SCIPEDIA

AI-Driven Multimodal Analysis of User Experience in Immersive Environments: A Case Study of The Sphere

Youjin Seo and Han Young Ryoo*

Department of Content Convergence, Ewha Womans University, Seoul, 03760, Republic of Korea

INFORMATION

Keywords:
AI-driven multimodal analysis
user experience
immersive environments

DOI: 10.23967/j.rimni.2025.10.75987

SCIPEDIA Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Revista Internacional Métodos numéricos para cálculo y diseño en ingeniería RIMNI

In cooperation with CIMNE

AI-Driven Multimodal Analysis of User Experience in Immersive Environments: A Case Study of The Sphere

Youjin Seo and Han Young Ryoo*

Department of Content Convergence, Ewha Womans University, Seoul, 03760, Republic of Korea

ABSTRACT

Artificial intelligence (AI) and sensing technologies are reshaping how people experience immersive environments. This study investigates how audiences perceive and emotionally respond to such environments through an AI-driven mixed-methods analysis. A dataset of 275 user-generated YouTube videos documenting experiences with The Sphere, an AI-convergent immersive environment, totaling over 3000 min of content and 24 million cumulative views, was analyzed to extract experiential themes, dominant emotions, and their relationships with public engagement metrics. The analysis identified seven key experiential themes: Awe of the Display, Personalized Spatial Audio Experience, Full-Body Sensory Engagement, Dynamic Visual Spectacle, Joyful Human–AI Encounter, Futuristic Spatial Design Experience, and Transformative Event Environment. Sentiment analysis revealed that fear was the most dominant emotion in textual narratives (42.3%), followed by surprise, sadness, happiness, and anger, whereas video-based analysis highlighted happiness (25.8%) and sadness (24.5%) as the most salient visual emotions. This contrast suggests that linguistic expressions emphasized feelings of awe and overwhelm, while visual cues reflected affective immersion and emotional depth. Regression results showed that Awe of the Display had the strongest positive impact on engagement (views, likes, comments), while Personalized Spatial Audio Experience showed a negative effect. These findings deepen the understanding of user experience in immersive environments and demonstrate how AI-assisted multimodal analysis can reveal the dynamics between audience perception and engagement in next-generation immersive environments.

OPEN ACCESS

Received: 12/11/2025

Accepted: 15/12/2025

Published: 03/02/2026

DOI
10.23967/j.rimni.2025.10.75987

Keywords:
AI-driven multimodal analysis
user experience
immersive environments

1 Introduction

The convergence of artificial intelligence (AI) and smart sensing technologies has fundamentally transformed how humans experience both digital and physical environments. As intelligent sensing systems and adaptive computation become embedded into architectural and cultural spaces, the environment is no longer a passive backdrop but an active experiential ecosystem capable of interpreting and responding to human activity in real time.

*Correspondence: Han Young Ryoo (hyryoo@ewha.ac.kr).

Human interaction with the world is inherently multimodal [1–3], integrating visual, auditory, tactile, and proprioceptive cues into a unified perceptual experience. This understanding shaped early research on multimodal interfaces in human–computer interaction (HCI) during the 1980s [4]. With advancements in AI and real-time data analytics, multimodal interaction has since expanded beyond screen- or device-based interfaces into AI-convergent immersive environments—spatial settings in which architecture, sensing technologies, display systems, and intelligent computation converge to produce adaptive, multisensory experiences.

The Sphere in Las Vegas represents one of the most advanced realizations of this transition. Featuring an exterior display composed of more than 1.2 million LEDs, a beamforming spatial audio system, and AI-driven content synchronization, the Sphere functions as an AI-convergent immersive environment that engages audiences on perceptual, cognitive, and emotional levels [5,6].

Such AI-convergent multimodal environments introduce a qualitatively new experiential structure that differs from conventional media or interface interactions. In these settings, the audience is no longer a passive recipient of content but an active participant who co-creates the experience through continuous perceptual and affective exchange with intelligent systems. Consequently, sensory immersion extends beyond visual spectacle to encompass multisensory engagement, emotional resonance, and dynamic feedback loops between users and the environment.

These transformations underscore the need for systematic empirical investigation. While prior research in HCI and immersive media has examined multimodal interaction at the device or interface level, far less is understood about how audiences actually perceive, interpret, and emotionally respond within large-scale AI-convergent environments. The experiential patterns that emerge in such settings—and their extension into social participation online—remain underexplored, representing a critical gap in understanding next-generation intelligent experiential systems.

Accordingly, this study conducts a data-driven multimodal analysis to explore audience experiences within the Sphere, a representative AI-Convergent Immersive Environment. Specifically, the study aims to (1) extract key experiential themes reflected in audience narratives, (2) analyze emotions present in textual and visual data, and (3) examine how these experiential and emotional factors influence public engagement metrics such as views, likes, and comments. Through this approach, the study seeks to elucidate how audiences' sensory and emotional responses are formed within AI-Convergent Immersive Environments and how these experiences expand into social participation within online spaces.

Ultimately, this research provides empirical insight into the experiential structure of next-generation intelligent environments by integrating textual narratives, visual emotion signals, and engagement behaviors. The findings contribute to understanding how AI-convergent environments mediate human perception, emotion, and participation, offering practical implications for the design of interactive experience spaces, cultural venues, and multisensory performance environments.

2 Review of Literature

2.1 Evolution of Immersive Experience Environments

Immersive experience environments have emerged as experience-centered spatial media enabled by the advancement of multimedia and display technologies. Early immersive environments primarily relied on visual stimuli to attract users' attention. However, as technology advanced, multiple sensory modalities—such as auditory, tactile, and olfactory—were integrated to create more realistic and engaging experiences. More recently, the convergence of AI and multimodal sensing has enabled

the realization of intelligent immersive environments that can perceive, analyze, and adapt to user responses in real time. The initial exploration of multimodal interfaces can be traced back to the "Put That There" demonstration [4], which combined voice and gesture commands to manipulate graphical objects, demonstrating that cross-modal integration enhances both the naturalness and efficiency of interaction. Building on this foundation, the QuickSet system extended this principle to mobile and applied contexts by synchronizing speech and gesture inputs in both temporal and semantic alignment to achieve automatic multimodal fusion [7,8]. Since the late 1990s, research has expanded beyond visual–auditory interaction to include haptics, thereby deepening the immersive experience. The combination of visual cues and haptic feedback in virtual environments has been shown to refine users' perception of distance, texture, and stiffness [9]. A pseudo-haptic feedback mechanism using miniature vibration motors and tactile displays further demonstrated that touch functions as a core element that enhances emotional and cognitive engagement rather than serving merely as a supplementary sense [9]. It has also been argued that integrating other modalities with visual stimuli amplifies the semantic meaning of interaction, and auditory information has been shown to improve the predictability and consistency of material perception [9]. In sound-centered research, spatial hearing has been emphasized as a crucial element in immersive experience design. Auditory cues in virtual acoustic environments not only convey the direction of a sound source but also shape the perceived depth and atmosphere of space [10]. Such auditory immersion complements visual-centered experiences and reinforces spatial realism. Recent studies have also focused on integrating olfactory modalities into immersive systems. Digital nature experiences incorporating olfactory stimuli positively influence immersion, realism, relaxation, and stress reduction [11]. Other studies found that VR environments augmented with scent promote psychological stability under high-anxiety conditions [11]. Multisensory systems integrating visual, auditory, tactile, and olfactory modalities have empirically confirmed that cross-modal interactions enhance holistic and affective immersion beyond individual sensory experiences [11]. While these studies demonstrate substantial technological progress, most have primarily emphasized system-level integration—focusing on sensors, displays, and feedback mechanisms—rather than the qualitative dimensions of human experience, including emotional, social, and cognitive engagement. To deepen the understanding of immersive environments, particularly AI-convergent environments in which human perception and adaptive intelligent systems co-create experience, future research must move beyond technological implementation and investigate the experiential mechanisms that shape how audiences interpret, feel, and participate within these environments.

2.2 Experience Research in Immersive Experience Environments

Recent technological innovations have fundamentally transformed how people perceive and respond within immersive experience environments. Moving beyond traditional visually centered and observational settings, contemporary immersive experiences increasingly integrate multisensory and affective elements, enabling users to engage with environments in a more embodied and emotionally involved manner. Large-scale media domes, virtual exhibitions, and interactive spatial installations exemplify this shift by combining visual, auditory, and spatial stimuli to create a strong sense of immersion. As immersive experience has gained importance across contemporary culture, art, and spatial design, research interest has likewise expanded beyond technical implementation toward understanding how users construct and experience meaning within immersive environments.

Within this research trajectory, a number of studies have examined user experience in extended-reality (XR) and augmented-reality (AR) contexts. For example, Warsinke et al. [12] revealed that in extended-reality (XR)-based immersive environments, user experience, presence, and emotional

responses significantly affected behavioral intentions. Dağ et al. [13] showed that immersive experience and user engagement in AR-based environments significantly influenced user satisfaction and participatory responses, highlighting the importance of sensory and affective dimensions in immersive interactions. However, few empirical studies have directly examined how experiential factors in immersive environments affect online engagement metrics—such as views, likes, and comments. Therefore, this study explores how sensory and emotional experiences within AI-Convergent Immersive Environments translate into audience engagement, conducting a data-driven analysis using the Sphere as a representative case.

2.3 AI-Driven Multimodal Approaches to Understanding Experience

Understanding user experience within immersive environments has long been a challenge, as such experiences are inherently subjective—shaped by individual perception, emotion, and interpretation. However, the increasing availability of large-scale multimodal data has enabled researchers to analyze how people collectively express, share, and interpret their sensory and emotional experiences in technologically mediated spaces. While earlier studies of immersive experience primarily relied on qualitative interviews or controlled experiments, advances in artificial intelligence (AI) and computational analytics now allow the extraction of emotional, cognitive, and behavioral patterns from vast amounts of user-generated text and visual content.

Building on these developments, recent progress in AI and big data analytics has given rise to AI-driven multimodal approaches for identifying meaningful patterns across large-scale textual, visual, and behavioral datasets. By integrating natural language processing (NLP), computer vision, and statistical modeling, these methods enable researchers to uncover emotional, cognitive, and experiential structures that were previously difficult to observe through traditional methods [14]. From a sociological perspective, individuals continuously perform and present themselves within their social environments—an idea that extends naturally to digital spaces, where users articulate and stage their experiences through multimodal forms of communication. Accordingly, analyzing user-generated content such as YouTube videos, comments, online reviews, and social media posts provides valuable insights into how audiences describe, visualize, and evaluate their experiences within immersive environments.

A growing body of research demonstrates the potential of AI-driven multimodal methods for uncovering emotional and cognitive dimensions of experience. For instance, Ko [15] analyzed YouTube comments to identify recurring affective themes and sentiment distributions, while Porreca et al. [16] applied text mining and emotion analysis to explore shifts in public emotions in response to social issues.

Building upon these emotion analysis studies, recent research has increasingly employed emotion and engagement analyses to examine affective involvement in interactive and immersive systems. Jeong et al. [17] found that emotional engagement significantly predicted user satisfaction in metaverse environments, while Guzman et al. [18] demonstrated that the polarity of emotional tone in collaborative online interactions influenced both motivation and usability. Moreover, integrating text mining with quantitative modeling has enabled researchers to explore how experiential themes relate to behavioral outcomes. Ahmed et al. [19] provided a comprehensive overview of emotion analysis across social networks, emphasizing its value in predicting user behavior, while Gandhi et al. [20] demonstrated how multimodal emotion analysis can link affective expressions with behavioral intentions. These methodological integrations bridge qualitative meaning and quantitative measurement—revealing not

only what users say about their experiences but also how these expressions connect to their intended actions.

Despite these methodological advances, relatively few studies have examined immersive experience environments using AI-driven multimodal techniques. Prior research has often focused on product reviews, health communication, or online fandoms, offering limited understanding of large-scale, multisensory media experiences. To address this gap, the present study integrates text mining, emotion analysis, and visual emotion analytics to investigate how audiences collectively articulate their experiences in AI-convergent immersive environments. Specifically, this study aims to bridge the gap between qualitative audience expressions and quantitative behavioral indicators by analyzing The Sphere as a case study, thereby revealing the relationships between experiential themes and behavioral engagement within large-scale immersive media contexts.

3 Study Design

This study aimed to understand how people experience AI-Convergent Immersive Environments and to explore their online responses to these experiences. As a representative example of this newly emerging type of environment, the Sphere in Las Vegas was selected as the focal case for analyzing user experience. To comprehensively examine how visitors' experiences are expressed and expanded into social engagement, the study employed a two-stage research design as illustrated in Fig. 1.

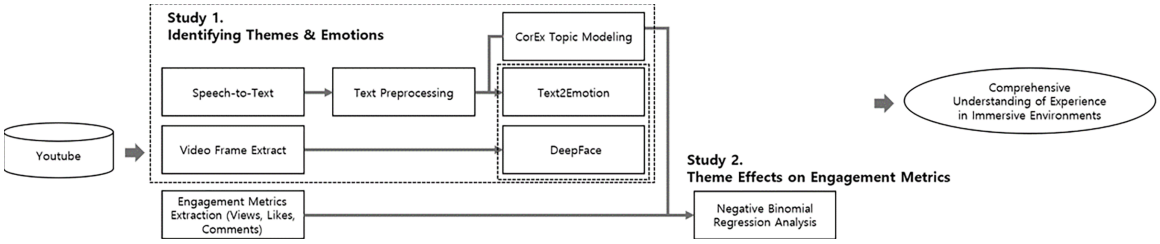

Figure 1: Overall research framework for the AI-driven analysis of user experience

Study 1 focused on identifying the key experiential themes and emotional tones expressed in user-generated content related to the Sphere. Textual and visual data were collected from YouTube videos, including both video transcripts (speech-to-text conversion) and video frames. After text preprocessing, Correlation Explanation (CorEx) topic modeling was applied to extract major experiential themes, while multimodal emotion analysis was performed using Text2Emotion (text-based) and Facial Emotion Recognition (FER) (video-based) to capture affective responses. A qualitative review was conducted to ensure that the computationally derived results accurately reflected the nuances of visitors' experiential narratives. Through this process, Study 1 explored the core themes and emotional characteristics that define user experiences within the Sphere.

Study 2 aimed to determine how these experiential themes influenced audience engagement with the Sphere. Public engagement metrics—including the number of views, likes, and comments—were extracted from YouTube and used as dependent variables. A negative binomial regression analysis was conducted to model the relationships between the experiential factors identified in Study 1 and the engagement indicators. This approach sought to empirically reveal how sensory and emotional experiences within AI-Convergent Immersive Environments translate into varying levels of online engagement and collective audience response.

3.1 Data Collection and Preparation

This study collected data from YouTube, one of the largest platforms for sharing travel-related experiences and impressions. Using the keyword "Sphere Las Vegas experience," search results were retrieved. YouTube's relevance-based ranking algorithm reflects multiple factors, including the match between the search term and video titles, descriptions, and tags, as well as user engagement indicators such as views, watch time, likes, and comments [21–23]. To minimize personalization effects during data collection, a new YouTube account was created with no watch history, no regional preferences applied, and the interface language set to English. All searches were performed while logged in to this new account using YouTube's default relevance-based sorting without additional filters. From the top 300 search results, we conducted a screening process to exclude non-English videos, content posted by news outlets or organizations, and animated material. To ensure the dataset captured genuine visitor experiences, videos were retained only if individuals verbally described or reflected on their own experiences at the Sphere.

Because YouTube's ranking algorithm is dynamic and frequently updated—which may cause search results to vary across users and time [24,25]—all data were collected within a single search session. The top 300 results were recorded, and their URLs were captured as the basis for constructing the dataset. All selected videos were downloaded using the yt-dlp library on 10 August 2024, to ensure consistency in collection timing.

A total of 275 videos met the inclusion criteria. A data cleaning procedure was applied to the extracted transcripts: sentences were converted to lowercase for standardization, and a manual review was conducted to remove extraneous or unrelated content, such as background noise descriptions, promotional remarks, or references to general Las Vegas information or local food.

The final dataset comprised 275 videos and 20,407 sentences. In addition, YouTube video metrics—including the number of views, likes, comments, video length, and upload recency (days since upload)—were collected to perform the subsequent negative binomial regression analysis.

3.2 Dataset Characteristics

Table 1 summarizes the descriptive statistics of the YouTube videos analyzed in this study. The dataset shows substantial variability in both video duration and engagement levels, providing a balanced representation of diverse visitor experiences. Engagement indicators—including the number of views, likes, and comments—were later employed as dependent variables in the regression analysis to examine how experiential themes influenced audience responses.

Table 1: Information of YouTube videos in our dataset

	Mean	S.D.	Median	Min	Max
Duration of videos	11.01	6.93	9.18	4.02	47.82
Views	82,092.19	342,369.23	1392.0	3.0	3,569,082.0
Likes	1965.33	11,556.07	30.0	0.0	168,247.0
Comments	128.23	646.8	6.0	0.0	8900.0

Note: Duration is measured in minutes.

4 Study 1: Identification of Key Themes and Dominant Emotions

4.1 Topic Modeling

To identify the dominant experiential themes expressed in audience narratives, this study employed Correlated Explanation (CorEx) topic modeling, a semi-supervised method that discovers coherent and interpretable latent structures by maximizing total correlation among word clusters [26]. Each YouTube transcript was treated as a single document, and a document–term matrix was constructed using CountVectorizer with the 2000 most frequent terms. Preprocessing included lowercasing, removal of numbers and punctuation, filtering out tokens shorter than three characters, and applying an expanded stopword list that excluded semantically uninformative words (e.g., function words, prepositions, conjunctions, conversational fillers such as oh, uh, yeah, just), while retaining Sphere-related technical terms (e.g., 4k, 8k, hdr, immersive, display).

Multiple theme solutions ranging from 6 to 12 clusters were trained and evaluated to identify the most appropriate thematic structure. Following prior work highlighting c_v as a reliable indicator of human-interpretable topic coherence [27], external c_v coherence was computed using Gensim's coherence model. The seven-theme solution produced the highest semantic coherence (c_v = 0.6389) and demonstrated strong interpretability and stability across repeated runs; therefore, it was selected as the final thematic structure [26].

Model training used an anchor strength of 3 and a random seed of 42. Semi-supervised theme formation was supported through the use of anchor words [26]. Following theme extraction, a binary coding rule was applied such that a theme was coded as present if at least one of its representative keywords appeared in a transcript, allowing each document to contain multiple experiential themes. The final seven experiential themes identified were: Awe of the Display, Personalized Spatial Audio Experience, Full-Body Sensory Engagement, Dynamic Visual Spectacle, Joyful Human–AI Encounter, Futuristic Spatial Design Experience, and Transformative Event Environment.

Table 2 summarizes the frequency and proportion of each theme within the dataset. The most dominant theme was Futuristic Spatial Design Experience, suggesting that the Sphere's architectural structure and spatial design were especially impactful in shaping audience responses.

Table 2: Key themes in textual transcripts of YouTube videos

Themes	Frequency	Percentage (%)
Awe of the display	1664	18.62
Personalized spatial audio experience	1154	12.91
Full-body sensory engagement	2336	26.04
Dynamic visual spectacle	806	9.02
Joyful human–AI encounter	2528	28.28
Futuristic spatial design experience	3385	37.88
Transformative event environment	681	7.62

Note: Themes may co-occur within a single transcript, so percentages reflect overlapping instances.

To further illustrate how each theme was expressed by users, Table 3 presents representative keywords along with summaries of illustrative excerpts that capture how each experiential dimension appeared in the transcripts.

Table 3: Themes, representative keywords, and summaries of illustrative excerpts

#	Themes	Representative keywords	Summary of illustrative excerpts
T01	Awe of the display	Screen, technology, display, resolution, led, square, largest, exterior, interior, wraparound	The transcripts consistently highlight the overwhelming scale, clarity, and brightness of the display, revealing that visual grandeur is a central source of astonishment.
T02	Personalized spatial audio experience	Sound, audio, speakers, beamforming, spatial, acoustic, surround, seat, beam, directly	Mentions of location-specific sound and beamforming effects indicate that listeners perceived the audio system as delivering individualized, precisely targeted auditory experiences.
T03	Full-body sensory engagement	Experience, multisensory, immersive, haptic, vibrations, smell, wind, touch, feel	Descriptions of wind, vibration, temperature shifts, and other bodily sensations suggest that multi-sensory stimulation plays a key role in shaping immersive responses.
T04	Dynamic visual spectacle	Video, visual, content, motion, animation, filmmakers, graphics	Frequent references to shifting colors, animations, and motion effects show that constant visual transformation contributes strongly to the sense of visual excitement.
T05	Joyful human–AI encounter	Interactive, avatar, robots, feedback, aura, people	Interactions with AI avatars—such as conversational exchanges or playful responses—are portrayed as enjoyable moments that evoke curiosity and a sense of novelty.
T06	Futuristic spatial design experience	Structure, design, innovation, architectural, architecture, spherical, unique	Observations of spherical architecture, lighting structures, and futuristic interior design indicate that the physical environment evokes a strong sense of entering a future world.
T07	Transformative event environment	Venue, space, events, seating, concerts, flexible, layout	Accounts of concerts and events describe how the venue visually and acoustically transforms into entirely new settings, emphasizing its capacity to create different experiential worlds.

4.2 Sentiment Analysis

To explore the affective dimensions embedded in visitors' experiential narratives, this study conducted a dual sentiment analysis that integrated both textual and visual modalities. This multimodal approach enables a more comprehensive understanding of emotional responses by capturing both

linguistic expressions and non-verbal cues. Because the YouTube videos used in this study were individually uploaded by users and therefore varied widely in filming conditions and quality—leading to issues such as missing frames, failed face detection, and poor lighting—we adopted an analysis strategy focused on averaging emotion estimates at the video level. This approach allowed us to derive the overall affective tendencies toward the Sphere while minimizing noise and inconsistencies introduced by heterogeneous video quality.

4.2.1 Textual Sentiment Analysis

The textual sentiment analysis was performed using the Text2Emotion algorithm, which detects five primary emotional categories—Happy, Sad, Angry, Fear, and Surprise—from natural language input [28,29]. The algorithm produces continuous probability scores between 0 and 1 for each emotional category by analyzing keywords, linguistic patterns, and contextual cues within the text. Rather than assigning a single dominant emotion to each sentence, all probability values were retained to construct emotion profiles at the video level. For each video transcript, emotion probabilities across all sentences were summed and subsequently normalized to derive the proportional contribution of each emotion to the overall affective composition. Standard deviations were also calculated to assess intra-video variability in emotional expression. This proportion-based aggregation approach provides a more stable representation of affective tendencies by reducing noise inherent in sentence-level fluctuations and capturing the broader emotional tone conveyed throughout the narrative.

Text2Emotion operates on a lexicon-driven classification mechanism, which makes its outputs sensitive to word-level associations and less capable of capturing nuanced, context-dependent, or multi-layered emotional states. Consequently, the resulting sentiment proportions should be considered approximations of affective tendencies rather than precise reflections of complex emotional expression. These aggregated emotion proportions are presented in Table 4 and were subsequently compared with video-based emotion estimates to support a multimodal interpretation of affective responses to the Sphere.

Table 4: Average emotional proportions derived from text and video analyses

	Textual narrative		Videos	
	Mean	SD	Mean	SD
Happy	15.4	7.5	25.8	9.1
Angry	5.3	3.9	17	5.6
Surprise	19.4	6	4.5	3
Sad	17.6	6.7	24.5	7.8
Fear	42.3	11.4	10.2	3.7

Note: Values represent mean percentages (%) and standard deviations (SD) of normalized emotion proportions derived separately from text transcriptions and video-based visual analyses.

4.2.2 Visual Sentiment Analysis

Since textual analysis alone cannot fully capture non-verbal affective expressions, a visual sentiment analysis was conducted to examine emotions conveyed through facial expressions. The DeepFace framework—integrating multiple state-of-the-art back-end architectures such as VGG-Face, FaceNet, OpenFace, DeepID, Dlib, and ArcFace—was employed due to its demonstrated

accuracy and generalizability in facial recognition and emotion detection tasks [30,31]. DeepFace was selected in particular because it has been widely validated in recent facial expression recognition studies, including work improving performance on challenging datasets such as FER-2013 [32]. Moreover, advancements in FER systems developed in complex AI- and XR-based environments [33] indicate that deep learning–based emotion recognition remains robust under less controlled conditions, supporting its applicability to user-generated YouTube videos.

To incorporate the temporal structure of spoken content into the visual analysis, transcript-based timestamped segments were synchronized with the video timeline. Subsequently, frames were sampled at 5-s intervals to capture the video's overall emotional distribution, reflecting the study's aim of characterizing aggregate affective tendencies. Emotion analysis was conducted only when DeepFace successfully detected a face. When multiple faces were present, the largest detected face—presumed to represent the primary speaker or focal subject—was selected, while background faces and non-human detections were excluded. Frames in which no valid face could be detected due to poor lighting, motion blur, occlusion, or extreme head pose were omitted to avoid distorting the aggregated affective estimates. Videos with fewer than three valid frames were excluded because reliable emotion inference was not feasible under such conditions.

For all valid frames, DeepFace's emotion recognition module detected facial landmarks and extracted expression-based features, classifying each frame into one of five core emotions (Happy, Sad, Angry, Fear, Surprise). Emotion proportions were then averaged at the video level, yielding an overall representation of each video's affective tendencies.

Fig. 2 illustrates the overall distribution of emotions across the two modalities. While the text- and video-based sentiment analyses exhibit generally similar structural patterns, the most notable divergence appears in the Fear category. Textual narratives show a substantially higher proportion of Fear, whereas video analysis reveals relatively higher proportions of Happy and Sad. This suggests that the two modalities capture different emotional perspectives of the same experiential context.

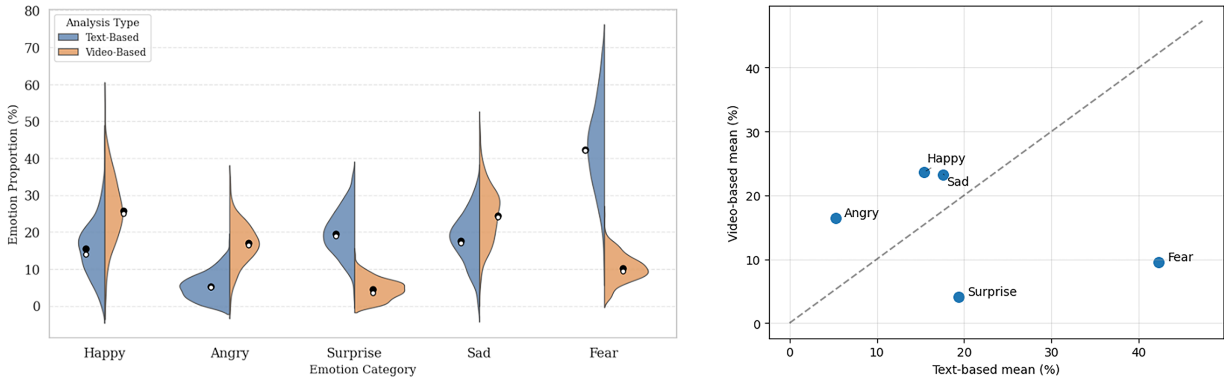

Figure 2: Comparative visualization of cross-modal emotion analyses. (a) Comparative distribution of emotions derived from textual and visual analyses; (b) Cross-modal correlation between mean emotion proportions

To quantitatively assess the degree of alignment between modalities, three cross-modal similarity metrics were calculated: Root Mean Square Error (RMSE = 0.175), Pearson Correlation (–0.419), and Cosine Similarity (0.669). These results indicate an average proportional difference of approximately ±17.5% between emotion distributions, a weak inverse correlation, and a moderate level of structural

similarity. Therefore, text- and video-based sentiment analyses should be utilized complementarily to achieve a more precise and multidimensional interpretation of audience emotions within multimodal affective analysis.

5 Study 2: Understanding How YouTube Content Influences Public Engagement Metrics

To examine how distinct experiential themes identified from YouTube video content relate to public engagement, a negative binomial regression analysis was conducted on three engagement metrics: view counts, likes, and comments. These dependent variables displayed the heavy-tailed distribution typical of user-generated content, in which a small subset of videos accumulates a disproportionately large share of engagement.

Across all three engagement metrics, the variance substantially exceeded the mean, revealing the heavy-tailed and strongly overdispersed nature of user-generated video data. Pearson-based diagnostics further confirmed violation of the equidispersion assumption, thereby justifying the adoption of negative binomial regression instead of a Poisson model [34,35]. To evaluate potential multicollinearity, Variance Inflation Factors (VIFs) were calculated, with all values falling between 1.08 and 1.80—well below the conventional threshold of 5.

To mitigate omitted-variable bias, two structural video characteristics were included as controls: video length and upload recency (days since upload). Video duration had a significant positive association with all engagement metrics, suggesting that longer videos may offer richer experiential content that sustains audience attention. Including upload recency helped adjust for differences in exposure time among videos.

Theme indicators were generated using a sentence-level, keyword-driven classification threshold. A theme was coded as present (1) if any sentence in a transcript contained its representative keywords; otherwise, it was coded as absent (0). The seven experiential themes included Awe of the Display (T01), Personalized Spatial Audio Experience (T02), Full-Body Sensory Engagement (T03), Dynamic Visual Spectacle (T04), Joyful Human–AI Encounter (T05), Futuristic Spatial Design Experience (T06), and Transformative Event Environment (T07).

Table 5 presents the regression coefficients (B), standard errors (SE), z-values, p-values, and 95% confidence intervals (CI). Because the negative binomial model uses a log-link function, each coefficient indicates the expected change in the logarithm of the engagement metric when a given theme is present.

The regression results revealed several noteworthy associations between experiential themes and public engagement metrics. Awe of the Display (T01) exhibited strong positive relationships across all three outcomes. Videos highlighting the Sphere's overwhelming display features generated significantly higher engagement—specifically, approximately 8.0 times more views, 4.7 times more likes, and 4.3 times more comments, as indicated by the positive coefficients for views ($B = 2.079$, $p < 0.001$), likes ($B = 1.551$, $p < 0.001$), and comments ($B = 1.457$, $p < 0.001$).

A similarly robust pattern was observed for Dynamic Visual Spectacle (T04). Videos emphasizing dynamic, motion-rich visual content were associated with substantial increases in engagement, corresponding to approximately 5.0 times more views, 13.6 times more likes, and 8.1 times more comments (views: $B = 1.599$, $p < 0.001$; likes: $B = 2.606$, $p < 0.001$; comments: $B = 2.089$, $p < 0.001$).

Table 5: Associations between themes and viewer engagement metrics: negative binomial regression results

Parameter	No of views					No of likes					No of comments				
	B	SE	z	*p*	95% CI	B	SE	z	*p*	95% CI	B	SE	z	*p*	95% CI
(Intercept)	7.07	0.569	12.419	<0.001	5.954–8.186	2.324	0.576	4.038	<0.001	1.196–3.452	0.983	0.6	1.639	0.101	−0.192–2.158
Video length	0.092	0.009	9.866	<0.001	0.074–0.110	0.136	0.009	14.522	<0.001	0.117–0.154	0.107	0.009	11.415	<0.001	0.089–0.126
Days since upload	0	0	−1.396	0.163	−0.001–0.000	−0.001	0	−3.887	<0.001	−0.002–0.001	−0.001	0	−2.861	0.004	−0.002–−0.000
Awe of the display	2.079	0.272	7.641	<0.001	1.546–2.612	1.551	0.275	5.646	<0.001	1.013–2.089	1.457	0.286	5.086	<0.001	0.896–2.018
Personalized spatial audio experience	−2.476	0.237	−10.431	<0.001	−2.942–2.011	−2.762	0.238	−11.603	<0.001	−3.228–−2.295	−2.394	0.24	−9.959	<0.001	−2.865–−1.923
Full-body sensory engagement	0.534	0.263	2.028	0.043	0.018–1.049	0.514	0.265	1.941	0.052	−0.005–1.032	0.144	0.271	0.532	0.595	−0.387–0.675
Dynamic visual spectacle	1.599	0.281	5.699	<0.001	1.049–2.150	2.606	0.285	9.154	<0.001	2.048– 3.164	2.089	0.301	6.946	<0.001	1.500–2.679
Joyful human–AI encounter	1.667	0.431	3.865	<0.001	0.822–2.513	1.169	0.434	2.693	0.007	0.318– 2.019	1.602	0.456	3.515	<0.001	0.709–2.495
Futuristic spatial design experience	−0.685	0.467	−1.466	0.143	−1.601–0.231	0.058	0.473	0.122	0.903	−0.870–0.985	−0.665	0.487	−1.366	0.172	−1.619–0.289
Transformative event environment	0.357	0.215	1.66	0.097	−0.065–0.778	0.342	0.216	1.584	0.113	−0.081–0.765	0.268	0.22	1.216	0.224	−0.164–0.699

Joyful Human–AI Encounter (T05) demonstrated significant positive relationships across all three engagement metrics. Videos featuring interactive human–AI exchanges were associated with approximately 5.3 times more views, 3.2 times more likes, and 5.0 times more comments, as indicated by the positive coefficients for views ($B = 1.667$, $p < 0.001$), likes ($B = 1.169$, $p = 0.007$), and comments ($B = 1.602$, $p < 0.001$).

Although Transformative Event Environment (T07) produced positive coefficients across all three models, these estimates did not reach statistical significance (views: $B = 0.357$, $p = 0.097$; likes: $B = 0.342$, $p = 0.113$; comments: $B = 0.268$, $p = 0.224$). Exponentiated coefficients suggested modest increases (1.31–1.43×), but these effects should be interpreted cautiously given the lack of significance.

In contrast, Personalized Spatial Audio Experience (T02) was consistently associated with significantly lower engagement. Negative coefficients for views ($B = -2.476$, $p < 0.001$), likes ($B = -2.762$, $p < 0.001$), and comments ($B = -2.394$, $p < 0.001$) indicated sharp decreases in all metrics. Exponentiation showed that videos emphasizing spatial audio characteristics received only 8.4% of the views, 6.3% of the likes, and 9.1% of the comments obtained by videos that did not mention such auditory features. These substantial reductions may reflect the comparatively lower visual immediacy or limited accessibility of audio-centric descriptions for general audiences.

6 Discussion

This study analyzed a unique dataset of 275 user-generated videos documenting individual experiences within AI-Convergent Immersive Environments, specifically focusing on the Sphere. Together, these videos contained 3026.78 min of footage and collectively accumulated 24,213,008 views, 577,788 likes, and 37,520 comments, demonstrating the strong public interest and engagement that AI-driven multisensory environments can evoke in digital spaces.

In Study 1, an integrated text- and vision-based analytical framework was developed to identify the key experiential themes and affective patterns emerging from users' narratives. This multimodal analysis revealed how emotional and perceptual responses are distributed within technologically mediated environments that merge large-scale visual displays, spatialized sound, and interactive AI systems. Through text mining, we identified the most salient experiential themes. The most prominent was the Futuristic Spatial Design Experience, where users described the Sphere's large-scale LED architecture as an entrance into an extraordinary and futuristic domain. The Joyful Human–AI Encounter theme reflected users' fascination with human-like AI agents such as the humanoid robot Aura, which elicited reactions of surprise and amusement. Rather than forming emotional attachment, users expressed admiration for the system's human-like responsiveness, highlighting their cognitive curiosity toward AI's representational naturalness. Additionally, the Full-Body Sensory Engagement theme captured how visitors articulated embodied immersion—describing how light, sound, vibration, and spatial dynamics coalesced into a physically encompassing sensory experience. These findings collectively underscore that the sense of presence within AI-Convergent Immersive Environments arises from multimodal integration that transforms perception into embodied engagement.

The emotion analysis indicates that users' affective expressions can differ substantially across textual and visual modalities. While text-based sentiment detection classified Fear (42.3%) and Surprise (19.4%) as dominant categories, video-based facial expression analysis identified Happiness (25.8%) and Sadness (24.5%) as comparatively prominent. These divergences suggest that each modality captures distinct types of emotional cues. Importantly, the high proportion of "fear" detected in textual narratives should not be interpreted as evidence of genuine fear. Rather, it may reflect a tendency of the classifier to map expressions associated with awe, wonder, or overwhelming sensory

intensity onto the "fear" category. This interpretation remains tentative and should be understood as a classifier-related explanation rather than a definitive psychological conclusion, highlighting how lexical and contextual limitations in text-based models may shape outputs. Accordingly, the textual and visual results are best viewed as complementary yet methodologically distinct signals, each offering only a partial perspective on users' affective responses within immersive environments.

This combined big-data approach demonstrates the methodological potential of computational multimodal analysis in user experience research. By analyzing a large corpus of naturalistic user expressions, the study overcomes the sample-size and ecological validity limitations of traditional laboratory-based presence studies. Moreover, the fusion of text mining with vision-based emotion analysis offers an advanced methodological pathway for quantifying affective engagement in AI-Convergent Immersive Environments.

In Study 2, negative binomial regression analysis was employed to examine how distinct experiential themes relate to public engagement metrics, including views, likes, and comments. The results (Fig. 3) showed that Awe of the Display and Dynamic Visual Spectacle exerted strong and statistically significant positive effects across all engagement indicators. Transformative Event Environment also exhibited positive associations, although these effects did not reach conventional significance thresholds.

These findings suggest that, even within multimodal environments, visually dominant experiences remain the primary drivers of public attention and interaction. Linguistic evidence from the transcripts reinforces this pattern: users frequently described their experiences in terms of light, color, motion, scale, and spatial transformation, reflecting perceptual immersion and sensory astonishment. The visual dimension of AI-Convergent Immersive Environments thus functions simultaneously as an informational and emotional interface, converting large-scale digital displays into embodied experiences that evoke fascination and technological awe.

The Joyful Human–AI Encounter theme showed significant positive associations with views, likes, and comments. However, the effects were more pronounced for views and comments than for likes, suggesting that AI-enabled interactions may primarily draw attention and stimulate conversational engagement rather than eliciting strong affective endorsement. This interpretation remains tentative, as the regression model does not directly capture users' psychological states; nevertheless, it aligns with descriptive observations that users often responded to AI agents with curiosity about their technological novelty.

In contrast, Personalized Spatial Audio Experience demonstrated strong negative associations across all engagement metrics. Although many users described the Sphere's beamforming and spatialized audio system as technically impressive, these auditory effects may be less effectively communicated through visually oriented online media. As a result, audio-focused descriptions—while central to the *in-situ* experience—appear less likely to contribute to online engagement.

Overall, these findings indicate that public engagement with AI-Convergent Immersive Environments is predominantly shaped by visually perceivable and emotionally salient cues, whereas non-visual modalities such as spatial audio may be underrepresented in digital communication. The results underscore the value of multimodal analytical frameworks capable of capturing how technological affordances, sensory integration, and affective perception jointly structure user experience and engagement in AI-driven media environments.

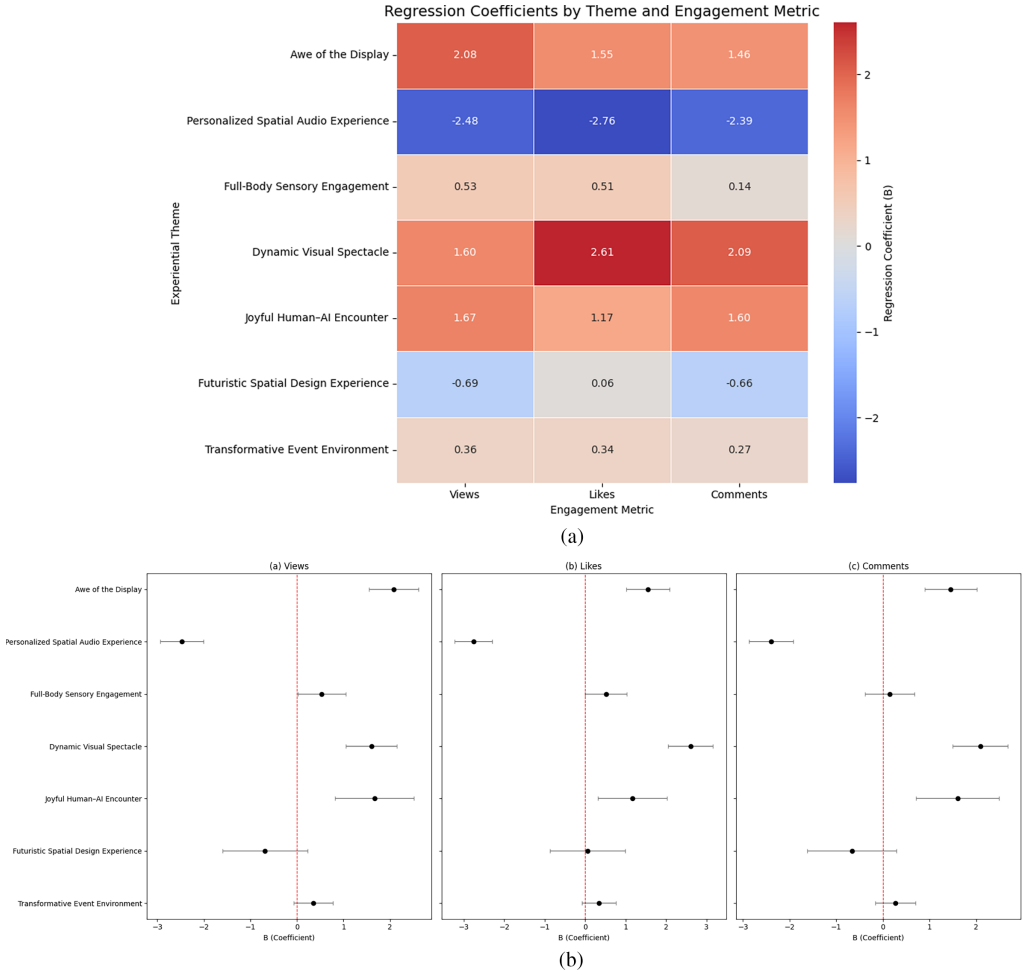

Figure 3: Relationships between experiential themes and engagement metrics. (a) Negative binomial regression coefficients by theme and engagement metric (heatmap); (b) Regression coefficients for Views with 95% confidence intervals (coefficient plot). Note. Panel (a) presents the regression coefficients (B) for each experiential theme across the three engagement metrics (Views, Likes, Comments). Panel (b) visualizes the coefficients for Views with their 95% confidence intervals, highlighting the direction and significance of each theme's effect

7 Conclusion

This study examined the newly emerging experiential environment of the Sphere, expanding the understanding of user experience within AI-Convergent Immersive Environments through the analysis of real visitor-generated content. By integrating text mining and emotion analysis of visual and auditory content, this study extracted the underlying themes and emotional tendencies embedded in the videos. Through this process, it revealed not only the narratives expressed by visitors but also the latent emotions, perceptions, and invisible affective contexts conveyed in the audiovisual materials. Furthermore, by empirically analyzing the relationships between the thematic content of videos and viewer engagement (views, likes, and comments), the study contributed to understanding the mechanisms of emotional response and audience participation within the YouTube platform.

These findings provided empirical evidence of which experiential themes attracted public attention and which core themes stimulated online engagement, offering a practical analytical foundation for the design and development of AI-Convergent Immersive Experience Environments in the future.

The theoretical and empirical contributions of this study are fourfold. First, this work advances experience research by examining AI-Convergent Immersive Environments at a naturalistic scale using large volumes of user-generated content. Unlike traditional survey-based approaches that rely on prestructured items and cognitively framed responses, this study utilizes organically produced YouTube videos, allowing direct observation of how experiences are expressed, interpreted, and circulated in real-world contexts. Second, the study proposes an integrated multimodal analysis pipeline that combines text mining, linguistic emotion detection, visual affect recognition, and engagement modeling. This pipeline demonstrates how heterogeneous data streams can be systematically synthesized to reveal latent experiential patterns that are not observable through single-modality methods. Third, the findings highlight the visual dominance of immersive experiences. Across modalities and analyses, visually salient themes—such as display awe and dynamic visual spectacle—emerged as the strongest predictors of public engagement. This empirical evidence refines theoretical understanding of how sensory hierarchies operate within AI-driven immersive environments. Fourth, by modeling how experiential themes influence views, likes, and comments, the conceptualizes online engagement as a broader participation ecosystem. This approach illustrates how individual perceptual responses can scale into collective attention, interpretive momentum, and social resonance within digital platforms.

Despite its contributions, this study has several limitations. First, because the dataset was constructed using YouTube's relevance-based ranking algorithm, the collected videos may be disproportionately biased toward high-visibility, English-language content. As a result, videos with lower algorithmic visibility or those reflecting region-specific linguistic and cultural variations may be underrepresented, thereby limiting the inclusiveness and representativeness of the dataset. Accordingly, the findings may not fully represent the experiences of the entire population of Sphere visitors, particularly those whose perspectives are shaped by regional, linguistic, or cultural differences that are less visible in algorithmically ranked content. This limitation should be taken into account when interpreting the generalizability of the result. Second, the emotion analysis relied on five basic affective categories, which may not fully capture the complex, multi-layered emotional states typically elicited in large-scale immersive environments—such as awe, sublimity, or overwhelming sensory intensity. Consequently, the emotion outcomes should be interpreted as approximations that reflect only a portion of the broader affective spectrum. Third, visual sentiment analysis was feasible only in segments where a face could be reliably detected, making it difficult to represent the full emotional nuance embedded throughout the entire narrative sequence of each video. Accordingly, the visual emotion estimates in this study should be understood as exploratory, video-level indicators rather than precise frame-aligned mappings. Lastly, the elevated proportion of "fear" in textual sentiment outputs should not be interpreted as conclusive evidence of negative emotional dominance. Rather, it may indicate that the classifier conflated awe-related, high-arousal positive emotions—such as wonder, astonishment, or sensory overwhelm—with the fear category, reflecting known lexical and contextual limitations of existing text-based models. This interpretation is therefore tentative and should be approached with caution.

Future research incorporating qualitative validation, expanded affective taxonomies, and more advanced emotion classification models may help distinguish awe-driven responses from genuinely negative affect with greater precision.

Acknowledgement: The authors would like to express their sincere gratitude to one another for their collaborative efforts throughout this research.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Youjin Seo and Han Young Ryoo; Methodology, Youjin Seo; Software, Youjin Seo; Validation, Youjin Seo and Han Young Ryoo; Formal Analysis, Youjin Seo; Investigation, Youjin Seo; Resources, Youjin Seo; Data Curation, Youjin Seo; Writing—Original Draft Preparation, Youjin Seo; Writing—Review and Editing, Youjin Seo and Han Young Ryoo; Visualization, Youjin Seo; Supervision, Han Young Ryoo; Project Administration, Han Young Ryoo. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, [Han Young Ryoo], upon reasonable request.

Ethics Approval: This study used only publicly available YouTube videos, and no identifiable personal information was collected or analyzed.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Bunt H, Beun RJ, Borghuis T. Multimodal human-computer communication: systems, techniques, and experiments. In: Maybury M, Wahlster W, editors. Lecture notes in computer science. Vol. 1374. Berlin/Heidelberg, Germany: Springer; 1998. doi:10.1007/bfb0052309.
2. Quek F, McNeill D, Bryll R, Duncan S, Ma XF, Kirbas C, et al. Multimodal human discourse. ACM Trans Comput-Hum Interact. 2002;9(3):171–93. doi:10.1145/568513.568514.
3. Turk M. Multimodal interaction: a review. Pattern Recognit Lett. 2014;36:189–95. doi:10.1016/j.patrec.2013.07.003.
4. Bolt RA. Put-that-there: voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques—SIGGRAPH '80; 1980 Jul 14–18; Seattle, WA, USA. p. 262–70. doi:10.1145/800250.807503.
5. Sphere Entertainment Co. Sphere Science & Technology Overview [Internet]. Las Vegas, NV, USA: Sphere Entertainment Co.; 2023–2024 [cited 2025 Dec 1]. Available from: https://www.thesphere.com/science.
6. Johnson A, Luthwaite A, Grow J, Motamedi S, Breen P. Immersion in Sphere: redefining live entertainment experiences [Internet]. Santa Clara, CA, USA: NVIDIA Corporation; 2024 [cited 2025 Dec 1]. Available from: https://www.nvidia.com/ko-kr/on-demand/session/gtc24-s63135/.
7. Oviatt S. Ten myths of multimodal interaction. Commun ACM. 1999;42(11):74–81. doi:10.1145/319382.319398.
8. Cohen PR, Johnston M, McGee D, Oviatt S, Pittman J, Smith I, et al. QuickSet: multimodal interaction for distributed applications. In: Proceedings of the Fifth ACM International Conference on Multimedia—MULTIMEDIA '97; 1997 Nov 9–13; Seattle, WA, USA. p. 31–40. doi:10.1145/266180.266328.
9. Jaimes A, Sebe N. Multimodal human-computer interaction: a survey. Comput Vis Image Underst. 2007;108(1–2):116–34. doi:10.1016/j.cviu.2006.10.019.
10. Begault DR. 3-D sound for virtual reality and multimedia. Cambridge, MA, USA: Academic Press; 1994.

11. Jung S, Karki N, Slutter M, Lindeman RW. On the use of multi-sensory cues in symmetric and asymmetric shared collaborative virtual spaces. Proc ACM Hum-Comput Interact. 2021;5(CSCW1):1–25. doi:10.1145/3449146.
12. Warsinke M, Vona F, Kojić T, Voigt-Antons JN, Möller S. Digital twins for extended reality tourism: user experience evaluation across user groups. In: Extended reality. Cham, Switzerland: Springer Nature; 2025. p. 22–41. doi:10.1007/978-3-031-97769-5_3.
13. Dağ K, Çavuşoğlu S, Durmaz Y. The effect of immersive experience, user engagement and perceived authenticity on place satisfaction in the context of augmented reality. Libr Hi Tech. 2024;42(4):1331–46. doi:10.1108/lht-10-2022-0498.
14. Mustaqim T, Umam K, Muslim MA. Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm. J Phys Conf Ser. 2020;1567(3):032024. doi:10.1088/1742-6596/1567/3/032024.
15. Ko HK. An analysis of YouTube comments on BTS using text mining. Rhizomatic Revolut Rev. 2020;1(1):1–10.
16. Porreca A, Scozzari F, Di Nicola M. Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination. BMC Public Health. 2020;20(1):259. doi:10.1186/s12889-020-8342-4.
17. Jeong D, Kim H, Yoon S. Unraveling the relationship between the dimensions of user experience and user satis-faction in metaverse: a mixed-methods approach. Korean Assoc Inf Syst J. 2023;32(3):19–39. doi:10.5859/KAIS.2023.32.3.19.
18. Guzman E, Azócar D, Li Y. Sentiment analysis of commit comments in GitHub: an empirical study. In: Proceedings of the 11th Working Conference on Mining Software Repositories; 2014 May 31–Jun 1; Hyderabad, India. p. 352–5. doi:10.1145/2597073.2597118.
19. Ahmed K, El Tazi N, Hossny AH. Sentiment analysis over social networks: an overview. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics; 2015 Oct 9–12; Hong Kong, China. p. 2174–9. doi:10.1109/SMC.2015.380.
20. Gandhi A, Adhvaryu K, Khanduja V. Multimodal sentiment analysis: review, application domains and future directions. In: 2021 IEEE Pune Section International Conference (PuneCon); 2021 Dec 16–19; Pune, India. p. 1–5. doi:10.1109/PuneCon52575.2021.9686504.
21. YouTube Help. Search and discovery on YouTube [Internet]. Mountain View, CA, USA: Google Support; 2023 [cited 2025 Dec 1]. Available from: https://support.google.com/youtube/answer/16090438.
22. Altun A, Askin A, Sengul I, Aghazada N, Aydin Y. Evaluation of YouTube videos as sources of information about complex regional pain syndrome. Korean J Pain. 2022;35(3):319–26. doi:10.3344/kjp.2022.35.3.319.
23. Chandrasekaran R, Konaraddi K, Sharma SS, Moustakas E. Text-mining and video analytics of COVID-19 narratives shared by patients on YouTube. J Med Syst. 2024;48(1):21. doi:10.1007/s10916-024-02047-1.
24. Rieder B, Matamoros-Fernández A, Coromina Ò. From ranking algorithms to 'ranking cultures'. Convergence Int J Res New Medium Technol. 2018;24(1):50–68. doi:10.1177/1354856517736982.
25. Arthurs J, Drakopoulou S, Gandini A. Researching YouTube. Convergence Int J Res New Medium Technol. 2018;24(1):3–15. doi:10.1177/1354856517737222.
26. Gallagher RJ, Reing K, Kale D, Ver Steeg G. Anchored correlation explanation: topic modeling with minimal domain knowledge. Trans Assoc Comput Linguist. 2017;5:529–42. doi:10.1162/tacl_a_00078.
27. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; 2015 Feb 1–6; Shanghai, China. p. 399–408. doi:10.1145/2684822.2685324.
28. Abu-Salih B, Alhabashneh M, Zhu D, Awajan A, Alshamaileh Y, Al-Shboul B, et al. Emotion detection of social data: APIs comparative study. Heliyon. 2023;9(5):e15926. doi:10.1016/j.heliyon.2023.e15926.
29. Di Sotto S, Viviani M. Health misinformation detection in the social web: an overview and a data science approach. Int J Environ Res Public Health. 2022;19(4):2173. doi:10.3390/ijerph19042173.

30. Taigman Y, Yang M, Ranzato M, Wolf L. DeepFace: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. p. 1701–8. doi:10.1109/CVPR.2014.220.
31. Serengil SI, Ozpinar A. LightFace: a hybrid deep face recognition framework. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU); 2020 Oct 15–17; Istanbul, Turkey. p. 1–5. doi:10.1109/asyu50717.2020.9259802.
32. Navyamol KT, Jose RT. Enhancing DeepFace algorithm performance for emotion detection: an adaptive vision preprocessing approach using FER-2013 dataset. Signal Image Video Process. 2025;19(14):1202. doi:10.1007/s11760-025-04676-6.
33. Kashef A, Wang Y, Assafi MN, Ma J, Wang J, Jones JA, et al. Developing A novel AI enabled extended reality system for real-time automatic facial expression recognition and system performance evaluation. Adv Eng Inform. 2025;65:103207. doi:10.1016/j.aei.2025.103207.
34. Cameron AC, Trivedi PK. Regression analysis of count data. 2nd ed. Cambridge, UK: Cambridge University Press; 2013.
35. Hilbe JM. Negative binomial regression. 2nd ed. Cambridge, UK: Cambridge University Press; 2011.