

Mental Health Discourse on TikTok: Interpreting Multimodal Short-Form Videos at Scale

Mingyue Zha¹, Ho-Chun Herbert Chang¹

¹ Program in Quantitative Social Science, Dartmouth College, Hanover, NH 03755 USA

Abstract

Short-form video platforms have become prominent spaces for mental health disclosure, peer support, and seeking information. They integrate text, visuals, and audio into a single stream of communication. However, much empirical research in digital mental health examines these modalities in isolation. This study introduces a pipeline combining automated multimodal feature extraction with Shapley value-based interpretability to analyze how text, visuals, and audio jointly influence engagement with mental health content. Applying this framework to 162,965 TikTok videos and 814,825 images about social anxiety disorder (SAD), we find that (1) visual facial expressions are more predictive of engagement than textual sentiment; (2) informational content attracts more attention than emotional support; and (3) cross-modal interactions exhibit threshold-dependent effects on reach. These findings advance our understanding of mental health communication in multimodal environments, demonstrating how platform algorithms privilege certain forms of mental health communication over others. Methodologically, this work contributes a reproducible and interpretable framework for multimodal research applicable across domains.

Introduction

Digital spaces present a double-edged sword for mental health. They provide community for information-seeking while amplifying anxiety around self-expression. The trade-off of information seeking and self-expression has not been widely explored, especially in multimodal contexts. Social media platforms such as TikTok increasingly prioritize multimodal communication over text-based exchanges by virtue of their **platform design** (Huang, Lei, and Chen 2025; Xie, Lin, and Bai 2025). As such, short-form video has emerged as a central medium for mental health disclosure, peer support, and informal health information-seeking. This shift calls for new forms of multimodal literacy to better understand how mental health content is produced and engaged with online (Niu et al. 2021).

Multimodal content is consequential for mental health communication. In terms of prevalence, 56% of Gen Z report using TikTok for health and wellness advice, with many specifically seeking mental health content (Hall and Keenan 2025). More fundamentally, multimodal content shapes both how emotional and mental states are conveyed by creators

and how they are perceived by audiences. Research on non-verbal communication shows that emotions such as distress, vulnerability, and reassurance are communicated not through words alone but through tone of voice, facial expression, pacing, and visual context (Horstmann 2003). Short-form video platforms are also engineered to reward content that sustains attention and drives engagement, with recommendation algorithms amplifying expressive visual and auditory signals regardless of content quality or clinical accuracy (Bickham et al. 2024). The visibility of mental health content is shaped not only by what creators say, but how effectively they deploy multimodal cues. Understanding which combinations of signals amplify or suppress algorithmic reach is important for assessing the supportive and potentially harmful dynamics of these platforms.

Despite the growing importance of multimodal content in mental health contexts, methodological tools for analyzing such content at scale remain underdeveloped. Existing research focuses on single modalities (most commonly text) (Islam et al. 2025), while multimodal studies often rely on small samples and manual annotation (Wu et al. 2025). As a result, there is limited consensus on how to systematically extract and interpret multimodal signals from a large dataset of short-form videos (Ali and Molla 2025).

Recent advances in machine learning, including pre-trained language models, image recognition, and audio embeddings, make such integration computationally feasible at scale (Wang et al. 2023). Emerging studies demonstrate that LLM-assisted content analysis can effectively extract and annotate visual features from online videos (Liu et al. 2025). Building on these capabilities, this paper introduces a scalable pipeline for multimodal analysis of short-form social media videos, integrating automated feature extraction across text, image, and audio modalities with Shapley value-based interpretability. This approach enables direct comparison of heterogeneous features within a common explanatory framework, attributing attention and engagement outcomes to both individual modal features and their interactions, within an interpretable framework designed to support social scientific inference.

Applying this pipeline to 162,965 TikTok videos and 814,825 frames related to social anxiety disorder (SAD), we find that facial expressions outperform textual sentiment in predicting viewership, informational content commands

greater attention than emotional support across both text and audio, and cross-modal combinations exhibit threshold-dependent effects on reach. Together, these results suggest TikTok functions as an informal health information system, one whose recommendation dynamics may privilege certain types of mental health communication over others. Our findings have implications on platform governance, content moderation design, and the well-being of vulnerable users.

Literature Review

Social Anxiety Disorder and Social Media

Social anxiety disorder (SAD) is among the most prevalent mental health conditions globally, affecting 7–13% of people across the world and ranking as the third most common psychiatric disorder after depression and substance use disorders (Stein and Stein 2008; Kessler et al. 2005). Characterized by intense, persistent fear of negative evaluation in social or performance situations, SAD is associated with substantial functional impairment across occupational, academic, and interpersonal domains (Fehm et al. 2005). Despite its prevalence and burden, SAD remains undertreated, with a majority of individuals never receiving formal diagnosis or care due to the disorder itself hindering help-seeking (Alonso et al. 2018). This treatment gap intensified interest in informal pathways through which SAD individuals seek information and find community, especially online.

Consistent with this, research finds that individuals high in social anxiety report greater comfort with online compared to face-to-face communication, and that computer-mediated interaction can serve as a lower-stakes arena for social skill development and connection-seeking (Prizant-Passal, Shechner, and Aderka 2016). Early “online disinhibition” frameworks proposed that the relative anonymity and asynchronicity of internet communication would reduce the social evaluative threat that defines SAD, potentially enabling more authentic self-expression (Suler 2004). However, longitudinal evidence suggests that excessive reliance on online interaction as a substitute for in-person contact may reinforce avoidance behaviors, thereby maintaining rather than alleviating anxiety over time (Weinstein et al. 2015). The platforms individuals with SAD turn to for safety may also cause harm.

The content produced within SAD-related digital communities reflects a heterogeneous communicative ecosystem that maps onto established typologies of social support. Cutrona and Suhr’s framework distinguishes emotional, informational, instrumental support (tangible assistance), and appraisal support (evaluative feedback for self-assessment) (Suhr et al. 2004). Research on health-related online communities has demonstrated that these support functions are not uniformly valued or algorithmically rewarded, and that the visibility of particular content types shapes community norms over time (Naslund et al. 2016). For SAD specifically, problems arise when informational content produced by unqualified peer creators may displace professionally vetted psychoeducation, while emotionally supportive exchanges—arguably the most clinically meaningful function of peer communities—may be systemati-

cally deprioritized by engagement-optimizing recommendation systems. Recent analyses of mental health content on TikTok have documented high rates of clinical inaccuracy and the widespread propagation of unsubstantiated coping advice, raising urgent questions about the role of platform design in shaping the information environment available to vulnerable users (Rathbone and Prescott 2017).

Mental Health Disclosure in Digital Spaces

Research on mental health disclosure online reveals users have complex motivations: individuals seek emotional validation, practical coping strategies, reduced stigma through normalized discussion, and connections with others sharing similar experiences (Andalibi, Ozturk, and Forte 2017).

Self-disclosure theory suggests that revealing personal information serves relationship-building functions. Online disclosures are often public, persistent, and algorithmically mediated, reaching audiences that extend far beyond immediate social networks (Bazarova and Choi 2014). The permanence and visibility of digital disclosures can amplify both benefits (finding community, reducing isolation) and risks (privacy concerns, unwanted attention, stigmatization). For individuals with social anxiety disorder specifically—characterized by intense fear of social evaluation and performance situations—the act of creating video content that foregrounds the self presents a notable paradox. Visual self-presentation, the core anxiety trigger for those with SAD, becomes the primary medium through which individuals discuss their social anxiety experiences.

Recent work examining TikTok mental health content specifically has found that algorithmic recommendation systems shape which mental health narratives become visible, with potential implications for both helpful peer support and misinformation spread (Bickham et al. 2024). Milton et al. (2023) documented how TikTok users with mental health conditions find community and validation through algorithmically curated content.

Multimodal Communication in Digital Spaces

The shift from text-dominant to multimodal social media platforms represents a fundamental transformation in online communication. Kress and van Leeuwen argue that contemporary communication operates through integrated semiotic systems rather than verbal language alone, where visual, textual, and auditory elements function as distinct but interconnected meaning-making resources (Kress and Van Leeuwen 2001). In digital contexts, these modalities interact, sometimes reinforcing and sometimes contradicting one another.

Short-form video platforms like TikTok, Instagram Reels, and YouTube Shorts exemplify this multimodal complexity. These platforms privilege visual performance and algorithmic curation over traditional text-based discourse. The affordances of these platforms—vertical video format, music integration, text overlay capabilities, rapid content consumption—create communicative environments where success depends on coordinating elements across modalities. Media richness theory suggests that face-to-face communication is traditionally considered the “richest” medium due to its integration of verbal, vocal, and visual cues (Daft and

Lengel 1986). Short-form video platforms approximate this richness in asynchronous, algorithmically mediated formats.

Multimodal Analysis of Mental Health Content: Current Approaches and Limitations

Research on social media viewership has largely proceeded along single-modality lines. Text-based studies examine how linguistic features such as sentiment, lexical diversity, and style predict popularity outcomes (Vilares, Alonso, and Gómez-Rodríguez 2015). Visual analysis focuses on image aesthetics, color composition, facial presence, and object recognition (Wu et al. 2017). Audio studies, though less common, investigate music characteristics, vocal prosody, and speech patterns (Li et al. 2024). These single-modality approaches have generated valuable insights but necessarily provide incomplete accounts of how multimodal content functions.

The few existing multimodal studies typically combine features from two modalities—most commonly text and images—and feed concatenated feature vectors into prediction models (Meghawat et al. 2018). While this approach improves predictive accuracy over single-modality models, it treats multimodal features as a flat input space and does not assess the contributions of individual modalities or cross-modal interactions. More critically, studies often rely on small samples with manual annotation, limiting the scale and generalizability of the analysis. As Chen et al. (2025) demonstrated, multimodal annotation requires substantial human labor, and the rapid evolution of platform norms means manually coded datasets quickly become outdated.

Recent work in computer science uses deep learning to analyze how visual, audio, and textual features jointly predict TikTok creator influence tiers (Tricomi et al. 2024). However, as is common with deep learning, embeddings sacrifice interpretability. Researchers cannot easily identify which features drive predictions or how features of different modalities interact. For social scientists interested in understanding the mechanisms through which content characteristics shape attention, prediction accuracy alone is insufficient.

Interpretability in Computational Social Science

For social scientists, interpretability is not merely a technical desideratum—it is fundamental to scientific generalizability and normatively, the design of interventions. Shapley values, derived from cooperative game theory, offer a principled approach to model interpretation (Winter 2002). The core insight is to treat each feature as a “player” contributing to a prediction “payout,” and to compute each feature’s contribution by considering its average marginal effect across all possible coalitions of other features. SHAP (SHapley Additive exPlanations) operationalizes this concept for machine learning models, providing a unified framework that decomposes any prediction into additive feature contributions (Lundberg and Lee 2017).

Recent applications of SHAP in computational social science have demonstrated its value for understanding feature importance in tabular data settings (Salih et al. 2025). However, SHAP has not been systematically applied to multimodal analysis in ways that leverage modality structure.

Standard global SHAP summaries (e.g., mean absolute attribution) exist, but they do not provide a coefficient-like directional summary conditioned on feature presence/intensity, which is often what social scientists want.

The existing literature in digital mental health has several gaps. First, little prior work has used a comprehensive automated pipeline for extracting theory-driven features across modalities from videos at scale. Second, while multimodal prediction models exist, they often lack the interpretability needed to understand how specific features within and across modalities shape outcomes. Third, mental health communication research has not systematically examined how affective expression diverges across modalities.

Research Questions

This study addresses these gaps by introducing an automated multimodal analysis pipeline that integrates zero-shot classification for semantic features with Shapley value-based interpretability. We examine three research questions:

1. **RQ1:** How do textual, visual, and audio sentiment diverge in characterizing mental health-related discourse?
2. **RQ2:** Which features across text, visual, and audio modalities most strongly predict viewership of mental health content?
3. **RQ3:** How do cross-modal combinations of textual, visual, and audio features influence attention dynamics in short-form mental health videos?

By answering these questions, we contribute both substantively (insight into multimodal mental health communication) and methodologically (a reproducible framework for interpretable multimodal analysis).

Methods

Overview of Multimodal Pipeline

Our proposed analytical pipeline integrates textual, visual, and audio information from social media videos into a unified, interpretable multimodal framework. Figure 1 provides a schematic overview of the pipeline.

Each video is decomposed into three modalities, which are processed independently using modality-appropriate feature extraction methods. Textual content is analyzed using sentiment analysis and zero-shot discourse labeling, visual frames are coded using facial analysis and zero-shot visual labeling, and audio tracks are analyzed for acoustic features with spoken content transcribed for downstream textual analysis. Across modalities, zero-shot labeling produces probabilistic feature representations that enable additive and comparable attribution across heterogeneous feature types. All features are concatenated into a unified multimodal model, and Shapley value analysis is applied to quantify feature- and modality-level contributions to model predictions. The full experimental pipeline was executed on a consumer-grade workstation equipped with an AMD Ryzen CPU and an NVIDIA GeForce RTX 4090 GPU.

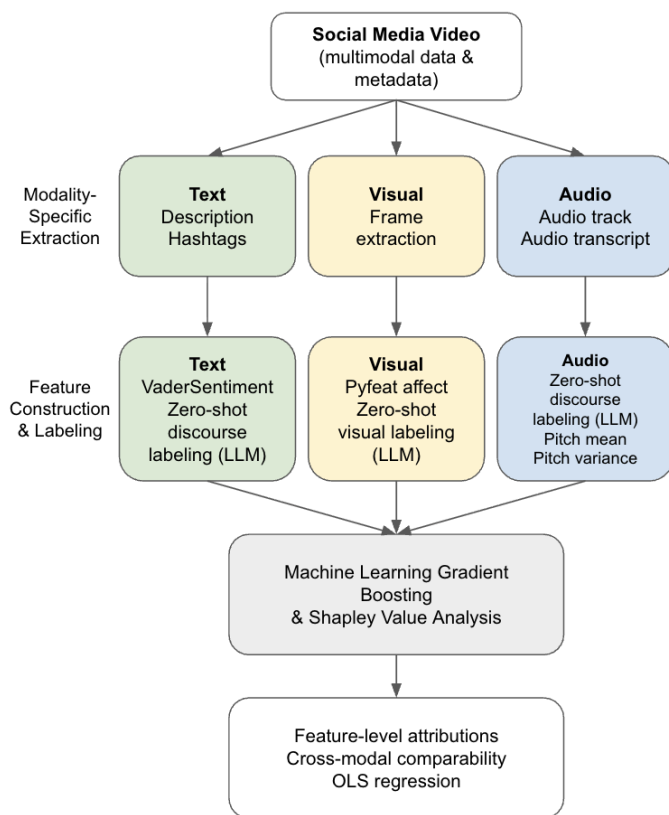


Figure 1: Overview of multimodal analysis pipeline.

Data Collection

We collected TikTok posts via the TikTok Research API. Starting with “social anxiety”, we used snowball sampling to construct a subset of keywords with high co-occurrence and strictly related to social anxiety: “social anxiety”, “socialanxiety”, “social phobia”, “socialphobia”, “social anxiety disorder”, “socialanxietydisorder”. This yielded 162,965 unique posts published in the U.S. between 01/01/2020 to 01/01/2025. The data includes date, username, video id, description, and crucially, engagement metrics such as view count, like count, and comment count. Posts were downloaded using the Python module Pyttok.

Data Augmentation with Language Models

We use zero-shot learning to annotate posts without task-specific labeled data. In zero-shot settings, models assign unseen labels by exploiting shared semantic structure or broad pretraining. Early approaches mapped inputs and labels into a common semantic space to enable transfer (Xian, Schiele, and Akata 2017). In NLP, large pretrained language models demonstrate strong zero-shot capabilities through instruction-following and prompt-based classification—reformulated as cloze or NLI tasks—enabling accurate labeling from natural-language descriptions alone (Radford et al. 2019; Brown et al. 2020). In our pipeline, we prompted GPT-4o with a detailed schema

and decision rules and constrained outputs to a JSON format.

Textual Modality

For every post, the model extracted basic identifiers (languages, hashtags, dates/times, user tags, emojis). For zero-shot labeling, in addition to labeling for typologies of social support (emotional, informational, instrumental, and appraisal), we created a custom macro-set that included the following categories: Humor/Satire, Interpersonal Relationships, Communication, Situational Stressor, Coping Strategies, Self Growth, Self Disclosure, Healthcare, Politics, Advocacy, and additional categories. We extracted mentions of mental health conditions. The full classification prompt is documented in Appendix A.

To quantify affective tone, we applied VADER (Valence Aware Dictionary and sEntiment Reasoner), a commonly-used, rule-based sentiment analysis tool optimized for short, informal text (Elbagir and Yang 2019). VADER produces a sentiment score from -1 to 1.

Audio Modality

Audio tracks were extracted from video files using pydub’s AudioSegment class and converted to uncompressed WAV format. Pydub is a widely-used, open-source Python library for audio manipulation (PyDub Developers 2025). These files served as inputs for audio feature extraction and analysis. We employed OpenAI Whisper (Radford et al. 2023) to generate transcripts from each audio track. Transcripts were analyzed with the same sentiment analysis (VADER) and classification prompt (Appendix A) as the textual modality.

Visual Modality

For each video, we extracted five frames evenly spaced across the video duration. This sampling strategy was chosen to balance representational coverage of visual content with computational efficiency and consistency across videos of slightly varying lengths. Evenly spaced sampling ensures that frames capture visual information from the beginning, middle, and end of each video. Compared to random sampling, this approach provides standardized coverage and reproducibility (Kandhare and Gisselbrecht 2024). Compared to dense frame extraction, it reduces computational cost (Kennedy et al. 2025). Our analysis targets stable, high-level visual attributes such as video format, production style, and facial expression aggregates, rather than temporal gestures or micro-expressions. Brkic et al. (2025) demonstrate that for high-level visual classification tasks, a small number of representative frames achieves performance comparable to dense sampling.

To systematically code higher-level visual characteristics, each frame was passed to GPT-4o-Mini using a structured visual prompt requesting probability scores for visual attributes. This includes visual format, content, and style categories: Selfie, Professional, Homemade, Meme, Full Shot, Close-Up, Wide, Split, Point-of-View, Real, AI, Use of Special Effects, Use of Text, and additional categories. The complete classification prompt is provided in Appendix B.

Frames containing faces were analyzed using open-sourced PyFeat (Cheong et al. 2023), which extracts Action Units (facial muscle movements) and classifies emotions into seven categories: happiness, sadness, anger, fear, surprise, disgust, and neutral. Each emotion was quantified with intensity values ranging from 0 to 1.

Zero Shot Validation

To assess the accuracy of the zero-shot classifications, two trained annotators independently labeled a random sample of 200 units per modality (200 descriptions, 200 transcripts, 200 frames). Validation results (Appendix C) demonstrate overall accuracy rates of 90% for descriptions, 86.5% for transcripts, and 83.5% for images, with inter-rater reliability assessed via Cohen’s κ yielding substantial agreement across modalities (descriptions: $\kappa = 0.74$; transcripts: $\kappa = 0.71$; images: $\kappa = 0.65$).

Predictive Modeling and Shapley Value Analysis

Modeling Framework To evaluate how multimodal features shape attention dynamics, we employed gradient boosting regression models using CatBoost (Prokhorenkova et al. 2018). Gradient boosting methods accommodate heterogeneous feature types, capture nonlinear relationships and higher-order interactions, and achieve strong predictive performance (Chen 2016). We used logged view count as the dependent variable, as it represents initial attention capture rather than post-viewing behaviors (likes, comments, shares). All input features were normalized to a 0-100 scale to ensure standardization across variables, where 100 represents full confidence of the feature being present.

We configured the CatBoost regressor with 1,000 iterations, a learning rate of 0.1, and a maximum tree depth of 6. We trained models on an 80-20 train-test split to evaluate predictive performance.

SHAP Value Analysis To interpret predictions from our gradient boosting models, we use SHAP (SHapley Additive exPlanations), which attributes each prediction to input features using Shapley values from cooperative game theory (Lundberg and Lee 2017). For a fitted model $f(\mathbf{x})$ and a background (reference) distribution over inputs \mathbf{X} , SHAP constructs an additive explanation model

$$f(\mathbf{x}) = \mathbb{E}[f(\mathbf{X})] + \sum_{i=1}^p \phi_i(\mathbf{x}), \quad (1)$$

where $\mathbb{E}[f(\mathbf{X})]$ is the baseline prediction and $\phi_i(\mathbf{x})$ is feature i ’s contribution (in the units of the model output) for observation \mathbf{x} . Shapley values are defined as the average marginal contribution of a feature across all subsets (coalitions) of the remaining features:

$$\phi_i(\mathbf{x}) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (2)$$

where $N = \{1, \dots, p\}$ are index features and $v(S)$ denote the model value when only features in subset $S \subseteq N$ are present.

SHAP can be further decomposed into main effects and pairwise interaction effects.

For each observation \mathbf{x} , SHAP interaction values allocate the deviation from the baseline prediction into (i) a main-effect contribution for each feature and (ii) additional terms capturing how pairs of features jointly contribute beyond their individual effects:

$$f(\mathbf{x}) = \mathbb{E}[f(\mathbf{X})] + \sum_{i=1}^p \phi_{ii}(\mathbf{x}) + \sum_{i < j} \phi_{ij}(\mathbf{x}). \quad (3)$$

where, $\phi_{ii}(\mathbf{x})$ denotes the main-effect attribution for feature i , and $\phi_{ij}(\mathbf{x})$ for $i \neq j$ denotes the interaction attribution between features i and j . In additive models without interactions between i and j , $\phi_{ij}(\mathbf{x})$ is approximately zero across observations.

Linearization of Estimator Surfaces

While SHAP beeswarm plots reveal heterogeneity across individual predictions, social scientists typically require aggregate measures of feature importance analogous to regression coefficients. We address this by computing feature-weighted SHAP contributions. We first standardize the underlying features. The centering operation is domain-specific (i.e. mean or media); we elect 0.5 as our annotations represent probabilities.

Feature-weighted SHAP contributions can therefore be expressed as:

$$\beta_{i,SHAP} = \sum_{k=0}^N x_k \phi_k \quad (4)$$

This aggregation weights each observation’s SHAP value by its feature magnitude, providing a summary measure of how strongly feature i influences predictions when present. Unlike raw SHAP means, this approach accounts for feature intensity: a feature with consistently high SHAP values only when strongly present (high x_k) receives appropriate weight.

Piecewise OLS on SHAP

To characterize threshold-dependent interactions between features from different modalities, we partitioned the feature space at median values (50 for probabilistic features) and estimated separate linear relationships within each quadrant. In the continuous domain, this corresponds to the SHAP value z with respect to variables x and y : $z'' = \frac{\partial^2 z}{\partial x \partial y}$.

Suppose there exists a median threshold x_0 such that the effect of y diverges once x exceeds this threshold:

$$\begin{cases} z'' > 0, & \text{if } x > x_0 \quad \& \quad y > y_0, \\ z'' < 0, & \text{if } x > x_0 \quad \& \quad y < y_0, \\ z'' = 0, & \text{if } x < x_0. \end{cases}$$

To characterize the analogous case in which the interaction is rotated along the y -axis, the conditions on x are inverted (i.e., $z'' = 0$ when $x > x_0$). The third scenario arises when the direction of interaction depends on the quadrant defined by the thresholds of both variables, yielding a sign change across quadrants:

$$\begin{cases} z'' > 0, & \text{if } xy > 0, \\ z'' < 0, & \text{if } xy < 0. \end{cases}$$

This approach reveals whether synergies emerge only when both features exceed thresholds (asymmetric interactions) or operate across all quadrants (symmetric interactions). Because SHAP values are centered at zero by default, the correlation coefficient directly estimates the linear effect of the interaction within each quadrant.

Within-Modality and Cross-Modal Analysis

Using CatBoost, we computed SHAP values separately within each modality to quantify how individual textual, visual, and audio features contributed to predicted view counts. This within-modality approach allowed us to identify the most influential features while preserving the internal structure of each modality. To enable cross-modal comparisons and assess interactions between features from different modalities, we employed XGBoost (Chen 2016), training models on the same 80-20 train-test split. The model was trained for up to 10,000 boosting rounds with early stopping based on test set performance, evaluated every 1,000 rounds to monitor convergence. Feature-level SHAP values were then analyzed to examine interactions between features from different modalities.

Results

Sentiment Across Modalities

First, we consider how sentiment features across modalities differentially predict exposure outcomes using SHAP values. Figure 2 presents a beeswarm plot of SHAP value distributions for sentiment features extracted from textual (caption), visual (facial expression), and audio (speech transcript) modalities. The color gradient shows feature magnitude (high values in red/pink, low in blue), while the horizontal position indicates impact on viewership. Facial expressions of happiness and neutral affect appear to predict higher viewership, with positive SHAP values concentrated right of zero, whereas caption and transcript sentiment measures show noisy, scattered patterns across both positive and negative ranges.

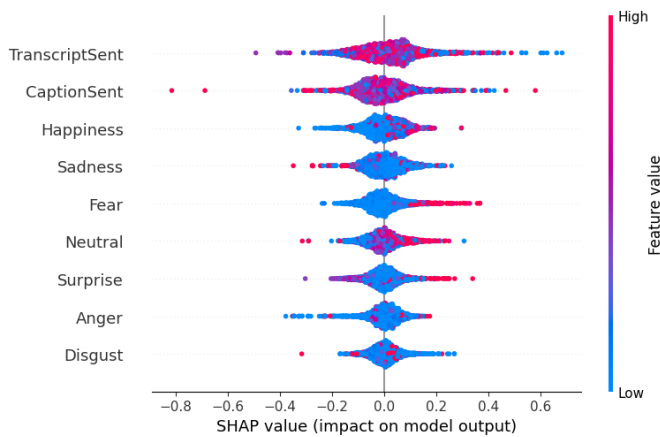


Figure 2: SHAP beeswarm plot of features from the visual, text, and audio modalities.

While the beeswarm plot is valuable for exploring distributional patterns, it lacks interpretability. For instance, although both transcript and caption sentiment have high feature importance, such importance is not necessarily clear. Utilizing feature-weighted SHAP (Eq. 4), direct interpretation of direction and magnitude becomes possible.

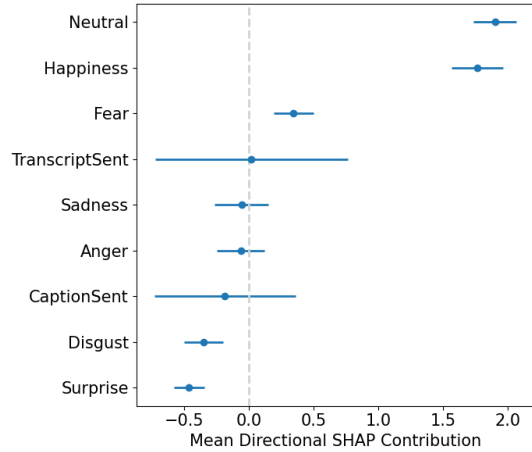


Figure 3: Point-weighted aggregate SHAP values as a forest plot.

Figure 3 reveals that facial expressions of happiness and neutral affect have clear positive SHAP contributions compared to textual and speech sentiment, whose SHAP values span a large range. This suggests that in multimodal content, creator facial expressions produce clearer influence on virality than textual and caption sentiment (Slepian and Carr 2019). This answers **RQ1**.

Feature Contributions Across Modalities

Building on the same analytical framework, we investigated which modality-specific features serve as the strongest predictors of viewership across textual, visual, and audio modalities. SHAP values were computed and weighted by feature importance for all three feature sets. Figure 4 displays the ten most influential features per modality, comprising the five features most positively and five most negatively associated with viewership. Predictive performance differed substantially across modalities: visual features demonstrated the highest explanatory power (Pearson $r = 0.289$), followed by caption features ($r = 0.170$) and audio features ($r = 0.100$). Full beeswarm plots and per-modality regression outputs are reported in Appendices D and E, respectively.

In Figure 4, analysis of textual and audio features both revealed that posts related to “informational support” demonstrated the most positive SHAP values, while posts labeled as “emotional support” yielded the lowest SHAP values. The preference for informational over emotional support content indicates that audiences prioritize actionable information over empathetic expression. This behavior aligns with emerging research on digital health-seeking and suggests opportunities for future investigation into how AI-driven content moderation and recommendation systems might be

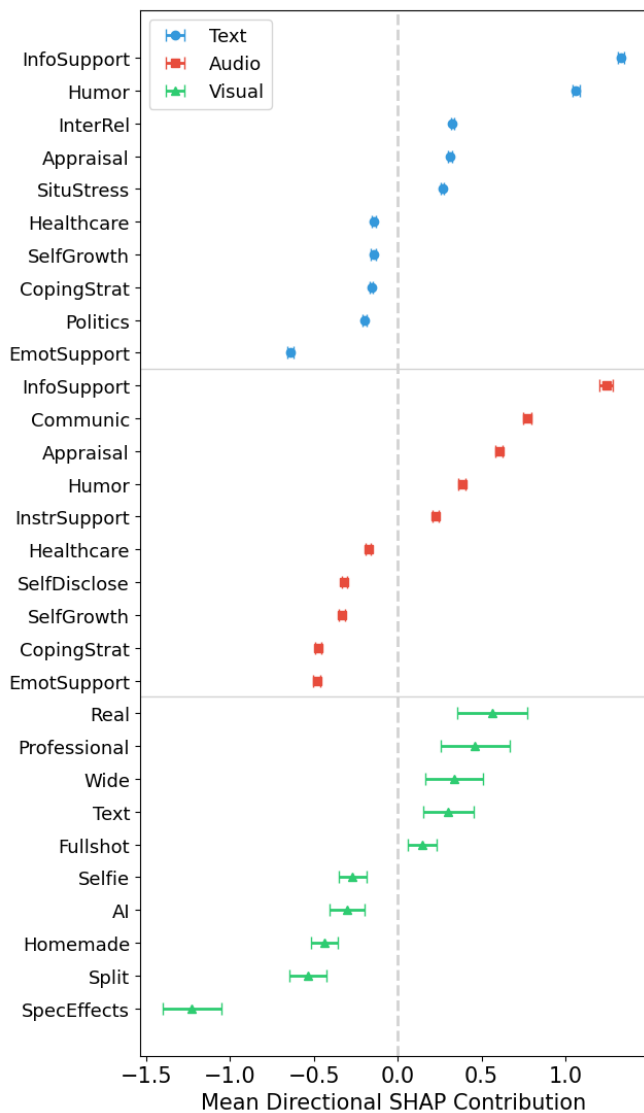


Figure 4: Mean directional SHAP contributions for textual (blue), audio (red), and visual (green) features.

audited or designed to support mental health information access while maintaining safety and accuracy standards. Results in the visual analysis corroborate these findings.

Videos appearing “real” (authentic footage) showed the strongest positive effect (Mean SHAP ≈ 0.5), followed by professional-appearing production quality, wide shots, and visible text overlays (Mean SHAP ≈ 0.3 – 0.4). Conversely, special effects showed the most dramatically negative contribution (Mean SHAP ≈ -1.3). Split-screen formats, homemade appearance, AI-generated content, and selfie framing all showed modest negative effects. Not only do audiences seem to favor informative content, but they also have preferences for high perceived authenticity and visual production polish in mental health content. This answers **RQ2**.

Cross-Modal Interactions

Through modeling with XGBoost, we find that the interaction effects between visual and textual modalities reveal that differing levels of feature interactions drive engagement synergistically while others reduce engagement.

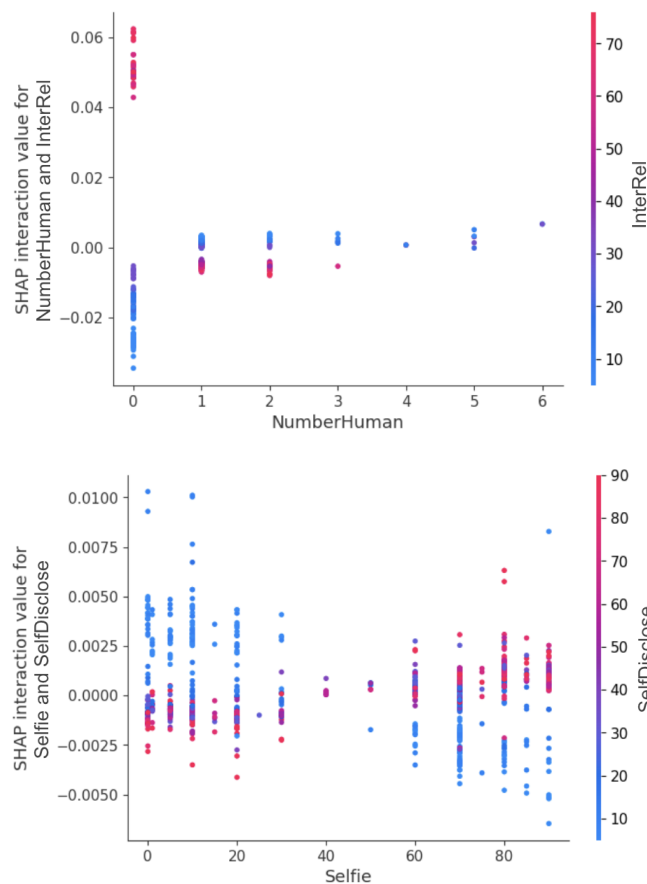


Figure 5: SHAP interaction plots illustrating cross-modal feature effects on viewership. Y-axis shows SHAP Values. X-axis shows one feature’s magnitude. Blue represents low feature magnitude and red represents high magnitude. a) SHAP interaction between number of people in frame and presence of interpersonal relationship narratives in caption. b) SHAP interaction between Selfie visual format and Self Disclosure narrative in caption.

Figure 5a illustrates the interaction between number of people present in image and textual probability of interpersonal relationship. Here, divergence occurs on the left side of the graph. When videos contain no visible humans, discussions of interpersonal relationships increase viewership, as shown by positive SHAP interaction values. However, when one or more people appear in the visual frame, interpersonal relationship themes have no discernible effect on engagement. This suggests that non-literal representations of social connection may enhance viewership.

Figure 5b illustrates how selfie framing and self-

disclosure interact to shape viewership through a threshold-dependent relationship. When selfie probability is low (left side of the graph), high self-disclosure (red/pink points) is associated with negative SHAP interaction values, indicating reduced viewership. Conversely, when selfie probability exceeds approximately 60 (right side of the graph), high self-disclosure (red/pink points) shifts to positive SHAP interaction values, indicating increased viewership. This pattern reveals that self-disclosure narratives require visual self-presentation to drive viewership effectively. Personal storytelling demands visual self-exposure to signal authenticity and build virtual connection.

Based on the scatter plot, it is difficult to precisely interpret the SHAP interaction plots. To mathematically model cross-modal interactions, we used piecewise regression analysis on SHAP interaction values across four quadrants, using feature medians as thresholds. Figure 6 presents regression coefficients (Beta) for the interaction pairs.

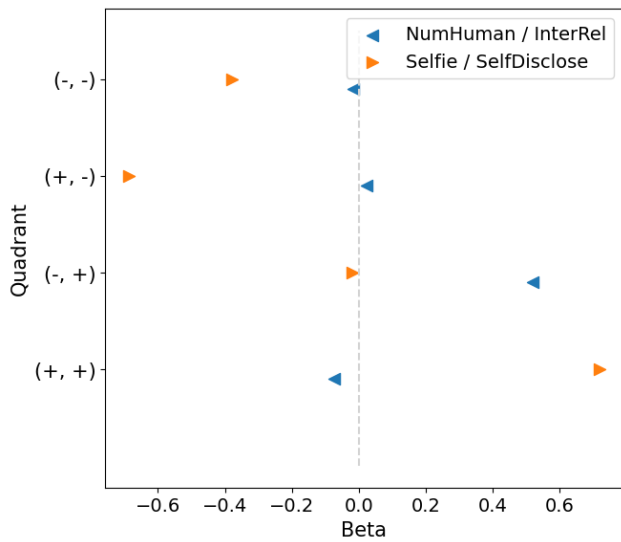


Figure 6: Piecewise regression coefficients from SHAP interaction values across quadrants. Each interaction pair is represented by a distinct marker. Four quadrants represent different combinations of feature values: (-, -) indicates both features below median threshold; (+, +) indicates both above threshold; (-, +) and (+, -) represent mixed conditions where one feature is above and one below.

As noted previously, positive effects emerge when humans are absent but discourse about interpersonal relationships is present (-, + quadrant: $\beta \approx 0.3$). All other quadrants have nearly no effect on viewership. This asymmetry suggests audiences may engage more readily with abstract relationship representations—text screenshots, empty spaces, metaphorical imagery—allowing viewers to project their own experiences onto ambiguous visual content. The selfie / self-disclosure interaction (green \times) demonstrates a pronounced positive and negative dependency: strongly negative when both features are absent (-, - quadrant: $\beta \approx -0.6$) and negative when self-disclosure occurs without visual self-

presentation (-, + quadrant: $\beta \approx -0.3$), but robustly positive when both are present (+, + quadrant: $\beta \approx 0.7$).

Piecewise regression reveals interaction structures that would be obscured by linear modeling, providing more nuanced understanding of how multimodal features shape attention dynamics. This answers **RQ3**.

Unimodal Measures are Inadequate in Multimodal Contexts

Lastly, we construct an example where unimodal measurements lack nuance. An assumption of existing digital mental health research is that coarse features—typically measured through textual lexicon—are effective proxies for making inferences on attention dynamics. Our findings challenge this assumption by showing that affective measures can diverge substantially across modalities within the same content.

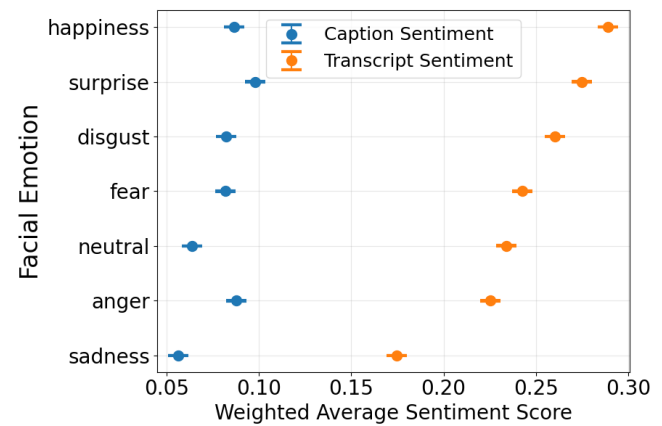


Figure 7: Comparison of Facial Emotion Scores with 95% Confidence Intervals against Caption and Transcript Text Sentiment

Figure 7 shows how facial emotions (as measured by Pyfeat) compare with textual sentiments from post caption and transcript sentiments. Since facial emotions are probabilities, points represent the average of emotions multiplied by the sentiment (Wei, Noh, and Chang 2025).

Figure 7 reveals inconsistent trends in the alignment between text sentiment, speech sentiment, and visual affect. Overall, captions are markedly flatter and more muted, with weighted average scores clustered in a narrow low range (≈ 0.05 – 0.10) across all emotion categories. In contrast, transcripts exhibit substantially higher and more differentiated sentiment weights (≈ 0.17 – 0.29), indicating stronger emotional expression in spoken content. Consequently, caption-based sentiment alone may systematically underestimate both emotional strength and emotional diversity relative to transcript-based analysis. The only consistent pattern observed across caption- and transcript-based sentiment is that sadness corresponds to the lowest sentiment in both written and spoken modalities.

Interestingly, recent work in political communication showed that the facial emotion of anger is correlated to the most negative textual sentiment (Wei, Noh, and Chang

2025). In mental health, sadness seemed to be most correlated with a negative sentiment. As Leo Tolstoy said, “Happy families are all alike; every unhappy family is unhappy in its own way.” The manifestation of negativity may be domain specific, with anger in politics sadness in mental health.

Discussion and Conclusion

The rise of short-form video content has introduced a new social media format that is becoming central to mental health sharing and information-seeking. Our analysis of over 162,000 TikTok posts reveals that the dynamics of mental health communication on this platform should not be reduced to single-modal analyses.

In affective analysis, the primacy of facial expressions over textual sentiment suggests visual self-presentation signals authenticity more effectively than text. Happiness and neutral expressions predicted increase in engagement, potentially because they signal approachability and emotional safety to audiences (Horstmann 2003). Additionally, individuals with heightened social anxiety exhibit greater sensitivity to negative facial expressions and tend to avoid faces that signals distress (Heuer, Rinck, and Becker 2007). The sensitivity to visual content poses a paradox at the heart of mental health discourse: users who need peer support may be systematically disadvantaged by engagement dynamics that reward emotional composure. The result is a visibility gap for users who receive support on short-form video content. Visual warmth may be more critical in engaging audiences than text in multimodal contexts, but could also disadvantage individuals seeking help.

The preference for informational over emotional support suggests that TikTok functions less as a digital support group and more as an informal health information system. In both textual and audio content, posts labeled as providing informational support predicted highest increase in viewership, while posts offering emotional support predicted greatest decrease. This suggests users actively seek actionable guidance rather than affective connection, treating short-form video platforms as educational resources for mental health literacy. However, unlike formal health information systems, TikTok operates with little clinical oversight and ethical safeguards. More than half of the most viewed mental health TikToks were characterized by psychiatrists as being misleading and misrepresenting mental health issues (Yeung, Ng, and Abi-Jaoude 2022). When peer-generated informational content consistently outperforms emotionally supportive content in algorithmic reach, platforms may be inadvertently amplifying health guidance from unqualified sources while suppressing the relational and empathic exchanges that have traditionally defined the value of online mental health communities (Marshall et al. 2024).

In visual content, visual authenticity and production quality are the factors driving the most viewership. This reinforces previous concerns. If audiences reward content that appears visually polished and credible, creators face incentives to present mental health information with confidence and authority, regardless of their actual expertise. Given this, platforms should consider how recommendation algorithms can amplify professionally vetted resources and suppress

health misinformation or unqualified advice. Auditing recommendation systems for the types of mental health content they surface, and designing ranking mechanisms that account for source credibility and information quality may be potential directions for platform-level intervention.

Methodological Contributions

This study addresses the challenge of systematically extracting, integrating, and interpreting heterogeneous features across modalities at scale by introducing a methodological framework that combines automated probabilistic feature extraction with Shapley value-based attribution, enabling researchers to quantify both individual feature contributions and cross-modal interactions in a unified framework.

Our approach makes three methodological advances. First, zero-shot probabilistic classification enables theory-driven feature extraction without fine-tuning, representing discourse categories and visual attributes as continuous probability distributions (0-100) rather than discrete labels. This preserves measurement uncertainty, a theoretical central value, while producing features directly comparable across modalities. Second, we extend SHAP interpretability through feature-weighted aggregation (β_{SHAP}), consolidating point-level attributions into regression-style coefficients that social scientists can interpret as clarity in direction of impact, while respecting the nonlinear complexity of gradient-boosting trees. Third, piecewise regression on SHAP interaction values reveals threshold-dependent cross-modal synergies by partitioning feature space at median thresholds and estimating separate effects within each quadrant. This formalizes when modalities must coordinate versus when they operate independently.

Applying this framework to 162,965 TikTok videos about social anxiety disorder demonstrates its empirical value. We show that sentiment diverges systematically across modalities, with facial expressions predicting engagement while textual sentiment produces noisy signals—evidence that single-modality proxies introduce systematic bias. Cross-modal interactions exhibit threshold effects. These threshold-dependent patterns would be impossible to detect through manual coding at scale or through learned embeddings without interpretable attribution. The pipeline used in this study can be generalized beyond mental health communication with domain-appropriate classification schemes while maintaining the same analytical pipeline.

Limitations and Future Work

Our study faces a few limitations that point to future refinements. First, future work can iterate on which features to include. In audio, for instance, we prioritized speech content over musical features, leaving features such as genre of music, pitch variation, and audio cadence unanalyzed. TikTok users’ preference for overlaying music with speech may also introduces noise into audio attribution, as our pipeline does not separate these sources; speech-music separation techniques would enable more precise labeling and analysis. Second, in visual analysis, treating videos as static snapshots can potentially obscure how multimodal composition evolves temporally within a given piece of content. Third,

while TikTok is the most popular short-form video platform, our exclusive focus on it can limit generalizability, as algorithmic recommendation systems, user demographics, and interface affordances differ across platforms. Cross-platform comparative studies would help distinguish universal multimodal principles from platform-specific effects in mental health communication.

In sum, this study demonstrates that in multimodal environments, analyzing any single channel in isolation risks missing mechanisms that drive visibility and engagement, while integrated analysis draws a fuller picture of the communication dynamics, risks, and opportunities. As computational social science continues to grapple with an increasingly rich and multifaceted media landscape, we hope the framework proposed here proves useful for future work in digital mental health and beyond.

References

- Ali, A.; and Molla, D. 2025. A Systematic Literature Review on Multimodal Text Summarization. *ACM Computing Surveys*, 58(3): 1–38.
- Alonso, J.; Liu, Z.; Evans-Lacko, S.; Sadikova, E.; Sampson, N.; Chatterji, S.; Abdulmalik, J.; Aguilar-Gaxiola, S.; Al-Hamzawi, A.; Andrade, L. H.; et al. 2018. Treatment gap for anxiety disorders is global: Results of the World Mental Health Surveys in 21 countries. *Depression and anxiety*, 35(3): 195–208.
- Andalibi, N.; Ozturk, P.; and Forte, A. 2017. Sensitive self-disclosures, responses, and social support on Instagram: The case of #depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 1485–1500.
- Bazarova, N. N.; and Choi, Y. H. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4): 635–657.
- Bickham, C.; Kazemi-Nia, K.; Luceri, L.; Lerman, K.; and Ferrara, E. 2024. Hidden in Plain Sight: Exploring the Intersections of Mental Health, Eating Disorders, and Content Moderation on TikTok. In *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.
- Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33.
- Chen, H.; Bisbee, J.; Tucker, J. A.; and Nagler, J. 2025. Labeling social media posts: does showing coders multimodal content produce better human annotation, and a better machine classifier? *Political Science Research and Methods*, 1–13.
- Chen, T. 2016. XGBoost: A Scalable Tree Boosting System. *Cornell University*.
- Cheong, J. H.; Jolly, E.; Xie, T.; Byrne, S.; Kenney, M.; and Chang, L. J. 2023. Py-feat: Python facial expression analysis toolbox. *Affective Science*, 4(4): 781–796.
- Daft, R. L.; and Lengel, R. H. 1986. Organizational information requirements, media richness and structural design. *Management Science*, 32(5): 554–571.
- Elbagir, S.; and Yang, J. 2019. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In *Proceedings of the international multicongference of engineers and computer scientists*, volume 122, 13–15. International Association of Engineers.
- Fehm, L.; Pelissolo, A.; Furmark, T.; and Wittchen, H.-U. 2005. Size and burden of social phobia in Europe. *European neuropsychopharmacology*, 15(4): 453–462.
- Hall, R.; and Keenan, R. 2025. More than half of top 100 mental health TikToks contain misinformation, study finds. *The Guardian*. Accessed: 2026-03-30.
- Heuer, K.; Rinck, M.; and Becker, E. S. 2007. Avoidance of emotional facial expressions in social anxiety: The approach–avoidance task. *Behaviour research and therapy*, 45(12): 2990–3001.
- Horstmann, G. 2003. What do facial expressions convey: Feeling states, behavioral intentions, or actions? *Emotion*, 3(2): 150.
- Huang, Q.; Lei, S.; and Chen, Z. 2025. Parasocial interaction and problematic use of short-form video applications: unveiling the mediating mechanism. *Frontiers in Psychology*, 16: 1584685.
- Islam, M. M.; Kakouros, S.; Heikkilä, J.; and Oussalah, M. 2025. Towards an Automated Multimodal Approach for Video Summarization: Building a Bridge Between Text, Audio and Facial Cue-Based Summarization. *arXiv preprint arXiv:2506.23714*.
- Kandhare, M.; and Gisselbrecht, T. 2024. An empirical comparison of video frame sampling methods for multi-modal rag retrieval. *arXiv preprint arXiv:2408.03340*.
- Kennedy, C. J.; Gajjar, B.; Chang, H.-C. H.; Unger, J. B.; and Vassey, J. 2025. Demographic trends in e-cigarette social media marketing: perceiving gender presentation and facial age via computer vision. *Nicotine and Tobacco Research*, ntaf057.
- Kessler, R. C.; Berglund, P.; Demler, O.; Jin, R.; Merikangas, K. R.; and Walters, E. E. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, 62(6): 593–602.
- Kress, G. R.; and Van Leeuwen, T. 2001. Multimodal discourse: The modes and media of contemporary communication. (*No Title*).
- Li, S.; Wu, S.; Liu, T.; Zhang, H.; Guo, Q.; and Peng, Z. 2024. Understanding the Features of Text-Image Posts and Their Received Social Support in Online Grief Support Communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 917–929.
- Liu, J.; Su, Y.; Seth, P.; et al. 2025. Can Large Language Models Grasp Concepts in Visual Content? A Case Study on YouTube Shorts about Depression. *arXiv preprint arXiv:2503.05109*.

- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Marshall, P.; Booth, M.; Coole, M.; Fothergill, L.; Glossop, Z.; Haines, J.; Harding, A.; Johnston, R.; Jones, S.; Lodge, C.; et al. 2024. Understanding the impacts of online mental health peer support forums: realist synthesis. *JMIR Mental Health*, 11: e55750.
- Meghawati, M.; Yadav, S.; Mahata, D.; Yin, Y.; Shah, R. R.; and Zimmermann, R. 2018. A multimodal approach to predict social media popularity. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, 190–195. IEEE.
- Milton, A.; Ajmani, L.; DeVito, M. A.; and Chancellor, S. 2023. “I see me here”: Mental health content, community, and algorithmic curation on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Naslund, J. A.; Aschbrenner, K. A.; Marsch, L. A.; and Bartels, S. J. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2): 113–122.
- Niu, S.; Bartolome, A.; Mai, C.; and Ha, N. B. 2021. #StayHome# WithMe: how do YouTubers help with COVID-19 loneliness? In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–15.
- Prizant-Passal, S.; Shechner, T.; and Aderka, I. M. 2016. Social anxiety and internet use—A meta-analysis: What do we know? What are we missing? *Computers in Human Behavior*, 62: 221–229.
- Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; and Gulin, A. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.
- PyDub Developers. 2025. PyDub: Manipulate audio with a simple and easy high-level interface. <https://www.pydub.com/>. Official homepage and documentation for the PyDub Python audio processing library.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8): 9.
- Rathbone, A. L.; and Prescott, J. 2017. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. *Journal of medical Internet research*, 19(8): e295.
- Salih, A. M.; Raisi-Estabragh, Z.; Galazzo, I. B.; Radeva, P.; Petersen, S. E.; Lekadir, K.; and Menegaz, G. 2025. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, 7(1): 2400304.
- Slepian, M. L.; and Carr, E. W. 2019. Facial expressions of authenticity: Emotion variability increases judgments of trustworthiness and leadership. *Cognition*, 183: 82–98.
- Stein, M. B.; and Stein, D. J. 2008. Social anxiety disorder. *The lancet*, 371(9618): 1115–1125.
- Suhr, J. A.; Cutrona, C. E.; Krebs, K. K.; and Jensen, S. L. 2004. The social support behavior code (SSBC). In *Couple observational coding systems*, 307–318. Routledge.
- Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3): 321–326.
- Tricomi, P. P.; Kumar, S.; Conti, M.; and Subrahmanian, V. 2024. Climbing the Influence Tiers on TikTok: A Multimodal Study. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1503–1516.
- Vilares, D.; Alonso, M. A.; and Gómez-Rodríguez, C. 2015. On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *Journal of the Association for Information Science and Technology*, 66(9): 1799–1816.
- Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; and Gao, W. 2023. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4): 447–482.
- Wei, C.; Noh, S.; and Chang, H.-C. H. 2025. Faces Speak Louder Than Words: Emotions Versus Textual Sentiment in the 2024 USA Presidential Election. In *Companion Proceedings of the ACM on Web Conference 2025*, 1390–1393.
- Weinstein, A.; Dorani, D.; Elhadif, R.; Bukovza, Y.; Yarmulnik, A.; and Dannon, P. 2015. Internet addiction is associated with social anxiety in young adults. *Annals of clinical psychiatry*, 27(1): 4–9.
- Winter, E. 2002. The shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054.
- Wu, B.; Cheng, W.-H.; Zhang, Y.; Huang, Q.; Li, J.; and Mei, T. 2017. Sequential prediction of social media popularity with deep temporal context networks. *arXiv preprint arXiv:1712.04443*.
- Wu, P.; Zou, S.; Chen, C.; and Song, Y. 2025. Hotbed of Stigmatization or Source of Support: A Multimodal Analysis of Mental Health-Related Videos on Douyin. *Computers in Human Behavior*, 108716.
- Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-Shot Learning — The Good, the Bad and the Ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4582–4591.
- Xie, X.; Lin, Y.; and Bai, Q. 2025. The recovery function of parasocial relationships for hopelessness on short-form video platforms: a moderated mediation study. *BMC Public Health*, 25(1): 3668.
- Yeung, A.; Ng, E.; and Abi-Jaoude, E. 2022. TikTok and attention-deficit/hyperactivity disorder: a cross-sectional study of social media content quality. *The Canadian Journal of Psychiatry*, 67(12): 899–906.

Appendix

Textual/Transcript Classification Prompt

System prompt

You are a public health researcher analyzing mental health discourse in social media post texts. Your task is to evaluate the following text:

The following formatting notes are extremely important to follow exactly correctly:

Please give probabilities as percentage likelihood (i.e. 0% if very unlikely and 100% if extremely likely)

Corresponding to the questions below, you will need to output a JSON object. Return the structured JSON only, with no additional text, descriptions, or explanations.

Mental health topics

Please provide the probability that the text's content relates to coping strategies?

Please provide the probability that the text's content relates to communication in social interactions?

Please provide the probability that the text's content relates to interpersonal relationships?

Please provide the probability that the text's content relates to self-growth? Please provide the probability that the text's content relates to situational stressors?

Please provide the probability that the text's content contains the use of humor?

Types of Social Support

Please provide the probability that the text's content relates to emotional support (e.g., expressions of empathy, love, trust and caring)?

Please provide the probability that the text's content relates to instrumental support (e.g., tangible aid and service)?

Please provide the probability that the text's content relates to informational support (e.g., advice, suggestions, and information)?

Please provide the probability that the text's content relates to appraisal support (e.g., information that is useful for self-evaluation)?

Based on the content, determine whether the text is seeking information, providing information, both, neither or unclear. Return this as a string: "seeking", "providing", "both", "neither", or "unclear".

Content Themes

Please provide the probability that the text's content relates to political issues.

Please also provide the probability that the text's content relates to healthcare.

Please also provide the probability that the text's content relates to large language models or AI chatbots.

Please also provide the probability that the text's content engages in advocacy, such as promoting a cause or calling for social or political change?

Please also provide the probability that the text's content aims to raise awareness, such as by educating others about a specific issue or condition?

Please also provide the probability that the text's con-

tent includes a call to action, such as encouraging the audience to take a specific step (e.g., donate, reach out, seek help)?

Please also provide the probability that the text's content involves self-disclosure through personal narrative, such as sharing one's own experiences or mental health journey?

If the text mentions any mental health conditions (e.g., depression, anxiety, PTSD), list all mentioned conditions as strings.

If the text mentions other social media profiles (e.g., @username), list all usernames mentioned.

If the text references social media platforms other than TikTok (e.g., Instagram, Twitter), list all platforms referenced.

Visual Classification Prompt

System prompt

You are a public health researcher analyzing mental health discourse in social media post visuals. Your task is to evaluate the following image.

The following formatting notes are extremely important to follow exactly correctly:

If an image includes any text, please extract it all into a string.

Please give probabilities as percentage likelihood (i.e. 0% if very unlikely and 100% if extremely likely)

Corresponding to the questions below, you will need to output a JSON object. Return the structured JSON only, with no additional text, descriptions, or explanations.

Basic Identification

Describe this image in 200 words or less. Return as string.

State whether the frame is most likely a real photo or AI-generated.

Please provide the probability that the frame contains text (label as Text:).

Please provide the probability that the frame uses special effects (label as SpecialEffects:).

Textual content:

If the frame includes any text, please extract it all into a string (label as TextContent:).

If it contains text, what language is it (label as Language:)? If there are multiple languages, please include them as multiple fields. If the frame includes any hashtags, please extract it all into a string (label as HashTags:).

If the frame includes any locations, please extract it all into a string (label as Location:).

If the frame includes any date, please extract it all into a string (label as Date:).

If the frame includes any link or QR code, please extract the link all into a string (label as Link:).

If the frame includes any song names or music, please extract it all into a string (label as Music:).

Visual content

If the image contains an individual, what objects are the person interacting with directly? List all objects as strings. If there are multiple individuals, list all (label as RelevantObjects:).

Do you recognize any individuals or public figures from the frame? If so, who are they? (label as KnownIndividuals:)

Where is the setting of the location (e.g., bedroom, car, park, office, etc.) (label as Location:)?

Please provide the probability that the style of video shot is a close-up (label as CloseUp:).

Please provide the probability that the style of video shot is a full shot (label as FullShot:).

Please provide the probability that the style of video shot is a two shot (label as Two Shot:).

Please provide the probability that the style of video shot is a point-of-view (label as POV:).

Please provide the probability that the style of video shot is a Wide (label as Wide:).

Please provide the probability that the frame is split into segments (label as Split:).

How many segments are the frame split into? (label as Segments:).

How are the segments split? Vertically, horizontally or other? (label as Orientation:).

Please provide the probability that the recording is a Selfie, i.e., very close-up and clearly hand-held (label as Selfie:)

Please provide the probability that the recording is Homemade, i.e., non-professional and clearly not hand-held, such as footage filmed on a tripod (label as Home-made:).

Please provide the probability that the type of recording is professional (label as Professional:).

Please provide the probability that the frame depicts a meme (label as Meme:)?

What cartoon or anime source is this meme from (Family Guys, The Simpsons..etc) (label as MemeSource:)?

Please provide the probability that the frame is a drawing/illustration? (label as Illustration:)

Please provide the probability that the frame shows someone dancing (label as Dance:).

Please provide the probability that the frame shows someone singing (label as Sing:).

Zero-shot Labeling Validation

Table 1: Percent Validity of Caption Labeling by Category

Category	Validity
Appraisal Support	75%
Coping Strategies	90%
Emotional Support	90%
Healthcare	100%
Humor	95%
Informational Support	90%
Interpersonal Relationships	100%
Political Issues	100%
Self-Growth	95%
Situational Stressors	65%
Overall	90%

Table 2: Percent Validity of Transcript Labeling by Category

Category	Validity
Appraisal Support	85%
Communication Probability	75%
Coping Strategies	90%
Emotional Support	70%
Healthcare	90%
Humor	85%
Informational Support	95%
Instrumental Support	85%
Self-Disclosure Narrative	100%
Self-Growth	90%
Overall	86.5%

Table 3: Percent Validity of Image Labeling by Category

Category	Validity
AI	60%
Full Shot	75%
Homemade	100%
Professional	70%
Real	100%
Selfie	75%
Special Effects	75%
Split	100%
Text	85%
Wide	95%
Overall	83.5%

Full Shap Beeswarm Plots

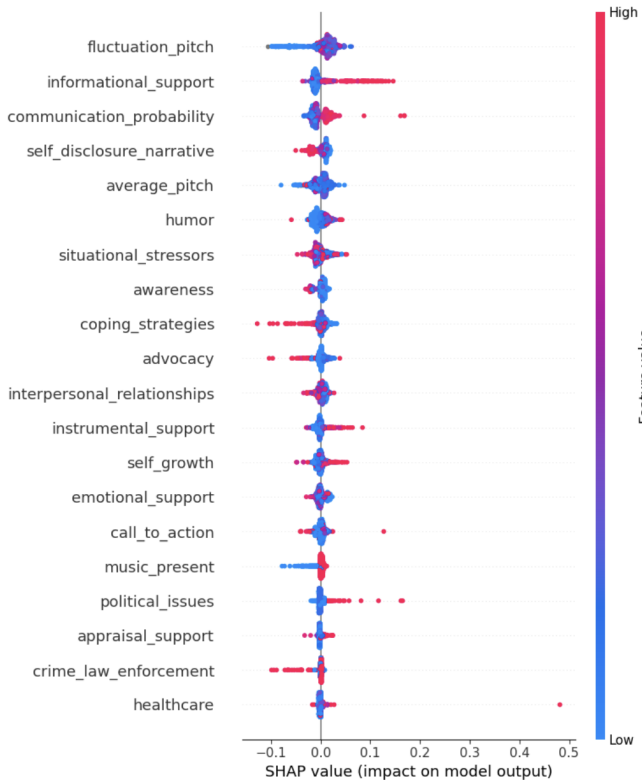
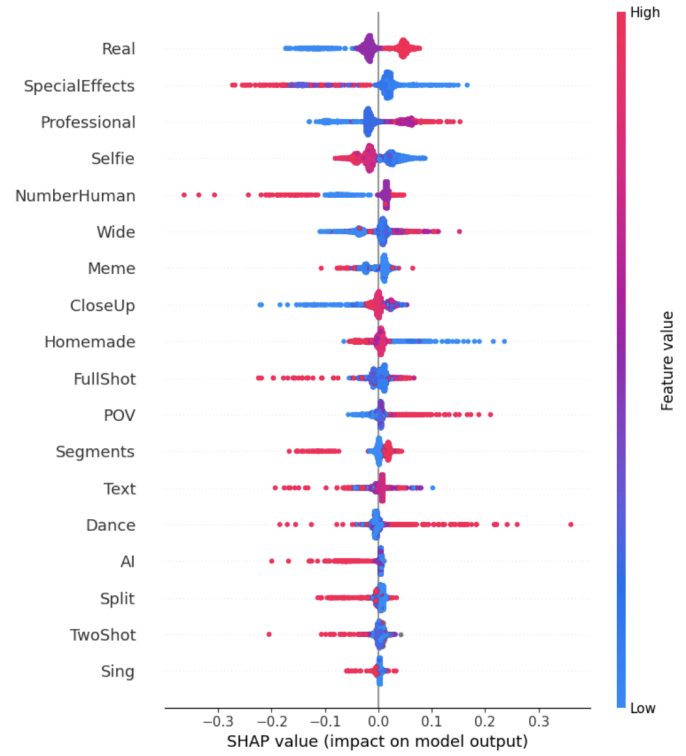
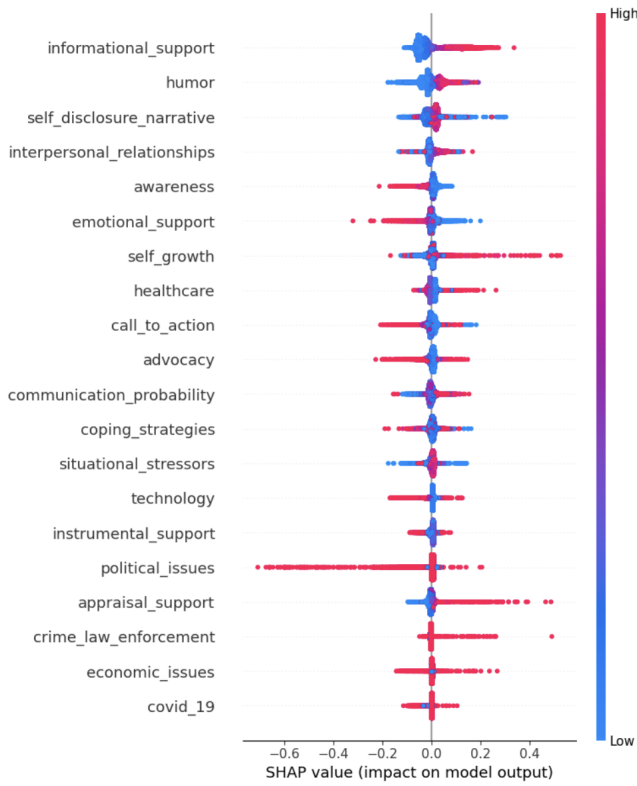


Figure 9: SHAP Summary Plot of textual, visual, and auditory features. Colors represent the feature magnitude, where blue represents low magnitude and red represents high magnitude. a) SHAP value distributions for probabilistic textual features. b) SHAP value distributions for probabilistic auditory features. c) SHAP value distributions for probabilistic visual features.

Regression Plots

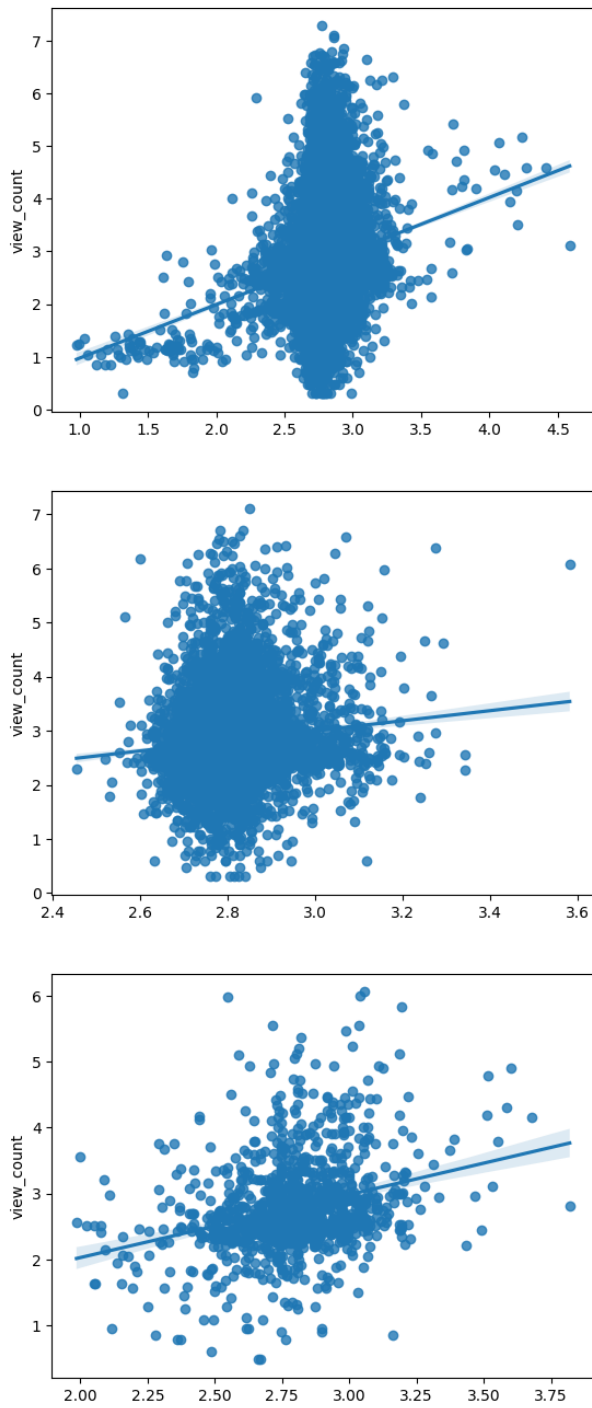


Figure 10: Regression plots for textual, visual, and auditory features. Scatter plots with fitted regression lines showing predicted logged viewership (x-axis) versus actual logged view counts (y-axis) derived from (a) textual, (b) visual, and (c) auditory features.