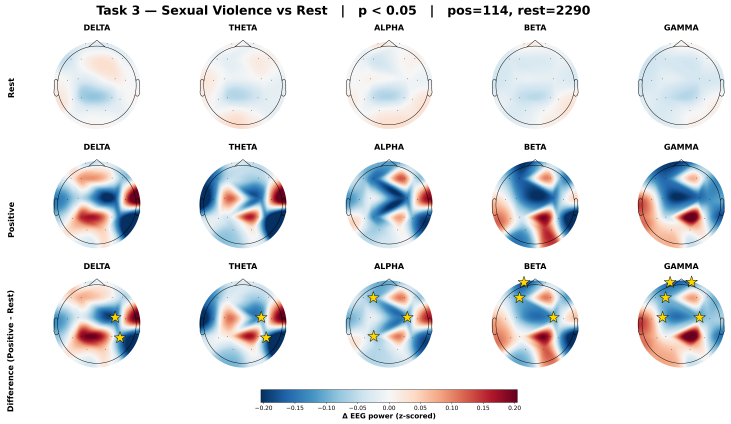
Classification Head and Training Strategy. The resulting text-aware physiological representations are aggregated via a learned weighted sum and concatenated with the main text embedding (`[CLS]` token) for final classification through a small Multilayer Perceptron (MLP) head. The model is trained in two phases for stability: first, only the fusion head and attention layers are trained with the XLM-R backbone frozen; then, the entire model is fine-tuned using discriminative learning rates and a weighted loss to account for class imbalance. The two-phase schedule prevents the



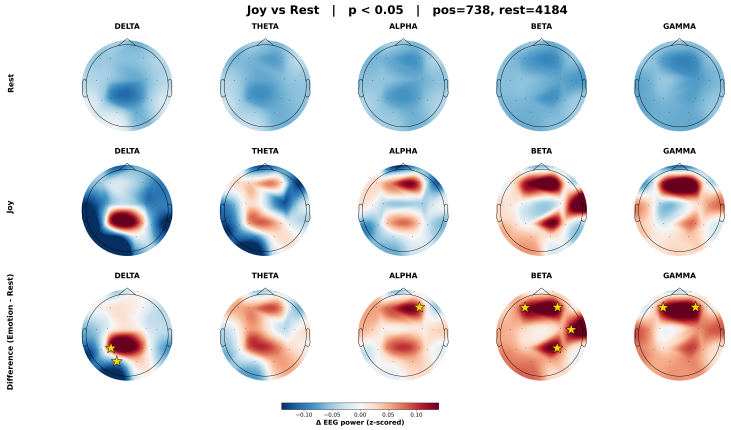
(a) T1: Sexist vs. Non-Sexist.



(b) T2: Direct vs. Judgmental.



(c) T3: Sexual Violence vs. Rest.



(d) Emotion: Joy vs. Rest.

Figure 1: EEG topographic maps of band-power differences across four representative contrasts. In each subfigure, rows correspond to the two conditions and their difference, and columns correspond to Delta, Theta, Alpha, Beta, and Gamma bands. Red indicates positive differences and blue indicates negative differences; yellow stars mark statistically significant electrodes ($p < 0.05$).

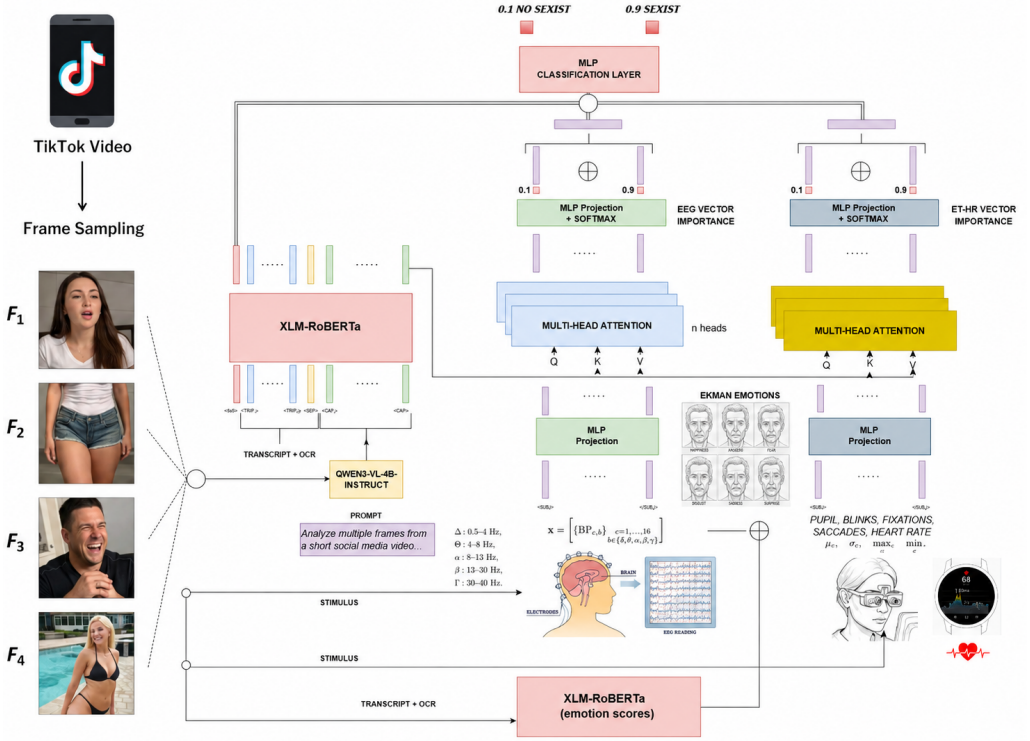


Figure 2: Architecture of the final hierarchical attention-based fusion model. It integrates enriched video text (transcript/OCR + VLM-generated frame description) with sequences of physiological reactions from several subjects (EEG and ET/HR). Cross-attention mechanisms allow the model to learn correlations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

large pre-trained encoder from being destabilized by gradients flowing from the initially untrained fusion components, and the class-weighted objective is particularly important for the long-tailed Task 3 categories identified in Table 1.

6. Results and Analysis

We evaluated our models using **10-fold cross-validation**, reporting AUC and category-level F1. Our multimodal results (**RQ2–RQ3**), summarized in Table 6, demonstrate the synergistic power of the fusion strategy. The content-only baseline is already strong, but progressively adding physiological signals yields consistent improvements across all tasks. Because the four rows of Table 6 differ only in the signals they incorporate, the table doubles as an incremental ablation: each successive row isolates the marginal contribution of frames, then EEG, then the combination of eye-tracking and heart rate.

For binary sexism detection (**Task 1**), the full model achieves an AUC of **0.790**, improving over both the text-only system and the text+frames baseline. Relative to the text+frames model, the addition of sensory variables yields a **3.9%** improvement. For the more nuanced intention detection (**Task 2**), the benefit of physiological signals is even clearer, increasing AUC from 0.740 to **0.788**, i.e., a relative gain of **6.5%**. The gains persist in fine-grained categorization (**Task 3**), where the full model reaches **0.703** AUC, corresponding to a relative improvement of **6.8%** over the text+frames baseline.

A consistent pattern emerges when the gains are read alongside the nature of each task. The smallest relative improvement appears in Task 1, the task with the strongest content-only baseline, where there is comparatively little residual ambiguity for physiological signals to resolve. The larger improvements in Tasks 2 and 3 are explained by different but complementary factors. For Task 3, the fine-grained categorization task, the gain aligns with the lowest inter-annotator agreement ($\kappa = 0.430$), suggesting that physiological signals are especially useful when human labels are themselves uncertain. For Task 2, however, the picture is different: inter-annotator agreement is actually the highest of the three tasks ($\kappa = 0.799$), so the gain cannot be attributed only to label noise. Instead, it likely reflects the intrinsic interpretative difficulty of distinguishing direct from judgmental sexism—a distinction that requires inferring the author's communicative intent rather than detecting surface-level content. The EEG beta/gamma signatures described in Section 4 are consistent with this interpretation, as they point to increased contextual and evaluative processing in judgmental content. In other words, physiological signals help not only where annotators disagree, but also where the content signal alone may be insufficient to recover the intended meaning.

To assess whether these AUC differences are statistically reliable, we conducted Welch's two-sample t-tests between successive model pairs (incremental ablation) and between the full model and the Text+Frames baseline (the comparison cited throughout the paper), using the per-fold mean and standard deviation reported in Table 6 with $n = 10$ folds. Results are shown in Table 7.

All key comparisons are highly significant ($p < 0.001$). The only exception is the Text→Text+Frames step for Task 2 ($p = 0.403$), showing that adding frames alone does not clearly improve intention detection. This makes sense because distinguishing direct from judgmental sexism depends more on interpreting communicative intent than on detect-

Table 6
Multimodal model performance (AUC) across all tasks. Rows form an incremental ablation: each adds one group of signals.

Model	T1	T2	T3
Text	0.705±0.013	0.735±0.012	0.596±0.016
Text + Frames	0.760±0.012	0.740±0.014	0.658±0.015
Text + Frames + EEG	0.776±0.011	0.772±0.011	0.679±0.014
Full (+ ET + HR)	**0.790±0.011**	**0.788±0.011**	**0.703±0.013**

Table 7
Statistical significance of AUC differences between model pairs (Welch's two-sample t-test, $n = 10$ folds). Δ is the difference in mean AUC. $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, n.s. $p\geq 0.05$.

	T1			T2			T3		
Comparison	Δ	t	p	Δ	t	p	Δ	t	p
Text → Text+Frames	+0.055	9.83	$<0.001^{***}$	+0.005	0.86	$0.403^{\text{n.s.}}$	+0.062	8.94	$<0.001^{***}$
Text+Frames → +EEG	+0.016	3.11	0.006^{**}	+0.032	5.68	$<0.001^{***}$	+0.021	3.24	0.005^{**}
+EEG → Full (+ET+HR)	+0.014	2.85	0.011^{*}	+0.016	3.25	0.004^{**}	+0.024	3.97	$<0.001^{***}$
Text+Frames → Full	+0.030	5.83	$<0.001^{***}$	+0.048	8.53	$<0.001^{***}$	+0.045	7.17	$<0.001^{***}$
Text → Full	+0.085	15.78	$<0.001^{***}$	+0.053	10.30	$<0.001^{***}$	+0.107	16.41	$<0.001^{***}$

ing explicit visual cues. The EEG beta/gamma patterns reported in Section 4 are consistent with this interpretation.

The largest category-level improvements are observed in *Objectification*, *Misogyny and Non-Sexual Violence*, and *Ideological Inequality*, i.e., the categories where contextual and perceptual ambiguity is usually highest. These results suggest that when content features are weak or equivocal, physiological signals provide valuable complementary evidence. Table 8 shows this pattern more clearly. The full model improves notably for *Ideological Inequality* and *Objectification*, while the gain is smaller for *Stereotyping and Dominance*, which is the most frequent and easier-to-detect category. For the rarest categories, the results are less stable: *Misogyny and Non-Sexual Violence* benefit from physiological signals, whereas *Sexual Violence* shows more variability, likely because it has fewer examples.

7. Discussion

The three research questions can now be answered jointly. Regarding **RQ1**, the analysis in Section 4 shows that ocular, autonomic, and neural responses differ significantly between sexist and non-sexist videos and, importantly, across types of sexism. The behavioral measures describe a monotonic cognitive-load gradient, and the EEG contrasts reveal a frequency-resolved structure in which low-frequency desynchronization tracks engagement with sexist and aversive content while higher-frequency synchronization tracks the contextual and evaluative processing that distinguishes judgmental from direct sexism. These are not merely incidental correlations: they recover, at the level of the body and brain, the same interpretative difficulty that the task structure imposes on human annotators.

Regarding **RQ2** and **RQ3**, the fusion results show that these signals are not only measurable but also *useful*. Adding physiology on top of an already strong text+frames baseline improves AUC across all three tasks, and the improvements are statistically meaningful relative to the reported cross-validation variance. The pattern of gains is instructive: for Task 3, the lowest-agreement task ($\kappa = 0.430$), physiology helps by providing evidence where human labels are themselves uncertain. For Task 2, despite its high inter-annotator agreement ($\kappa = 0.799$), physiology yields the second-largest gain—not because annotators disagree, but because the task demands intent inference that surface content cannot supply, and which the beta/gamma EEG signatures capture directly. Task 1, with its strong content-only baseline, shows the smallest but still consistent gain. This alignment between *where content is structurally insufficient* and *where physiology helps* is the central empirical message of the paper: physiological responses encode interpretative effort that neither label noise nor content features alone can recover.

The agreement analysis in Table 5 provides an additional layer of validation. The substantial T1 agreement between the physiological subjects and the external EXIST annotators indicates that the people who produced our signals were not idiosyncratic outliers but largely shared the crowd's notion of what counts as sexist. The very high T2 agreement, particularly in Spanish, further confirms that the physiological subjects were strongly aligned with the official annotations when distinguishing direct from judgmental sexism in the video subset.

Taken together, these results support a moderation paradigm in which physiological evidence is used as a complementary, content-aligned signal during model development, rather than as a profiling mechanism applied to end users. The most natural deployment is one in which the model trained with physiological supervision is later used on content alone, having internalized the perceptual cues during training.

8. Limitations

Several limitations should be kept in mind when interpreting these results. First, the physiological data come from

Table 8
F1-score for Task 3 categories (10-fold cross-validation).

Category	Text	+Frames	+EEG	Full
Ideolog. Ineq.	0.603±0.017	0.645±0.017	0.664±0.015	**0.705±0.014**
Stereotyping	0.830±0.012	0.830±0.012	**0.858±0.011**	0.855±0.011
Objectif.	0.412±0.020	0.441±0.019	0.484±0.018	**0.506±0.017**
Sexual Viol.	0.299±0.022	**0.345±0.020**	0.296±0.021	0.321±0.020
Misogyny NSV	0.273±0.021	0.294±0.020	0.342±0.019	**0.349±0.019**

10 subjects across 15 sessions; although this yields many video-viewing instances, the number of subjects remains relatively small, so subject-specific idiosyncrasies cannot be fully ruled out despite the harmonization procedures. Second, our ground-truth labels are derived by **majority voting**, which discards disagreement that may itself be informative; explicitly modeling annotator disagreement is therefore a promising direction for future work. Third, the EEG analysis is conducted at the **scalp level**: the topographic maps describe distributed electrophysiological patterns and must not be interpreted as precise source-localization results. Fourth, although the agreement between physiological subjects and external EXIST annotators is high, the models are trained and evaluated against a **majority-vote gold label** produced by annotators who are different from the subjects participating in the physiological recordings. Thus, physiological signals should not be understood as direct ground truth for each subject's personal judgment, but rather as perceptual evidence that is generalized to an external consensus label. This mismatch between subject-specific perception and dataset-level majority labels remains an important limitation. Fifth, the data were collected in a **controlled laboratory setting**, which differs from naturalistic, in-the-wild consumption on mobile devices and may limit the external validity of the cognitive-load estimates. Finally, the present model does not yet incorporate the **audio** channel, which is a meaningful source of information in short-form videos.

9. Conclusion and Future Work

This work validates a human-centered paradigm for sexism detection in short social media videos. The results indicate that physiological responses to TikTok videos provide a rich and objective signal of perception that complements content analysis. By integrating EEG, ET, and HR with enriched textual-visual representations generated from video frames, our multimodal fusion model achieves consistent gains over strong baselines across binary detection, intention classification, and fine-grained categorization.

Crucially, the inclusion of sensory variables improves over the text+frames baseline by **3.9%** in binary sexism detection, **6.5%** in intention detection, and **6.8%** in fine-grained categorization. These gains show that physiological signals are not merely auxiliary information, but a meaningful source of complementary evidence in cases where content alone remains ambiguous.

The physiological analysis also offers a more interpretable perspective on why the task is hard. Judgmental sexism requires more contextual disambiguation than direct sexism; sexual violence triggers strong right-lateralized desynchronization patterns; and emotionally salient content such as joy yields robust positive oscillatory differences. These findings reinforce the value of incorporating human responses into computational models of harmful content.

Future work should move beyond majority-vote labels to model subjective disagreement more explicitly, since disagreement is especially informative in subtle and context-dependent cases of sexism. Another important direction is to explore finer temporal alignment between specific video segments and physiological reactions, so that the model can identify not only whether a video is difficult to interpret, but also which moments trigger greater cognitive or emotional response. In addition, future models should incorporate the audio modality, which is central in TikTok videos and may provide cues that are not captured by text or frames alone. Finally, this human-centered paradigm should be evaluated on larger multilingual video collections and in more realistic moderation scenarios.

10. Ethical Considerations

The physiological data used in this study were collected from 10 subjects of different nationalities who provided informed consent for anonymous research use. The study, including the collection of physiological responses to the video stimuli, was approved by the Research Ethics Committee of the Universitat Politècnica de València (reference P13_26-12-2025). The video stimuli came from a pre-existing research dataset used in the EXIST shared tasks. Demographic and physiological data were anonymized to protect subject privacy. The resulting models and resources are intended to be used only in systems that prioritize transparency, user agency, privacy protection, and robust appeal processes. Physiological signals should be understood as a research and model-development resource, not as a mechanism for profiling end users in deployment settings. We further note that physiological measurements are sensitive personal data; their collection, storage, and reuse should remain governed by explicit consent and strict access control, and any operational moderation system derived from this work should rely on content-level models rather than on the continuous monitoring of individuals.

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).

References

[1] I. Arcos, P. Rosso, Sexism identification on TikTok: A multimodal AI approach with text, audio, and video, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024, pp. 61–73. doi:10.1007/978-3-031-71736-9_2.

[2] L. De Grazia, D. Sánchez Villegas, D. Elliott, M. Farrús, M. Taulé, Beyond binary classification: Detecting fine-grained sexism in social media videos, 2026. URL: https://arxiv.org/abs/2602.15757. arXiv:2602.15757.

[3] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, Memeweaver: Inter-meme graph reasoning for sexism and misogyny detection, in: Findings of the Association for Computational Linguistics: EACL, 2026. Also available as arXiv:2601.08684.

[4] J. T. Nockleby, Hate Speech, 2 ed., Macmillan, New York, 2000.

[5] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, E. Sartori (Eds.), Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2193–2210. URL: https://aclanthology.org/2023.semeval-1.305. doi:10.18653/v1/2023.semeval-1.305.

[6] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 442–449. doi:10.1007/978-3-031-88720-8_65.

[7] A. Jha, R. Mamidi, When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O'Connor, O. Tsur, A. S. Doğruöz (Eds.), Proceedings of the Second Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 7–16. URL: https://aclanthology.org/W17-2902/. doi:10.18653/v1/W17-2902.

[8] R. P. Pelle, V. P. Moreira, 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), Sociedade Brasileira de Computação, São Paulo, Brazil, 2017, pp. 510–519. doi:10.5753/brasnam.2017.3260.

[9] F. Gasparini, E. Fersini, Z. Perjési, M. Anzovino, E. Messina, M. Pievani, Multimodal classification of sexist advertisements, in: Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE), volume 1, SciTePress, 2018, pp. 399–406. URL: https://www.scitepress.org/papers/2018/68594/68594.pdf. doi:10.5220/0006859403990406.

[10] D. Grosz, P. Conde-Cespedes, 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, Springer-Verlag, Berlin, Heidelberg, 2020, p. 104–115. URL: https://doi.org/10.1007/978-3-030-60470-7_11. doi:10.1007/978-3-030-60470-7_11.

[11] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Trans. Web 15 (2021). URL: https://doi.org/10.1145/3457189. doi:10.1145/3457189.

[12] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Information Processing & Management 60 (2023) 103474. URL: https://www.sciencedirect.com/science/article/pii/S030645732300211X. doi:https://doi.org/10.1016/j.ipm.2023.103474.

[13] L. Li, L. Fan, S. Atreja, L. Hemphill, "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media, ACM Trans. Web 18 (2024). URL: https://doi.org/10.1145/3643829. doi:10.1145/3643829.

[14] G. Abercrombie, N. Vitsakis, A. Jiang, I. Konstas, Revisiting annotation of online gender-based violence, in: G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, S. Tonelli (Eds.), Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 31–41. URL: https://aclanthology.org/2024.nlperspectives-1.3/.

[15] V. Khurana, Y. Kumar, N. Hollenstein, R. Kumar, B. Krishnamurthy, Synthesizing human gaze feedback for improved NLP performance, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1895–1908. URL: https://aclanthology.org/2023.eacl-main.139/. doi:10.18653/v1/2023.eacl-main.139.

[16] A. Kar, I. Pathak, S. Mukherjee, S. Chatterjee, Decoding complexity and emotion: A computational linguistic approach to sentence comprehensibility and writer affect, International Journal of Innovative Science and Research Technology (2025) 423–428. doi:10.38124/ijisrt/25jul551.

[17] A. Azarbarzin, M. Ostrowski, P. Hanly, M. Younes, Relationship between arousal intensity and heart rate response to arousal, Sleep 37 (2014) 645–653. doi:10.5665/sleep.3560.

[18] P. Lin, L. Yang, Building a virtual psychological counselor by integrating EEG emotion detection with large-scale NLP models, in: A. Wang (Ed.), Third International Conference on Biological Engineering and Medical Science (ICBioMed2023), volume 12924, International Society for Optics and Photonics, SPIE, 2024, p. 129241X. URL: https://doi.org/10.1117/12.3013169. doi:10.1117/12.3013169.

[19] F. Zermiani, P. Dhar, E. Sood, F. Kögel, A. Bulling, M. Wirzberger, InteRead: An eye tracking dataset of interrupted reading, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9154–9169. URL: https://aclanthology.

org/2024.lrec-main.802/.

[20] 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, European Language Resources Association, Marseille, France, 2020, pp. 138–146. URL: https://aclanthology.org/2020.lrec-1.18/.

[21] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, K. Zhu, Qwen3-vl technical report, 2025. URL: https://arxiv.org/abs/2511.21631. arXiv:2511.21631.

[22] Q. Sun, T. Chen, Y. Zhang, Y. Zhang, J. Yue, J. Jiao, Z. Fu, Multihateloc: Towards temporal localisation of multimodal hate content in online videos, in: Proceedings of the ACM Web Conference (WWW), 2026. Also available as arXiv:2512.10408.

[23] Y. Harada, Y. Oseki, Cognitive feedback: Decoding human feedback from cognitive signals, in: Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), Association for Computational Linguistics, Suzhou, China, 2025, pp. 209–219. doi:10.18653/v1/2025.hcinlp-1.17.

[24] I. Arcos Gabaldón, P. Rosso, E. Gomis Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals, in: Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), 2026, pp. 9830–9840. URL: https://doi.org/10.63317/2w2vaanu3pe7. doi:10.63317/2w2vaanu3pe7.

[25] G. G. M. Dalveren, N. E. Cagiltay, Using eye-movement events to determine the mental workload of surgical residents, Journal of Eye Movement Research 11 (2018). doi:10.16910/jemr.11.4.3.

[26] A. Takemoto, A. Nakazawa, T. Kumada, Non-goal-driven eye movement after visual search task, Journal of Eye Movement Research 15 (2022). doi:10.16910/jemr.15.2.2.

[27] J.-P. Fortin, N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman, R. T. Shinohara, Harmonization of cortical thickness measurements across scanners and sites, NeuroImage 167 (2018) 104–120. URL: https://www.sciencedirect.com/science/article/pii/S105381191730931X. doi:https://doi.org/10.1016/j.neuroimage.2017.11.024.

[28] M. De Rivecourt, M. N. Kuperus, W. J. Post, L. J. M. Mulder, Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight, Ergonomics 51 (2008) 1295–1319. doi:10.1080/00140130802120267.

[29] G. Marquart, C. Cabrall, J. de Winter, Review of eye-related measures of drivers' mental workload, Procedia Manufacturing 3 (2015) 2854–2861. doi:10.1016/j.promfg.2015.07.783.

[30] W. van Winsum, The effects of cognitive and visual workload on peripheral detection in the detection response task, Human Factors 60 (2018) 855–869. doi:10.1177/0018720818776880.

[31] A. H. Young, J. Hulleman, Eye movements reveal how task difficulty moulds visual search, Journal of Experimental Psychology: Human Perception and Performance 39 (2013) 168–190. doi:10.1037/a0028679.

[32] G. G. Berntson, J. T. Bigger, D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, M. W. van der Molen, Heart rate variability: Origins, methods, and interpretive caveats, Psychophysiology 34 (1997) 623–648. doi:10.1111/j.1469-8986.1997.tb02140.x.

[33] H. W. Kim, E. J. Cheon, D. S. Bai, Y. H. Lee, B. M. Koo, Stress and heart rate variability: A meta-analysis and review of the literature, Psychiatry Investigation 15 (2018) 235–245. doi:10.30773/pi.2017.08.17.

[34] S. Delliaux, A. Delaforge, J.-C. Deharo, G. Chaumet, Mental workload alters heart rate variability, lowering non-linear dynamics, Frontiers in Physiology 10 (2019) 565. doi:10.3389/fphys.2019.00565.

[35] S. Benedetto, M. Pedrotti, L. Minin, T. Baccino, A. Re, R. Montanari, Driver workload and eye blink duration, Transportation Research Part F: Traffic Psychology and Behaviour 14 (2011) 199–208. URL: https://www.sciencedirect.com/science/article/pii/S136984781000094X. doi:https://doi.org/10.1016/j.trf.2010.12.001.

[36] G. Pfurtscheller, F. H. Lopes da Silva, Event-related eeg/meg synchronization and desynchronization: basic principles, Clinical Neurophysiology 110 (1999) 1842–1857. doi:10.1016/S1388-2457(99)00141-8.

[37] J. J. Foxe, A. C. Snyder, The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention, Frontiers in Psychology 2 (2011) 154. doi:10.3389/fpsyg.2011.00154.

[38] J. F. Cavanagh, M. J. Frank, Frontal theta as a mechanism for cognitive control, Trends in Cognitive Sciences 18 (2014) 414–421. doi:10.1016/j.tics.2014.04.012.

[39] T. Harmony, The functional significance of delta oscillations in cognitive processing, Frontiers in Integrative Neuroscience 7 (2013) 83. doi:10.3389/fnint.2013.00083.

[40] A. K. Engel, P. Fries, Beta-band oscillations–signalling the status quo?, Current Opinion in Neurobiology 20 (2010) 156–165. doi:10.1016/j.conb.2010.02.015.

[41] B. Spitzer, S. Haegens, Beyond the status quo: A role for beta oscillations in endogenous content (re)activation, eNeuro 4 (2017) ENEURO.0170–17.2017. doi:10.1523/ENEURO.0170-17.2017.

[42] P. Fries, Rhythms for cognition: Communication through coherence, Neuron 88 (2015) 220–235. doi:10.1016/j.neuron.2015.09.034.

[43] A. M. Bastos, J. Vezoli, C. A. Bosman, J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P. De Weerd, H. Kennedy, P. Fries, Visual areas exert feedforward and feedback influences through distinct frequency channels,

Neuron 85 (2015) 390–401. doi:10.1016/j.neuron.2014.12.018.

[44] M. Codispoti, A. De Cesarei, V. Ferrari, Alpha-band oscillations and emotion: A review of studies on picture perception, Psychophysiology 60 (2023) e14438. doi:10.1111/psyp.14438.

[45] L. Luther, J. M. Horschig, J. M. van Peer, K. Roelofs, O. Jensen, M. A. Hagenaars, Oscillatory brain responses to emotional stimuli are effects related to events rather than states, Frontiers in Human Neuroscience 16 (2023) 868549. doi:10.3389/fnhum.2022.868549.

[46] L. I. Aftanas, A. A. Varlamov, S. V. Pavlov, V. P. Makhnev, N. V. Reva, Event-related synchronization and desynchronization during affective processing: emergence of valence-related time-dependent hemispheric asymmetries in theta and upper alpha band, International Journal of Neuroscience 110 (2001) 197–219. doi:10.3109/00207450108986547.

[47] G. L. Ahern, G. E. Schwartz, Differential lateralization for positive and negative emotion in the human brain: Eeg spectral analysis, Neuropsychologia 23 (1985) 745–755. doi:10.1016/0028-3932(85)90081-8.