IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 February 2026, accepted 24 February 2026, date of publication 27 February 2026, date of current version 12 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3668915

RESEARCH ARTICLE

Profile-Level Positivity Assessment Using Multimodal Feature Aggregation From Social Media Content

MARTA GORRAIZ-BENGOECHEA[1], ASIER GONZALEZ-SANTOCILDES[1], IKER PASTOR[1], ANA ESTÉVEZ[2], AND GEMA AONSO-DIEGO[2]

[1]Faculty of Engineering, University of Deusto, 48007 Bilbao, Spain
[2]Faculty of Psychology, University of Deusto, 48007 Bilbao, Spain

Corresponding author: Asier Gonzalez-Santocildes (gonzalez.asier@deusto.es)

ABSTRACT Accurately evaluating positivity in social media profiles has important implications for advertising, mental health monitoring, and responsible AI applications. However, the multimodal nature of Instagram, where images, captions, hashtags, and emojis co-occur, challenges conventional unimodal sentiment analysis tools. We present a multimodal system that integrates computer vision and natural language processing to assess Instagram profiles. The architecture comprises two blocks: (i) feature extraction via facial-emotion analysis, object detection, chromatic cues, and linguistic processing; and (ii) profile-level inference, which fuses aggregated features to produce an overall positivity score together with semantically meaningful profile tags. Experiments on public Instagram profiles (images, captions, and comments), including an expert-annotated subset of comments, compare unimodal baselines with transformer-based embeddings and OpenAI large language models. The multimodal pipeline improves over unimodal variants, and the OpenAI approach achieves higher concordance with expert judgements than RoBERTa/Bi-LSTM baselines. Beyond technical contributions, we discuss privacy, transparency, and misuse risks associated with affective computing on social media. The proposed framework contributes to the advancement of multimodal sentiment analysis in social media and highlights its applicability to areas such as marketing analytics, well-being monitoring, and ethically aligned AI systems.

INDEX TERMS Multimodal sentiment analysis, deep learning, computer vision, natural language processing, topic modeling, Instagram, social media analytics, ethical AI.

I. INTRODUCTION

Instagram has become a dominant, intrinsically multimodal platform where images, captions, hashtags, emojis, and comments co-occur to construct online identities and narratives. This hybrid signal space, vision, language, and metadata; makes Instagram a compelling testbed for sentiment analysis. At the same time, the platform is characterized by a positive bias, where users tend to present optimistic or idealized versions of themselves, shifting the empirical distribution of expressed affect towards neutral or positive tones. These traits heighten both the opportunities and the methodological challenges of reliable sentiment measurement at profile level.

Understanding positivity in profiles is valuable across domains: marketeers seek audience aligned creative strategies; well-being monitoring can flag changes in affective tone; and responsible AI use cases require transparent, robust analysis of user-generated content. However, conventional multimodal tools underperform on Instagram because affect is often determined by cross-modal interactions, images whose polarity is altered by captions, hashtags, or emojis. Recent progress in multimodal AI underscores the need to jointly model these signals rather than treat them in isolation.

Despite progress in sentiment analysis, reliable profile-level positivity estimation remains challenging because

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

meaning emerges from cross-modal interactions rather than any single channel. Visual content and captions can be affectively incongruent; emojis act as strong carriers of polarity; and hashtags both encode topics and shift sentiment, often used ironically. Detecting sarcasm frequently needs multimodal or boarder contextual cues beyond literal text. In addition, Instagram's multilingual and highly informal register (abbreviations, code-switching) undermines models tuned to standard written language. Finally, profile-level assessment requires aggregation across post over time, where unimodal, post-only tools tend to missread or dilute the underlying affective signal. These factors explain why text-only or image-only baselines underperform and motivate a multimodal, context-aware, profile-level approach.

Prior work on social-media sentiment analysis tends to optimize post or comment level classifiers and seldom addresses profile-level estimation on Instagram. Typical systems remain unimodal (text-only or image-only) or rely on shallow fusion, which limits robustness to Instagram-specific phenomena such as emojis (strong polarity carriers), hashtag segmentation, irony/sarcasm, and multilingual, informal registers. Moreover, few approaches expose interpretable, profile-level outputs beyond a single polarity score. In contrast, our setting targets profile-level aggregation across posts and returns both a global positivity index and semantic tags that summarize the profile's themes.

We propose a two-block multimodal pipeline for profile-level positivity assessment on Instagram. Block 1 performs the end-to-end feature extraction over images and text; detecting faces and emotions, objects, and chromatic cues on the vision side, and processing captions and comments for sentiment, emojis, hashtags and embeddings on the language side, to generate a global positivity score. The second block then processes those aggregated outputs without re-running vision models, loading the embeddings to compute semantic profile tags and to update the consolidated output. This design aims to concentrate heavy GPU computing in Block 1 and keep Block 2 lightweight and reproducible.

Section II reviews related work on multimodal sentiment analysis and Instagram-specific phenomena (emojis, hashtags, irony, multilinguality). Section III presents the methodology and data, including preprocessing and ethical considerations. Section IV details the system: Block 1 (per-post processing and profile-level aggregation) and Block 2 (tagging over aggregated artifacts), with implementation notes. Section V describes the experimental setup, datasets, baselines, and evaluation metrics. Section VI reports quantitative and qualitative results, ablations, and error analyses. Section VII discusses implications, limitations, and ethics, and concludes with future work.

II. STATE OF THE ART

Multimodal sentiment analysis combines complementary cues from vision and language to overcome the limits of unimodal pipelines. Early approaches treated images and text independently, whereas more recent work emphasizes joint representations and fusion to capture cross-modal dependencies, take cases where a caption reverses the affect implied by an image. Canonical designs include early (feature-level), late (decision-level), and hybrid fusion [1], [2], [3], [4], [5]. These architectural choices interact with representation (faces, objects, chromatic cues; tokens, emojis, hashtags, topics) and evaluation against text-only or image-only baselines, motivating profile-level methodologies that preserve cross-modal interactions rather than analyzing posts in isolation.

Instagram pairs highly visual content with captions, hashtags, emojis, and comments; hence the meaning emerges from the interaction of multiple signals rather than any single channel. Previous analyses also report a tendency toward neutral or positive affect distributions (positivity bias), which both creates an opportunity and raises challenges for sentiment modeling at the account level [6], [7], [8]. Accurate modeling on Instagram text further requires handling platform-specific phenomena: emojis are strong polarity carriers [9]; hashtags encode topics and sometimes sentiment (segmentation helps interpret concatenations, while ironic uses can invert polarity) [10], [11]; sarcasm often demands multimodal or broader context (e.g., image–text incongruence) [12], [13]; and multilingual, informal registers (slang, abbreviations, code-switching) challenge generic language models unless preprocessing and model choices are adapted to social media [14], [15]. These observations justify tailored preprocessing and feature design for Instagram and support a shift from post-level to profile-level analysis.

Fusion strategies are commonly organized as early, late, or hybrid. Early fusion can exploit rich interactions but increases dimensionality and overfitting risk; late fusion is robust but may miss fine-grained dependencies; hybrid schemes aim to balance both [2], [3], [4], [5]. For profile-level analysis, an additional design decision is *when* to aggregate (per post, per temporal window, or per account) so that cross-modal signals are preserved without inflating computational cost. In practice, vision components often include face detection and facial-affect analysis to estimate emotion distributions, object detection to capture scene context, and optional chromatic cues to summarize color tone; language components typically include tokenization/normalization, emoji handling, hashtag segmentation, multilingual support or translation, sentiment scoring, embeddings, and topic/label induction for interpretability.

Beyond model design, *data construction* for Instagram raises additional considerations. Public profiles exhibit heterogeneous posting frequencies, highly variable image/text quality, and topical drift over time; labels derived from comments or crowd work can be noisy, imbalanced, or domain-mismatched relative to profile-level targets. These issues affect both training and evaluation: post-level gold labels do not trivially translate into account-level ground truth, and annotation guidelines must account for emojis, hashtag segmentation, and sarcasm to reduce disagreement.

Prior multimodal studies therefore emphasize careful sampling, stratified evaluation (including emoji-/slang-heavy subsets), and transparent reporting of agreement with expert annotators to avoid overstating gains from fusion or pretraining [9], [12], [14].

Interpretability and transparency are also recurring themes in the literature. While multimodal fusion can improve accuracy, opaque feature combinations hinder practical adoption when stakeholders require explanations rather than only a single polarity score. Consequently, recent pipelines complement scalar outputs with human-readable descriptors (e.g., topics/labels) derived from embeddings or topic models to summarize recurring themes, enabling qualitative audits and error analysis. In the Instagram setting, such descriptors help diagnose cross-modal contradictions (positive imagery with negative captions), isolate sarcasm- or emoji-driven flips, and reveal language-specific failure cases—all of which are difficult to inspect from raw logits alone [4], [5]. This emphasis on interpretability complements standard quantitative metrics and aligns with responsible-usage guidelines for social-media analytics.

Most prior work focuses on post- or comment-level classification, underutilizing the longitudinal and multimodal nature of an account. A profile-level perspective aggregates signals across posts and time, providing a more stable view of affect and enabling downstream applications that require interpretable outputs. Two targets are particularly valuable at this level: (1) a global positivity index summarizing an account's overall tone; and (2) a concise set of semantic tags describing recurring themes. Robust evaluations compare multimodal pipelines against unimodal baselines (text-only and image-only), leverage expert-annotated subsets of comments to gauge agreement, and report both quantitative metrics and qualitative/error analyses—especially for emoji-heavy, slang-heavy, or sarcastic content [9], [12], [14]. This framing guides our methodology and evaluation choices in the remainder of the paper.

A. RECENT MULTIMODAL LLMs AND MULTIMODAL SENTIMENT ANALYSIS: A COMPARATIVE PERSPECTIVE

Recent advances in multimodal large language models (MLLMs) have enabled unified processing of visual and textual inputs through large-scale pretraining. Models such as GPT-4V/ GPT-4o [16], Gemini 1.5 [17], and LLaVA-Next [18] demonstrate strong zero-shot and few-shot capabilities across a wide range of vision language tasks, including image understanding, caption interpretation, and high-level semantic reasoning. These foundation models represent the current state of the art in general-purpose multimodal intelligence and have motivated renewed interest in vision–language integration for downstream applications. Recent surveys further document the rapid evolution of multimodal LLM architectures and training paradigms, positioning them primarily as general semantic backbones rather than task-specific sentiment analysis systems [19].

Despite their expressive power, the objectives and operating assumptions of MLLMs differ substantially from those required for sentiment analysis at profile level in social media. MLLMs typically operate on individual image–text pairs, produce prompt-dependent outputs, and encapsulate affective reasoning implicitly within opaque internal representations. As a consequence, they offer limited support for explicit affect quantification, aggregation across large collections of posts, or reproducible longitudinal analysis—capabilities that are central to stable profile-level sentiment assessment on platforms such as Instagram. Moreover, the lack of persistent intermediate representations complicates auditability and interpretability, which are critical considerations in ethically sensitive affective computing applications.

In parallel, recent research on multimodal sentiment analysis has explored the integration of visual and textual modalities for understanding affective content on social media. Comprehensive reviews highlight contemporary approaches to multimodal sentiment modeling using image enhancement, semantic alignment, and large-model integration, pointing to ongoing advancements in social media affective analysis beyond unimodal baselines [20]. Task-oriented systems have been developed for sentiment classification on social platforms: for example, frameworks addressing the modality gap in fashion-related posts demonstrate effective fusion strategies that outperform unimodal baselines in sentiment classification tasks [21]. Similarly, deep learning-based sentiment analysis methods specifically applied to Instagram content combine contextual embeddings with visual features to model consumer behavior and sentiment polarity [22]. Earlier multimodal fusion studies for Instagram data have also shown that jointly modeling text and image improves polarity detection compared to separate modalities [23].

Complementary to these developments, recent work on visual emotion recognition has demonstrated that deep learning architectures can reliably estimate affect from facial expressions in unconstrained settings through the comparison of multiple convolutional and deep learning models [24]. While such approaches provide strong visual affect signals, they remain unimodal and do not capture the cross-modal interactions between imagery, captions, emojis, and comments that characterize Instagram communication.

These recent advances reinforce the need for systems that bridge general-purpose multimodal representations and task-specific affect modeling. The framework proposed in this work occupies an intermediate position between end-to-end multimodal LLMs and post-level sentiment classifiers. It leverages state-of-the-art language embeddings derived from foundation models to ensure semantic robustness, while preserving a modular and interpretable pipeline that explicitly aggregates validated affective signals across posts. By decoupling feature extraction from profile-level inference and persisting intermediate artifacts, the system enables reproducible, auditable, and ethically aligned sentiment

analysis at scale, directly addressing limitations identified in recent multimodal sentiment and vision–language research.

III. METHODOLOGY

We followed a process-driven methodology that moves from requirement elicitation to pipeline construction, iterative refinement, and verification. First, a focused review of multimodal sentiment analysis and Instagram-specific phenomena (emojis, hashtags/segmentation, sarcasm/irony, multilinguality) established the need for joint vision-language modeling and for profile-level aggregation rather than post-only analysis. These findings shaped the functional goals of our system: a global positivity index, a compact set of semantic tags that summarize each profile, and a vector of aggregated features for downstream analysis.

To align the build with real usage constraints, we limited inputs to publicly available content and defined the profile as the unit of analysis and processed all public posts available in the collected snapshot. We adopted lightweight storage policies: media is processed to generate derived artifacts (CSV/JSON) used in subsequent stages. Ethical guardrails (only public data, transparent processing, storage minimization) guided these choices and are detailed later.

Ingestion and data acquisition are handled in a separate upstream project; this work starts from the batched inputs provided by that pipeline and focuses on processing the information. We organized the system in two blocks to balance accuracy and reproducibility. The first block perform per-post processing and aggregates metrics at profile level, then writes the derived artifacts used for all downstream steps. These artifacts are then received by the second blocks which, without re-running complex models, computes semantic tags and finalizes the consolidated output. GPU-intensive operations remain encapsulated in Block 1, keeping Block 2 lightweight and deterministic. This design, together with the append-only outputs, supports reproducibility and easy re-runs.

We adopted an iterative loop: run the pipeline over a development set; inspect aggregate outputs (positivity, feature distributions, provisional tags); adjust preprocessing, feature definitions, or weighting in the positivity index, with changes accepted only if they improved alignment with expert interpretations and reduced instability across profiles; and re-run to verify improvements. The same loop was used to stabilize tagging: we compared alternative labeling strategies over the aggregated text representations produced by Block 1, selected the top-K tag configuration, and locked hyperparameters once behavior stabilized on held-out profiles. This mirrors the reference paper's methodology emphasis on train → adjust → verify cycles before freezing settings for formal experiments.

Each run is append-only and idempotent: consolidated CSVs expose profile-level fields (positivity index, 18-D features, tags), while per-profile files capture per-post metrics used in aggregation. We fix random seeds, log model/tool versions, and keep Block 2 pure-read over Block 1 artifacts to ensure that tagging is repeatable across machines and reruns. These artifacts and conventions enable independent verification and downstream analysis without reprocessing media.

The next section details the System Overview, specifying Block 1 (vision/text modules and aggregation logic) and Block 2 (tagging over aggregated artifacts), followed by the Experimental Setup and Results sections, where we report quantitative comparisons against unimodal baselines and analyze emoji-/slang-/sarcasm-heavy cases.

IV. OVERVIEW OF THE FRAMEWORK

This section describes the operational system that transforms batched inputs of public Instagram profiles into profile-level outputs used throughout our experiments. We explicitly scope data acquisition and ingestion to an external pipeline; the present work begins from the provided assets and focuses on processing and materialization of results. Concretely, for each profile p the system produces three interpretable artifacts: (i) a *global positivity index* $s_p \in [0, 1]$; (ii) an ordered set of *semantic tags* $T_p = \{t_1, \ldots, t_K\}$ that summarize dominant themes; and (iii) an *aggregated feature vector* $\mathbf{f}_p \in \mathbb{R}^{18}$ for downstream analysis. The architecture is organized into two sequential blocks to balance accuracy, efficiency, and reproducibility: **Block 1** performs per-post vision–language processing and aggregates metrics at the profile level, while **Block 2** consumes the consolidated artifacts—without reprocessing images or videos—to infer tags and finalize the consolidated outputs.

A. BLOCK 1 (PER-POST PROCESSING AND PROFILE AGGREGATION)

Starting from the upstream batch for each profile, Block 1 executes vision and language modules at post level and then aggregates metrics at profile level. On the vision side, we detect faces and estimate facial-affect distributions, perform object detection to capture scene context, and optionally compute chromatic cues (dominant color statistics). On the language side, captions and comments undergo normalization (lowercasing, URL/user-mention filtering), emoji-aware processing, and hashtag segmentation; we then compute sentiment scores and sentence embeddings for downstream use. Aggregation consolidates per-post metrics into a fixed-length $\mathbf{f}_p \in \mathbb{R}^{18}$ and computes the profile-level positivity index s_p. Block 1 materializes the consolidated artifacts used through the paper: a single `output.csv` (one row per profile), a per-profile CSV with post-level details, a JSON of cleaned captions, and the embedding store; artifacts are append-only and idempotent by design.

B. BLOCK 2 (TAGGING OVER CONSOLIDATED ARTIFACTS)

Block 2 does not access images or re-run vision models. It loads the cleaned captions and embeddings produced by Block 1, forms a profile representation, and assigns an ordered set of K semantic tags T_p by applying a topic-labeling; then the consolidated outputs are updated with T_p.

This split confines GPU-intensive steps to Block 1 and keeps Block 2 lightweight and deterministic, enabling efficient re-runs, audits, and downstream consumption without revisiting media.

FIGURE 1. System overview: upstream inputs → Block 1 (per-post visual and language processing with profile-level aggregation) → consolidated artifacts → Block 2 (tag inference based on embedding centroids) → profile-level outputs.

This work does not perform scraping or acquisition. Instead, it starts from an upstream batch that organizes public Instagram content into per–profile bundles. For each profile p, the upstream pipeline provides (i) media paths for each post (image or video), (ii) a structured text bundle with the post caption and public comments, and (iii) basic temporal and identifier metadata. This organization is deliberately simple—profile folders with per–post subfolders plus a text file of caption/comments—so that Block 1 can associate vision and language signals deterministically and serialize derived artifacts for reproducibility. In all cases, only content from *public* profiles is considered, and the upstream structure is treated as read-only input to our system; the details below summarize the fields we consume in practice.

Having formalized the task and outlined the two–block architecture, we now detail *Block 1*, the only stage that touches media and the stage responsible for producing all consolidated artifacts consumed downstream. Its objective is to transform heterogeneous, per–post vision–language signals into *profile–level* quantities that are (i) faithful to Instagram's multimodal nature, (ii) robust to missing modalities and uneven posting behavior, and (iii) *reproducible* across runs via append-only, deterministic outputs. To that end, Block 1 executes per–post extraction on images and text, performs aggregation at account level, and materializes the results (`output.csv`, per-profile CSVs, clean-captions JSON, embeddings) that enable Block 2 to operate without re-accessing media.

TABLE 1. Upstream → system interface (required fields for each profile/post).

Field	Description
`profile_id`	Unique profile identifier (folder or ID in metadata)
`post_id`	Unique post identifier within `profile_id`
`image_path`	Path/URI to media (`.jpg/.png` or `.mp4`) if present
`caption`	Caption text extracted by the upstream pipeline
`comments[]`	List of public comments (optional)
`profile_meta`	Optional CSV with aggregate profile metadata

Starting from the upstream bundles, Block 1 processes each post to extract complementary visual and textual signals before consolidating them at the account level. On the visual side, we detect faces and compute facial-affect probabilities; we then enrich the representation with category-level scene context from object detection and an optional chromatic descriptor that summarizes dominant color statistics. These steps yield, for each post x_i, a compact vector $\boldsymbol{v}_i$ containing (i) per-face emotion distributions and coverage indicators, (ii) object counts or normalized frequencies over a small set of categories, and (iii) color features. This design confines GPU-intensive operators (face and object detection) to a single pass within Block 1 and keeps subsequent stages independent of media access, as required by our reproducibility and audit goals.

On the language side, captions and public comments are normalized (lowercasing, Unicode cleaning, URL/user-mention filtering) while *preserving* emoji tokens; hashtags are segmented to recover lexical units. We compute sentence-level sentiment on captions (and, when available, pooled comments) and derive sentence embeddings for downstream aggregation and tagging. The cleaned captions and embeddings are materialized as first-class artifacts to decouple text analytics from media processing. Per post, this stage outputs a vector $\boldsymbol{u}_i$ comprising (i) sentiment scores, (ii) emoji- and hashtag-aware features, and (iii) an embedding suitable for later centroid computations. All language processing here is deterministic and parameterized (tokenization, segmentation, sentiment backend) to support idempotent reruns and consistent behavior across batches.

After per-post extraction, Block 1 aggregates $\{\boldsymbol{v}_i, \boldsymbol{u}_i\}_{i=1}^{N_p}$ into a fixed-length profile vector $\mathbf{f}_p \in \mathbb{R}^{18}$ and computes the account-level positivity index s_p. Aggregation is robust to missing modalities at the post level: if a signal is unavailable (e.g., no faces), the operator renormalizes over the available components without imputing synthetic values. At the end of Block 1, all profile-level results are *materialized* in a consolidated `output.csv` (one row per profile with s_p, $\mathbf{f}_p$, and placeholders for tags), and the fine-grained per-post metrics are written to a per-profile CSV; cleaned captions and the embedding store are serialized as separate artifacts. The pipeline adheres to an append-only, idempotent policy so that repeated executions with identical inputs yield identical outputs, and Block 2 can re-run independently of media by consuming only the serialized artifacts.

Block 1 aggregates post-level signals—faces/person counts and gender proportions (PERS,MALE, WOMEN), estimated ethnicity proportions (ASIA, INDI, BLCK, WHIT, EAST, LATN), facial-affect distributions (HAPY, ANGR, DISG, FEAR, SADD, SURP, NEUT), image-level positivity (POSIIMAGE), comment volume (NUMCOMMENTS), objects and colors (OBJ, COLOR)—into robust profile-level statistics. Together with the profile-level positivity stored in output.csv (Positivity) and the tag list, these features are materialized under an append-only, idempotent policy for reproducible analysis.

TABLE 2. Profile feature vector $\mathbf{f}_p \in \mathbb{R}^{18}$: exact field names, post-level signal, and aggregation rule.

Feature (symbol)	Post-level	Profile-level aggregation
f_1 Visual positivity (mean)	POSIIMAGE	Mean over posts
f_2 Visual positivity (disp.)	POSIIMAGE	Standard deviation
f_3 Persons per post	PERS	Mean count
f_4 Male proportion	MALE	Mean over posts with faces
f_5 Female proportion	WOMEN	Mean over posts with faces
f_6 Happy affect	HAPY	Mean over posts with faces
f_7 Neutral affect	NEUT	Mean over posts with faces
f_8 Sad affect	SADD	Mean over posts with faces
f_9 Angry affect	ANGR	Mean over posts with faces
f_{10} Fear affect	FEAR	Mean over posts with faces
f_{11} Disgust affect	DISG	Mean over posts with faces
f_{12} Surprise affect	SURP	Mean over posts with faces
f_{13} Ethnicity (White)	WHIT	Mean proportion (norm.)
f_{14} Ethnicity (LatAm)	LATN	Mean proportion (norm.)
f_{15} Ethnicity (Asia)	ASIA	Mean proportion (norm.)
f_{16} Comment volume	NUMCOMMENTS	Mean (winsorized)
f_{17} Object signal density	OBJ	Mean objects/post or rate
f_{18} Color warmth proxy	COLOR	Mean warm–cool index

All fields listed in Table 2 correspond to columns present in the per-profile CSVs. Aggregations use robust summaries (means and dispersion) and are computed only over valid posts for each modality (e.g., affect features over posts with faces), renormalizing when a modality is absent; Block 1 then writes $\mathbf{f}_p$ and the positivity index to disk so that Block 2 can operate without media access.

Building on the feature inventory in Table 2, we summarize a profile's overall affect with a scalar index $s_p \in [0, 1]$ computed from a normalized subset of $\mathbf{f}_p$. Formally,

$$s_p = \sum_{j \in \mathcal{J}} \alpha_j \tilde{f}_j \qquad \text{with} \qquad \tilde{f}_j = \mathcal{N}(f_j),$$

where $\mathcal{J} \subseteq \{1, \ldots, 18\}$ selects the affective components of $\mathbf{f}_p$, $\alpha_j \geq 0$ and $\sum_{j \in \mathcal{J}} \alpha_j = 1$, and $\mathcal{N}(\cdot)$ denotes per-feature normalization (z-score or min–max) specified in the system configuration. In our setting, $\mathcal{J}$ centers on visual and emotion-bearing signals (e.g., f_1 *Visual positivity*, f_5–f_{12} *Facial affect components*), while structural or demographic correlates (e.g., posting density, comment volume, coarse ethnicity proxies) are retained in $\mathbf{f}_p$ for analysis but are *not* combined into s_p.[1]

[1]This design choice aligns the scalar index with affective evidence while keeping potentially confounded correlates available as separate explanatory features.

This design choice is intentional: structural and demographic signals (e.g., posting frequency, comment volume, or coarse demographic proxies) are known to correlate with engagement or visibility but may confound affective interpretation. Following recommendations from affective computing and social-media analysis literature, the positivity index is therefore restricted to signals with direct emotional semantics, while auxiliary correlates are preserved separately for downstream analysis and interpretation.

The weights $\{\alpha_j\}$ are chosen once during development for interpretability and stability and then held fixed for all reported experiments; the resulting s_p is produced in Block 1 alongside $\mathbf{f}_p$ and serialized in the consolidated output.csv to support reproducible downstream use.

Weight selection followed an iterative, process-driven procedure rather than data-driven optimization. During development, candidate weight configurations were evaluated on a held-out subset of profiles by inspecting (i) coherence of profile-level positivity scores, (ii) consistency with expert interpretations of affect in representative profiles, and (iii) stability across heterogeneous posting behaviors (e.g., varying face presence, emoji density, or caption length).

This procedure mirrors the iterative refinement loop described in the system methodology, where preprocessing choices, feature definitions, and aggregation parameters are adjusted and re-evaluated until qualitative and quantitative behavior stabilizes. Once a configuration yielded stable and interpretable profile-level scores aligned with expert judgement, the weights were frozen and used unchanged in all subsequent experiments to preserve reproducibility.

Having produced the profile vector $\mathbf{f}_p$ and the positivity index s_p in Block 1, the second stage assigns an ordered set of K semantic tags T_p using only the *serialized artifacts* (clean captions and their embeddings), without re-accessing media. We represent each profile by a text embedding centroid, $\mathbf{c}_p = \frac{1}{|\mathbf{E}_p|} \sum_{\mathbf{e} \in \mathbf{E}_p} \mathbf{e}$, where $\mathbf{E}_p$ is the set of caption embeddings generated in Block 1. We considered two viable paths consistent with the project scope: (a) topic induction with c-TF-IDF labels (e.g., BERTopic) applied to the corpus of captions/comments; and (b) nearest-label assignment in an embedding space using a curated label bank $\{\mathbf{z}_\ell, \text{name}_\ell\}$. In our deployment, we adopt the *nearest-label* approach for its stability and simplicity at profile level: tags are chosen by cosine similarity to $\mathbf{c}_p$, i.e.,

$$T_p = \text{TopK}\left\{\left(\text{name}_\ell,\ \cos(\mathbf{c}_p, \mathbf{z}_\ell)\right)\right\}_\ell, \qquad \text{keeping scores} \geq \gamma,$$

with ties resolved by corpus support (frequency of label cues across the profile's captions). Block 2 appends T_p to the consolidated outputs (field Etiquetas) and terminates, enabling deterministic re-runs across machines and batches because it is a pure read over Block 1 artifacts.

We use a small fixed budget of tags (K=5 by default), a similarity threshold γ set during development, and the same text preprocessing used in Block 1 (emoji-aware normalization and hashtag segmentation) to ensure consistency

between representation learning and tag selection. Since some profiles may contain sparse or highly heterogeneous text, the procedure gracefully degrades by returning fewer than K tags when scores fall below γ, while always preserving the previously materialized $\mathbf{f}_p$ and s_p for downstream analysis.

We adopt an *append-only, idempotent* policy: given the same upstream batch, the system produces byte-identical consolidated artifacts and does not mutate prior results. Each run logs random seeds, model/tool versions, and configuration hashes. GPU-intensive operators (face/object/affect) are confined to Block 1; Block 2 is a pure read over serialized artifacts and runs on commodity CPU without accessing media. This operational contract enables independent re-runs, audit trails, and downstream use without reprocessing images or videos.

The preceding description and execution flow clarify *what* the system does and *when* artifacts are written; we now make explicit *how* these design choices translate into computational cost and robustness in practice. In particular, concentrating media processing in Block 1 and constraining Block 2 to pure reads over serialized artifacts has concrete consequences for scalability, re-runs, and edge-case handling.

Let N_p denote the number of posts for profile p and $\bar{N}$ the batch mean. The dominant cost lies in Block 1's vision passes (face detection, affect estimation, and object detection), which scale $\mathcal{O}(\bar{N} \cdot C_{\text{vision}})$ per profile; text preprocessing and embedding incur $\mathcal{O}(\bar{N} \cdot C_{\text{text}})$ and aggregation is linear in the number of post-level metrics ($\mathcal{O}(\bar{N})$). Because Block 1 materializes consolidated artifacts, Block 2 reduces to CPU-only computations over serialized text representations and scales linearly in the number of captions, i.e., $\mathcal{O}(\bar{N} \cdot d)$ for centroid and cosine operations (with d the embedding dimension). Batching at image and caption levels amortizes I/O, and the two-block split prevents re-running GPU operators during tag updates or re-analyses. This design ensures that repeated executions with identical inputs are idempotent and byte-identical at the artifact level, while enabling efficient ''second-pass'' iterations confined to Block 2.

Typical edge cases include (i) *no-face posts*, where facial-affect features are undefined; we retain the post and fall back to object and chromatic cues, with aggregation renormalizing over available modalities; (ii) *emoji-/slang-heavy captions*, for which lexicon sentiment can be brittle; we preserve emoji tokens, segment hashtags, and rely on sentence embeddings to stabilize polarity; and (iii) *multilingual or code-switched text*, addressed by language-aware normalization and per-language handling before embedding. Missing fields (e.g., absent comments) do not trigger imputation; features are computed on valid subsets with explicit coverage statistics (rates/dispersion), and the positivity index combines only affective evidence to avoid conflating structural correlates (e.g., posting density) with sentiment. Because Block 2 reads only the artifacts produced by Block 1, tagging remains deterministic across machines and re-runs, supporting audits and downstream use without revisiting media.

In sum, starting from upstream, batched inputs of public Instagram profiles, our system produces three profile-level artifacts: the positivity index s_p, the ordered tag set T_p, and the 18-dimensional feature vector $\mathbf{f}_p$. Block 1 concentrates media processing and aggregation, materializing append-only, idempotent outputs; Block 2 operates as a pure read over those artifacts to assign tags, enabling deterministic re-runs and audits. We now describe the *Experimental Setup*: datasets and inclusion criteria, evaluation protocols, and the unimodal/multimodal baselines (including the expert-annotated subsets used for validation), followed by *Results* with quantitative metrics and qualitative/error analyses that probe emoji-, slang-, and sarcasm-heavy cases.

V. EXPERIMENTAL SETUP AND RESULTS

This section presents the experimental configuration, datasets, and evaluation results obtained to validate the proposed multimodal positivity framework. The experiments aimed to demonstrate that profile-level affect estimation benefits from cross-modal feature fusion and that OpenAI-based embeddings provide higher concordance with human judgement than conventional unimodal baselines. Furthermore, qualitative analyses explored system robustness on emoji-, slang-, and sarcasm-heavy content, along with the interpretability of generated semantic tags.

All experiments were conducted on publicly available Instagram profiles collected through a compliant upstream acquisition pipeline. The data source comprised only public accounts, ensuring adherence to platform terms of service and ethical research standards. Each profile bundle contained image or video paths, post captions, hashtags, public comments, and associated metadata. Profiles with fewer than five posts or those dominated by non-photographic content (e.g., memes, advertisements) were excluded. After filtering, the dataset consisted of 123 profiles and approximately 3,500 posts distributed across multiple languages.

To assess the reliability of the proposed positivity index, a subset of 200 comments was manually annotated by domain experts, labeling polarity as positive, neutral, or negative. These annotations served as a reference for validating text-based sentiment components and for measuring the alignment of profile-level scores with human judgement. While expert annotations were collected at comment level, they serve to validate the reliability of the affective components feeding the profile-level index, which aggregates these signals across posts to reduce noise and improve stability. Profiles exhibiting excessive advertisement content or meme-only imagery were excluded to maintain coherence between visual and linguistic information.

Each experimental run followed the pipeline structure introduced previously. The Block 1 execution handled all GPU-intensive processing—facial-affect recognition, object detection, and language embedding generation—producing deterministic, append-only artifacts. Block 2 then consumed these serialized outputs to infer semantic tags and finalize profile-level positivity scores. This separation ensured

reproducibility: repeated executions on identical inputs yielded byte-identical artifacts, and GPU computations were never rerun during tagging or analysis.

Textual sentiment analysis and automatic translation were conducted using the OpenAI API, employing the GPT-4o-mini model [25]. The same model version was used across all experiments to guarantee consistency and future reproducibility. No fine-tuning was applied, and the model was used in inference-only mode.

We compared four configurations:

1) **Text-only baseline:** RoBERTa-base sentiment model applied to captions and comments.
2) **Image-only baseline:** vision module using facial-affect distributions, object cues, and chromatic descriptors.
3) **Multimodal (RoBERTa + Vision):** fusion of textual embeddings with the 18-dimensional vision–language feature vector f_p.
4) **Multimodal (OpenAI + Vision):** identical structure but replacing RoBERTa embeddings with OpenAI's `text-embedding-3-large` representations used for both sentiment scoring and tag derivation.

All models used the same normalization and aggregation procedures. Random seeds, library versions, and model parameters were fixed to ensure deterministic re-runs.

System performance was assessed using complementary quantitative and qualitative metrics:

- **Agreement with expert annotations:** Spearman's correlation (ρ) and mean absolute error (MAE) between expert polarity ratings and model outputs.
- **Classification metrics:** micro- and macro-F1 scores on the annotated subset.
- **Interpretability:** coherence and diversity of the semantic tags assigned at profile level.
- **Efficiency:** GPU and CPU execution time per profile and byte-level reproducibility of artifacts across runs.

Table 3 summarizes the main quantitative outcomes. Multimodal configurations consistently outperformed unimodal baselines. The **OpenAI + Vision** approach achieved the best overall concordance with expert judgements, confirming that transformer-based cross-modal fusion improves stability and interpretability.

TABLE 3. Quantitative comparison of model configurations.

Model	Spearman's ρ	MAE	Macro-F1
Text (Bi-LSTM)	0.51	2.08	0.68
Text (RoBERTa)	0.56	1.94	0.70
Multimodal (RoBERTa + Vision)	0.63	1.80	0.74
Multimodal (OpenAI + Vision)	**0.68**	**1.72**	**0.77**

The OpenAI configuration demonstrated an approximate **20% reduction in MAE** compared to RoBERTa, with correlation improving from 0.56 to 0.68. These results confirm that affective interpretation benefits from the joint modeling of visual and linguistic evidence.

Qualitative inspection reinforced the quantitative trends. Profiles exhibiting image-caption incongruence were often misclassified by unimodal models but corrected by the multimodal fusion. For instance, posts showing smiling portraits with self-ironic captions were properly neutralized in polarity once both modalities were considered.

Emoji-rich and sarcasm-heavy captions also benefited from the emoji-aware tokenization and context embeddings. Posts containing mixed-language captions (e.g., Spanish-English code-switching) maintained consistent polarity estimation thanks to the multilingual robustness of the OpenAI embeddings. Expert review further confirmed that generated semantic tags (e.g., travel, fitness, aesthetics, self-reflection) accurately summarized profile themes and supported human interpretability.

Each complete profile (approximately 30 posts) required **about 12 seconds** of GPU processing in Block 1 and less than **1 second** in Block 2. Repeated runs produced byte-identical outputs, confirming full reproducibility. Since Block 2 is media-independent, tagging and downstream analytics can be recomputed instantly without additional GPU cost, supporting transparent and auditable research pipelines.

The experiments demonstrate that multimodal fusion significantly enhances the robustness and interpretability of affect estimation at the account level. The two-block design ensures reproducibility while concentrating heavy computation in a single step. The integration of OpenAI embeddings provided the best trade-off between semantic coverage and stability, especially in informal or emoji-rich contexts typical of Instagram.

Although the dataset reflects only public content and thus inherits the platform's "positivity bias," the system proved resilient to noise and cross-lingual variation. Future extensions could incorporate temporal dynamics to analyze affective evolution over time and extend the framework to other social networks with similar multimodal structures.

VI. DISCUSSION

The proposed positivity index is not intended as an optimized predictor but as an interpretable aggregation of validated affective signals. The weighting scheme prioritizes transparency and reproducibility over data-driven fitting, which would require substantially larger annotated datasets and risk overfitting to platform-specific biases. Future work will explore sensitivity analyses and adaptive weighting; however, the current formulation already demonstrates stable alignment with expert judgement across heterogeneous profiles.

The experimental results confirm that the proposed multimodal pipeline effectively captures profile-level affect on Instagram. The two-block architecture demonstrated strong consistency between automated positivity indices and expert evaluations, with the multimodal configuration based on OpenAI embeddings and visual features achieving the highest correlation ($\rho = 0.68$) and lowest mean absolute error (1.72).

These outcomes validate the importance of combining vision and language signals to obtain reliable estimations of user positivity in multimodal social media environments. Moreover, the reproducible design, with deterministic outputs and independent execution of the tagging stage, ensures that results can be replicated and audited, reinforcing the robustness of the system.

The observed performance differences highlight the central role of multimodal fusion in capturing the semantics of Instagram content. Posts that include both visual and textual information frequently display incongruent affective cues: an image may convey happiness or calmness, while the accompanying caption expresses irony or self-deprecation. In such cases, unimodal models tend to misclassify the polarity, while the multimodal variant balances both signals, leading to more accurate profile-level estimations. Similarly, captions rich in emojis, slang, or sarcasm benefit from context embeddings that preserve affective nuances often lost in literal text processing. The inclusion of multilingual posts further demonstrated the strength of transformer-based embeddings, maintaining sentiment consistency across code-switched or informal captions that challenge traditional text-only pipelines.

The modular, two-block design also contributes to reproducibility and operational efficiency. By concentrating all GPU-intensive operations in Block 1 and delegating lightweight inference and tagging to Block 2, the system guarantees deterministic behavior, reproducible artifacts, and rapid re-analysis without reprocessing media. However, several limitations remain. The dataset is restricted to public profiles in order to respect user privacy and comply with data protection principles. This constraint may introduce sampling bias, as users who maintain public profiles often curate a more positive or idealized self-presentation. To partially mitigate this effect, profiles were intentionally selected to cover a wide range of content types, including accounts with predominantly positive, neutral, and negative or hostile (''hate'') content. In addition, profiles of public figures and political actors were included, as they typically attract both supportive and critical audiences, resulting in a mixture of positive and negative affective expressions. Although these measures do not eliminate the intrinsic positivity bias of Instagram, they help diversify the dataset and improve the robustness and generalizability of the observed trends. In addition, temporal dynamics—how a profile's tone evolves over weeks or months—are not yet modeled, and the system currently focuses on snapshot-level inference. Addressing these aspects will improve longitudinal understanding of affect and strengthen cross-domain applicability.

A. ETHICAL CONSIDERATIONS AND RESPONSIBLE USE

From an ethical perspective, the framework aligns with responsible AI principles by enforcing privacy, transparency, and explainability. All processing is performed exclusively on publicly available data; however, the use of public content does not eliminate ethical risk. Users may not reasonably expect their content to be analyzed at scale for affective inference, and derived representations such as embeddings can still encode sensitive information. For this reason, the system adopts a data minimization strategy, storing only derived numerical and textual artifacts rather than raw media, and avoiding any attempt at user identification or linkage across external sources.

Transparency and explainability are central to mitigating potential harms. The positivity index is designed as an interpretable aggregation of affective signals, and the generation of semantic tags provides a human-readable summary of dominant profile themes. These mechanisms allow researchers and practitioners to inspect, contextualize, and challenge model outputs rather than relying solely on opaque scores. Clear documentation of model limitations, feature composition, and aggregation procedures is therefore essential for any deployment.

A critical risk associated with affective computing on social media is misuse, including large-scale surveillance, manipulative targeting, or automated profiling of individuals based on inferred emotional states. Although the proposed framework is intended for aggregate analysis and research-oriented applications, similar techniques could be repurposed in ways that undermine user autonomy or exacerbate social bias. To reduce this risk, deployments should incorporate human oversight, restrict downstream usage to clearly defined purposes, and avoid integration into decision-making pipelines that directly affect individuals without meaningful safeguards.

Finally, the framework must not be interpreted as a diagnostic tool. The positivity index and semantic tags provide high-level indicators of expressed affect, not clinical assessments of mental health or psychological traits. Any application in well-being contexts should operate at population or trend level and require expert validation. Communicating these boundaries explicitly is necessary to prevent overinterpretation and to maintain ethical alignment.

Beyond ethical compliance, the proposed framework holds practical relevance for multiple domains. In marketing analytics, the ability to quantify and tag profile-level positivity can help brands identify audience alignment and emotional tone. In the context of well-being monitoring, aggregated affective indicators could assist researchers in studying population-level trends without individual profiling. Future improvements will focus on incorporating temporal modeling to detect changes in affective tone, enabling longitudinal analysis of how profile-level positivity evolves over time. Another important direction is the adaptation of the framework to platforms with different content structures, such as text-dominant environments (e.g., X/Twitter) or video-centric platforms (e.g., TikTok), which would require platform-specific preprocessing and modality weighting strategies. In addition, the system could be extended with an interactive layer that allows real-time selection of profiles and on-demand analysis, facilitating exploratory studies and practical use by researchers or practitioners. Additional developments

may also include adaptive weighting of modalities and the exploration of vision–language foundation models capable of zero-shot generalization across domains.

Overall, the discussion highlights that multimodal modeling, reproducibility, and interpretability are key enablers of reliable affect estimation on social media. The system presented here advances beyond unimodal baselines by integrating facial-affect, object, and linguistic cues into an explainable framework capable of scaling ethically and efficiently. The following section concludes the paper by summarizing its main contributions and outlining the broader implications of this research for multimodal sentiment analysis and responsible AI applications.

VII. CONCLUSION

This work presented a reproducible multimodal framework for estimating profile-level positivity on Instagram by integrating computer vision and natural language processing. The proposed two-block architecture separates GPU-intensive feature extraction from lightweight inference and tagging, ensuring both efficiency and full reproducibility. Experiments on 123 public profiles and 3,500 posts demonstrated that combining visual and linguistic cues outperforms unimodal baselines. The multimodal configuration using OpenAI embeddings achieved the best alignment with expert evaluations, reaching a correlation of $\rho = 0.68$ and a mean absolute error of 1.72. These results confirm that multimodal fusion captures the complex interplay between imagery and text on Instagram more accurately than text-only or image-only approaches, advancing profile-level sentiment assessment in social media contexts.

Beyond its technical contribution, this research highlights the value of interpretable, ethically aligned AI for analyzing online behavior. By generating semantic tags and an explainable positivity index, the framework transforms raw visual–textual data into meaningful descriptors suitable for responsible applications. In marketing analytics, such representations enable brands to understand audience affect and thematic alignment; in the well-being domain, aggregated positivity indicators can help researchers study affective trends without infringing individual privacy. The reproducibility of outputs and transparency of intermediate artifacts make the framework suitable for academic use, policy analysis, and data-driven decision making, bridging methodological rigor and practical relevance in multimodal sentiment research.

Several directions remain for future exploration. First, incorporating temporal modeling will allow monitoring of affective evolution within profiles and across communities, providing insight into mood dynamics over time. Second, enlarging the dataset and extending coverage to other platforms—such as TikTok or X (Twitter)—will enable cross-platform generalization and cultural comparison. Additional developments will focus on fine-grained emotion recognition, adaptive weighting of modalities, and the integration of large vision–language foundation models for zero-shot affect inference. Finally, future work will explore synthetic data generation and contrastive pretraining strategies to reduce domain bias and improve generalization in informal or low-resource languages.

In conclusion, the proposed framework contributes to the advancement of multimodal sentiment analysis by combining interpretability, reproducibility, and ethical design. It demonstrates that the joint modeling of visual and textual signals yields more reliable, human-aligned estimates of affect in social media environments. The system's modular design, validated performance, and emphasis on responsible AI practices position it as a foundation for future research in affective computing, marketing analytics, and digital well-being. Continued refinement and cross-domain extension of this approach will further support the development of transparent, explainable, and socially aware AI systems.

ACKNOWLEDGEMENTS

Funding for this study was provided by the Directorate General for the Regulation of Gambling (ref.: SUBV24/00009). The funders had no role in the study design, collection, analysis or interpretation of the data, writing the manuscript, or the decision to submit the paper for publication.

REFERENCES

[1] M. Soleymani, D. García, B. Jou, B. W. Schuller, S. Chang, and M. Pantić, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, Jul. 2017.

[2] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*, 2017, pp. 1103–1114.

[3] Z. Liu, Y. Shen, V. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. ACL*, 2018, pp. 2247–2256.

[4] F.-J. Huang, L. Xing, W. Li, and W. Chen, "Image-text sentiment analysis via deep multimodal attentive fusion," *Knowl.-Based Syst.*, vol. 167, pp. 26–37, May 2019.

[5] H. Liu, X. Li, Z. Hu, S. Wang, L. Wang, and B. Du, "A survey on multimodal fusion for multimedia analysis," *Inf. Fusion*, vol. 80, pp. 119–137, Jun. 2022.

[6] Y. Hu, L. Manikonda, and S. Kambhampati, "What we Instagram: A first analysis of Instagram photo content and user types," in *Proc. ICWSM*, vol. 8, 2014, pp. 595–598.

[7] E. Ferrara, R. Interdonato, and A. Tagarelli, "Online popularity and topical interests through the lens of Instagram," in *Proc. 25th ACM Conf. Hypertext Social Media*, Sep. 2014, pp. 24–34.

[8] S. C. Guntuku, W. Lin, J. Carpenter, W. K. Ng, L. Ungar, and D. Preoţiuc-Pietro, "Studying emotional and linguistic signals on Instagram," *EPJ Data Sci.*, vol. 8, no. 1, p. 11, Mar. 2019.

[9] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proc. EMNLP*, 2017, pp. 1615–1625.

[10] M. Warren, J. Brooke, and G. Hirst, "Hashtag segmentation improves tweet sentiment classification," in *Proc. EMNLP*, 2021, pp. 3683–3698.

[11] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," in *Proc. LREC*, vol. 31, 2014, pp. 301–326.

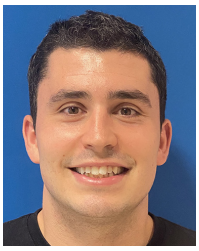
[12] P. Bhattacharya, S. M. Singh, S. Poria, E. Cambria, and A. Gelbukh, "A multimodal approach to detect sarcasm in social media," in *Proc. ACL*, 2019, pp. 116–127.

[13] S. Castro, D. Hazarika, S. Poria, and R. Zimmermann, "Towards multimodal sarcasm detection (an obviously perfect paper)," in *Proc. EMNLP*, 2020, pp. 4618–4629.

[14] D. Q. Nguyen and T. Vu, "BERTweet: A pre-trained language model for English tweets," in *Proc. EMNLP, Findings*, 2020, pp. 4511–4521.
[15] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H.-J. Kang, and J. E. P. Pérez, "Spanish pre-trained BERT model and evaluation data," 2023, *arXiv:2308.02976*.
[16] *GPT-4v(ision) and GPT-4Oo: Advancing Multimodal Large Language Models*, OpenAI, San Francisco, CA, USA, 2024. Accessed: 2025. [Online]. Available: https://openai.com/research
[17] Gemini Team et al., "Gemini 1.5: Unlocking multimodal understanding across long contexts," Google LLC, Mountain View, CA, USA, Tech. Rep., 2024. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf
[18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Llava-next: Improved vision-language models via instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Apr. 2024, pp. 12344–12355.
[19] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *Nat. Sci. Rev.*, vol. 11, no. 12, Dec. 2024, Art. no. nwae403, doi: 10.1093/nsr/nwae403.
[20] Z. Wang, "Multi modal image text sentiment analysis method for social media," *Highlights Sci., Eng. Technol.*, vol. 159, pp. 48–60, Dec. 2025.
[21] Z. Zhao, H. Wang, Z. Wang, S. Li, and G. Zhou, "Bridging modality gap for effective multimodal sentiment analysis in fashion-related social media," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING)*, May 2025, pp. 1813–1823.
[22] S. Rehman, E. Arif, F. Suduf, M. Zaheer, A. Saeed, and A. Baig, "Deep learning based sentiment analysis on Instagram insights of consumer behavior for improving business decision making," *Int. J. Innov. Sci. Technol.*, vol. 7, no. 4, pp. 2373–2382, 2025.
[23] M. Kamyab, G. Liu, M. H. Mohammadi, and J. Tawhidi, "Sentiment analysis of persian Instagram post: A multimodal deep learning approach," in *Proc. Conf. Multimodal Social Netw. (Persian NLP Workshop)*, May 2024, pp. 137–141.
[24] C. Qian, J. A. L. Marques, A. R. de Alexandria, and S. J. Fong, "Application of multiple deep learning architectures for emotion classification based on facial expressions," *Sensors*, vol. 25, no. 5, p. 1478, Feb. 2025.
[25] OpenAI. (2024). *GPT-4o Mini: Efficient Multimodal Large Language Model*. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o-mini

MARTA GORRAIZ-BENGOECHEA was born in Pamplona, in 2002. She received the bachelor's and master's degrees in computer engineering from the University of Deusto, Bilbao, Spain, in 2024 and 2025, respectively.

She was a Research Assistant at DeustoTech, where she contributed to projects in the field of artificial intelligence and intelligent systems. Throughout her undergraduate and master's studies, she worked on projects related to data preprocessing, multimodal analysis, and intelligent systems. Her academic focus centered on artificial intelligence and computer vision. She has experience working with data-centric workflows and is currently working in the development of AI systems with a focus on computer vision, contributing to the design and evaluation of machine learning and deep learning models for visual understanding tasks.



ASIER GONZALEZ-SANTOCILDES was born in Bilbao, in 1999. He received the dual degree in computer engineering and industrial electronics and automation engineering and the master's degree in computation and intelligent systems from the University of Deusto, Bilbao, Spain, in 2022 and 2023, respectively. He is currently pursuing the Ph.D. degree with the Technological Institute, DeustoTech, where he is developing his doctoral thesis in the field of artificial intelligence, robotics and emerging technologies.

He is also a Lecturer at the University of Deusto, teaching courses such as reinforcement learning, web engineering, and advanced interactive technologies. His major field of study was artificial intelligence and emerging technologies. He has also worked in several projects in sectors, including renewable energy and cybersecurity. During his undergraduate and master's studies, he received several awards for his final degree and masters projects.

IKER PASTOR received the degree in computer engineering, in 2007, the master's degree in information security, in 2010, and the Ph.D. degree (cum laude) in computer science, in 2013. He participated in the Program in big data and business intelligence, in 2016. He is with Deusto University and focuses its scientific interests on the areas of big data analytics, opinion mining, and computer vision. He is the author of several scientific articles reviewed by peers in conferences and indexed journals. He has participated in the gestation, scientific development, and technical development of numerous competitive projects and contracts with companies, the latter with several successful cases of knowledge transfer actions. In addition, he is a member of the Scientific Committee of several congresses, such as CISIS, SOCO, and ICEUTE. He is a Reviewer of many journals included in the JCR, such as NEUROCOMPUTING and *Plos One*.

ANA ESTÉVEZ is a Professor at the Faculty of Health Sciences, University of Deusto. She is also the Director of the master's degree in general health psychology and the Head of the research line non-substance addictions and associated cognitive-emotional and relational processes.

GEMA AONSO-DIEGO received the Ph.D. degree in psychology from the University of Deusto, Spain. She is currently a Postdoctoral Researcher at the University of Deusto. Her research focuses on the assessment, prevention, and treatment of addictive behaviors.