International Journal of Multidiscipline

TLEP – International Journal of Multidiscipline
(Technology, Language, Education, and Psychology)
ISSN: 2488-9342 (Print) | 2488-9334 (Online)

Open Access | Peer-Reviewed | Monthly Publication | Impact factor: 8.497 / 2025

Multimodality in Internet Discourse: The Interplay of Textual, Visual, and Paralinguistic Elements

Vohidova Tamanno Saidjonovna
Kokand University, World Language Department
vohidova.t@gmail.com

Abstract

Contemporary internet discourse is characterized by a complex semiotic environment where meaning is co-constructed through multiple modes of communication. This paper investigates the intricate interplay of textual, visual, and paralinguistic elements in shaping meaning on social media platforms. Moving beyond analyses that treat these modes as discrete, this study adopts an integrated social semiotic framework to explore how they combine, compete, and synergize to create coherent and nuanced communicative acts. Using a qualitative content analysis of a curated corpus of 200 posts from the social media platform X (formerly Twitter), this research examines the functional hierarchy and interdependence of different modes in contexts of humor, irony, and emotional expression. The results reveal a clear trend towards multimodal dominance, where visual elements (images, GIFs) and paralinguistic cues (emojis, typography) frequently override or fundamentally alter the literal meaning of accompanying text. Findings indicate that in affective and ironic discourse, the visual mode often serves as the primary carrier of pragmatic meaning, while paralinguistic features act as crucial disambiguating agents. This study contributes to the field of computer-mediated communication by providing empirical evidence that multimodality in internet discourse is not merely additive but transformative, creating composite meanings that are irreducible to their individual components. The implications for digital literacy, sentiment analysis, and communication theory are discussed.

Keywords: multimodality, internet discourse, social media, social semiotics, paralinguistic cues, visual communication, computer-mediated communication (CMC)

Introduction

The digital revolution has fundamentally reshaped the landscape of human communication, transforming it from a predominantly text-centric medium into a vibrant, dynamic, and inherently multimodal ecosystem. In the early days of computer-mediated communication (CMC), discourse was largely constrained by the technological limitations of the time, resulting in text-only environments that scholars characterized as lean and lacking the rich nonverbal cues of face-to-face interaction. However, the proliferation of broadband internet, the rise of visually oriented social media platforms, and the development of sophisticated mobile technologies have rendered this initial characterization obsolete. Today's internet discourse, particularly on platforms such as Instagram, TikTok, and X (formerly Twitter), is a complex tapestry woven from text, static and moving images, audio clips, and a sophisticated arsenal of paralinguistic symbols like emojis, stickers, and typographic alterations. This integration of multiple semiotic resources, or **multimodality**, is no longer an ancillary feature of online communication but its defining characteristic. Understanding how these diverse modes interact is therefore crucial to comprehending the nature of contemporary social interaction and meaning-making.

This paper addresses a critical gap in the existing literature by investigating the functional interplay and hierarchical relationships between textual, visual, and paralinguistic elements in online social discourse. While a significant body of research has examined individual modes in isolation—for instance, the semiotics of internet memes, the pragmatic functions of emojis, or the rhetoric of viral images—fewer studies have adopted a holistic framework to analyze how these modes work in concert within a single communicative act. The central premise of this research is that the meaning of a social media post is not simply the sum of its parts; rather, it emerges from a complex, synergistic process of inter-modal negotiation. A text may be affirmed, contradicted, nuanced, or entirely inverted by an accompanying image, GIF, or string of emojis. This dynamic interplay necessitates an analytical approach that moves beyond text-centric biases and acknowledges the capacity of all modes to carry significant semantic and pragmatic weight.

The theoretical foundation for this study is rooted in social semiotics, particularly the work of Kress and van Leeuwen, who argue that all modes of communication are shaped by social and cultural practices and possess their own "grammars" for making meaning. They posit that different modes have different **affordances**—that is, they are suited for different kinds of communicative work. Text, for instance, excels at conveying propositional logic and narrative sequence, while images are powerful in their capacity for simultaneous representation and affective appeal. Paralinguistic cues, such as emojis, function to inject emotional tone and illocutionary force into otherwise ambiguous text. The central problem this paper seeks to address is how these distinct modal affordances are orchestrated in the constrained and fast-paced environment of social media. This inquiry is guided by the following research questions: 1) How do textual, visual, and paralinguistic elements interact to create cohesive, singular meanings in social media posts? 2) In communicative contexts requiring nuanced interpretation, such as humor and irony, what is the functional hierarchy of these different modes? 3) To what extent do paralinguistic cues, particularly emojis, serve to modify, disambiguate, or reinforce the meaning of accompanying text and images?

By answering these questions, this paper argues that in contemporary internet discourse, visual and paralinguistic modes are not merely supplementary to text but are often the primary carriers of meaning, fundamentally reshaping illocutionary force and pragmatic intent. Through a qualitative analysis of a corpus of social media posts, this study will demonstrate that the relationship between modes is often one of transformation rather than simple reinforcement. This research aims to provide a more nuanced understanding of multimodal communication online, offering insights that are pertinent not only to linguistics and communication studies but also to fields such as digital literacy, human-computer interaction, and artificial intelligence, where the accurate interpretation of human sentiment and intent is paramount. The following sections will detail the theoretical background, outline the methodology employed, present a detailed analysis of the findings, and discuss their broader implications.

Literature Review

The study of multimodality in digital spaces builds upon a rich history of research in social semiotics and computer-mediated communication (CMC). The foundational work in this area is unequivocally that of Gunther Kress and Theo van Leeuwen. In seminal texts such as *Reading Images: The Grammar of Visual Design* (2006), they extended the principles of systemic functional linguistics to visual communication, arguing that images, like language, possess a structured grammar for representing reality (the **ideational**

metafunction), enacting social relations (the **interpersonal metafunction**), and creating coherent messages (the **textual metafunction**). Their framework provides the essential vocabulary for analyzing how visual elements are composed and how they relate to other modes. Central to their theory is the concept of **modal affordance**, which suggests that each mode offers unique potentials for meaning-making. For example, the affordance of writing is its capacity for detailed, linear exposition, whereas the affordance of images lies in their ability to present complex information synthetically and evoke immediate emotional responses. This theoretical lens is critical for understanding why users select specific modal combinations to achieve their communicative goals online.

Early CMC research, conducted during the era of text-only interfaces, often centered on the perceived leanness of the medium. Theories like Social Presence Theory and Media Richness Theory posited that the absence of nonverbal cues in CMC hindered socio-emotional communication. However, Joseph Walther's Social Information Processing (SIP) theory (1992) offered a corrective, arguing that users adapt to the medium, developing text-based strategies over time to convey the relational information that is typically carried nonverbally. This adaptation is evident in the conventionalized use of emoticons, acronyms, and other textual cues. The contemporary digital environment, however, has moved far beyond these early constraints. The current landscape is one of hyper-richness, where the challenge is not compensating for a lack of cues but interpreting a potential overabundance of them.

More recent scholarship has begun to tackle the specific semiotic resources of the modern internet. Limor Shifman's (2014) work on internet memes identified them as complex cultural texts that blend image and text to spread through imitation and remixing, often carrying shared cultural norms and ideological stances. Similarly, research on Graphics Interchange Formats (GIFs) has explored their function as forms of "affective currency" that convey nuanced emotional reactions and cultural references more effectively than text alone. A significant stream of research has focused on the role of emojis. Scholars like Dresner and Herring (2010) have argued that emojis and emoticons are not merely decorative but serve crucial **pragmatic functions**, often indicating the illocutionary force of an utterance (e.g., distinguishing a joke from a serious statement). They function as digital analogues to body language and intonation, re-introducing a paralinguistic layer to textual communication. This body of work confirms that non-textual elements are integral, not peripheral, to meaning. However, despite the depth of research into these individual modes, a significant gap remains in understanding their holistic integration. Much of the literature treats these elements as separate phenomena, rather than as components of a unified multimodal message. This study seeks to bridge this gap by employing a social semiotic framework to analyze the complete communicative artifact—the social media post in its entirety—to understand how the interplay between text, image, and paralinguistic symbols co-constructs a single, coherent meaning.

Methodology

This study employed a qualitative content analysis methodology, guided by a social semiotic framework, to investigate the multimodal construction of meaning in internet discourse. This approach was chosen for its suitability in conducting an in-depth, interpretive analysis of communicative artifacts. The primary objective was not to quantify frequencies for statistical generalization but to identify, describe, and interpret the patterns of interplay between different semiotic modes within their authentic context.

The data for this research comprised a corpus of 200 public posts collected from the social media platform X (formerly Twitter). This platform was selected due to its widespread use and its inherent support for multimodal content, seamlessly integrating text (up to 280 characters), images, videos, GIFs, and emojis. To ensure the relevance and thematic coherence of the data, the sampling strategy targeted posts associated with a single, high-engagement cultural event that transpired over a one-week period in May 2024. The corpus was compiled using a specific, trending hashtag related to the event, and posts were selected to represent a variety of modal combinations, including text-only, text-plus-emoji, text-plus-image, and text-plus-image-plus-emoji. All collected data were anonymized by removing usernames and other personally identifiable information to adhere to ethical research standards for public data.

The unit of analysis was the individual post, considered as a complete multimodal ensemble. The analytical framework was adapted from Kress and van Leeuwen's (2006) model of semiotic analysis, focusing on the ideational, interpersonal, and textual metafunctions. A detailed coding scheme was developed to deconstruct each post. For the **ideational** component, the analysis noted what was being represented in the text and the image. For the **interpersonal** component, the analysis focused on how the post enacted a relationship with the viewer, examining elements such as tone (conveyed by text and emojis), gaze and social distance in images, and the overall pragmatic function of the post (e.g., to inform, entertain, persuade). For the **textual** component, the analysis centered on the composition of the modes, specifically the relationship between image and text (e.g., does the text anchor the meaning of the image, or does the image illustrate the text?) and the placement and function of emojis. Each post was coded by two independent researchers to ensure analytical rigor, and any discrepancies in coding were resolved through discussion to reach a consensus.

Results and Analysis

The analysis of the 200-post corpus revealed systematic patterns in the integration of textual, visual, and paralinguistic modes. The findings demonstrate a clear preference for multimodal communication over text-only formats and highlight the functional primacy of non-textual elements in conveying pragmatic and affective meaning. The results are organized into three sections corresponding to the research questions: patterns of modal integration, the functional hierarchy of modes in specific contexts, and the disambiguating role of paralinguistic cues.

Patterns of Modal Integration

A foundational finding of this study is that multimodal communication is the default, rather than the exception, in the sampled discourse. Text-only posts were significantly underrepresented in the corpus, indicating that users instinctively combine semiotic resources to construct their messages. The distribution of modal combinations across the corpus is detailed in Table 1.

Table 1: Frequency of Modal Combinations in the Sampled Corpus (N=200)

Modal Combination	Frequency	Percentage
Text-Only	14	7.0%
Text + Paralinguistic (Emojis)	58	29.0%
Text + Visual (Image/GIF)	42	21.0%
Text + Visual + Paralinguistic	86	43.0%
Total	**200**	**100.0%**

As Table 1 illustrates, the most prevalent format, accounting for 43.0% of the corpus, was the combination of all three modes: text, visuals, and paralinguistic cues. When combined with posts using Text + Emojis (29.0%) and Text + Visuals (21.0%), it becomes evident that a striking 93.0% of all analyzed posts were multimodal in nature. This overwhelming prevalence suggests that users perceive text alone as insufficient for nuanced communication in this environment. The integration of visual and paralinguistic layers appears to be a conventionalized strategy for adding the emotional resonance, personality, and contextual clarity that might otherwise be absent. This finding challenges any residual text-centric assumptions about digital communication and establishes the multimodal ensemble as the primary unit of analysis for understanding meaning-making on these platforms.

The Functional Hierarchy in Meaning-Making

Beyond the frequency of combination, the analysis revealed a clear functional hierarchy among the modes, which often shifted depending on the pragmatic intent of the message, particularly in contexts of humor, irony, and strong emotional expression. In these instances, the visual and paralinguistic modes frequently functioned not as supplements to the text, but as the primary determinants of the post's meaning, often overriding or completely inverting the literal semantics of the written words.

A powerful example of this was observed in posts intended to be ironic. One post contained the textual caption, "My week is off to a fantastic start." By itself, this text is unambiguously positive. However, it was paired with a popular GIF of a cartoon character frantically trying to put out multiple fires in a room. In this multimodal construction, the visual element does not simply complement the text; it completely subverts it. The **ideational** meaning of the GIF (chaos, disaster) overwrites the positive **ideational** meaning of the text, creating a new, singular meaning: "My week is off to a disastrous start." The humor and the true meaning of the post are carried almost entirely by the visual mode. The text serves merely as an ironic setup for the visual punchline.

This pattern of visual dominance was also apparent in the expression of complex emotions. A post featuring the text "Just got the project feedback" could be interpreted in numerous ways—as positive, negative, or neutral. However, the inclusion of an image, specifically a meme showing a person smiling faintly while a single tear rolls down their cheek, provides immediate and nuanced emotional clarification. The visual communicates a complex blend of disappointment, resignation, and the social pressure to appear fine. This affective nuance would be difficult and verbose to convey through text alone, demonstrating the unique **affordance** of the visual mode in representing synthetic emotional states. The analysis of such instances led to the categorization of dominant modes in specific pragmatic contexts, as summarized in Table 2.

Table 2: Dominant Mode in Conveying Primary Pragmatic Intent

Pragmatic Context	Dominant Mode	Explanatory Analysis
Humor & Irony	Visual (65%)	Visuals (especially GIFs and memes) provide the punchline or ironic counterpoint that subverts the literal meaning of the accompanying text.
Emotional Expression	Paralinguistic (55%)	Emojis provide a direct and unambiguous signal of the author's affective state, often clarifying or specifying the

		emotion implied by the text/image.
Information Dissemination	Textual (80%)	In posts primarily intended to inform (e.g., news updates, announcements), the text carries the core propositional content, with images serving an illustrative or attentional role.

As shown in Table 2, in 65% of posts coded as humorous or ironic, the visual mode was determined to be the primary carrier of the core message. In contrast, for posts focused on factual information, the textual mode remained dominant in 80% of cases. This demonstrates a clear division of labor among the modes, where users intuitively leverage the specific affordances of each to achieve different communicative ends.

The Role of Paralinguistic Cues as Disambiguators

The third major finding relates to the crucial function of paralinguistic cues, predominantly emojis, as powerful disambiguating agents. While visuals often set the overall frame of meaning, emojis frequently operate at a more granular level to fine-tune interpersonal tone and illocutionary force. They are rarely merely decorative; instead, they are functionally integral to how a message is to be received. Consider the textual phrase, "I can't wait." This statement is inherently ambiguous. Appending it with a smiling face with hearts emoji (😍) signals sincere and joyful anticipation. Appending it with an eye-rolling emoji (🙄) transforms the statement into a sarcastic complaint about something undesirable. The emoji does not just add emotion; it dictates the fundamental interpretation of the textual proposition. In our corpus, 38% of posts containing emojis used them in this disambiguating function, clarifying an otherwise neutral or ambiguous text.

Furthermore, emojis were found to moderate the relationship between text and image. In one analyzed post, a user posted a photo of a perfectly organized desk with the caption, "Finally productive." This text-image combination could be interpreted as a sincere expression of pride. However, the user added a clown face emoji (🤡) at the end. This single paralinguistic cue reframes the entire post as self-deprecating humor, implying that the "productivity" is a facade or a foolish endeavor. The emoji functions as an interpretive key, instructing the reader to view the text-image relationship through a lens of irony. Without this cue, the intended meaning would be lost. This highlights the sophisticated, multi-layered negotiation that occurs between all three modes in the construction of a single, coherent message.

Discussion

The results of this study offer substantial evidence that multimodality is not an optional or decorative feature of contemporary internet discourse but is central to its communicative logic. The overwhelming prevalence of multimodal constructions (93% of the corpus) confirms that users have moved far beyond the adaptive strategies described in early CMC theories like SIP. Instead of compensating for a lean medium, users are now leveraging a rich semiotic environment, skillfully orchestrating text, visuals, and paralinguistic cues to create meanings that are more nuanced, efficient, and affectively potent than what could be achieved through text alone. This reality requires a theoretical shift away from text-centric models of communication toward more integrated frameworks that recognize the co-equal status of all semiotic modes.

The findings regarding the functional hierarchy of modes contribute significantly to our understanding of this new communicative logic. The observation that visual elements often dominate in contexts of humor and irony aligns with the social semiotic concept of **modal affordance**. The affordance of language is linearity and logic, which makes it an ideal tool for setting up an expectation. The affordance of the image, however, is its capacity for immediate, synthetic presentation. In an ironic post, the text builds a linear expectation ("My week is off to a fantastic start") that the image can instantly and powerfully subvert (the dumpster fire GIF). This interplay leverages the distinct strengths of each mode to create a communicative effect—irony—that is a product of their combination. This synergistic meaning-making, where the whole is greater than the sum of its parts, is a hallmark of sophisticated multimodal discourse.

Furthermore, the role of paralinguistic cues as powerful disambiguators has profound implications. The finding that a single emoji can fundamentally alter the illocutionary force of a statement supports the work of scholars like Dresner and Herring (2010) but extends it by showing how these cues mediate not just text, but the entire text-image relationship. This suggests a three-tiered interpretive process for many social media posts: the text provides a propositional base, the image provides a contextual and affective frame, and the emoji provides a final, precise interpretive key for understanding tone and intent. The failure to account for all three layers can lead to a complete misinterpretation of the message. This has critical implications for the field of artificial intelligence, particularly in the development of sentiment analysis and content moderation algorithms. Systems trained primarily on textual data will inevitably fail to grasp the true sentiment or intent of a message where meaning is carried or inverted by non-textual elements. Effective AI must develop the capacity for holistic, multimodal interpretation.

Finally, this study has limitations that open avenues for future research. The corpus was limited to a single platform, X, and focused on a single cultural event. Future research should conduct cross-platform analyses to explore whether these patterns of modal hierarchy hold true on more visually-centric platforms like Instagram or video-based platforms like TikTok. Additionally, the cultural context of the data is inherently Anglophonic and Western; cross-cultural research is needed to investigate how the use and interpretation of multimodal ensembles vary across different linguistic and cultural backgrounds. A longitudinal study could also track the evolution of these multimodal conventions over time as platforms and user behaviors continue to change.

Conclusion

This research set out to investigate the intricate interplay of textual, visual, and paralinguistic elements in contemporary internet discourse. Through a qualitative analysis of a corpus of social media posts, this study has demonstrated that online communication is a fundamentally multimodal phenomenon where meaning emerges from the dynamic and synergistic relationship between different semiotic modes. The investigation yielded three primary conclusions. First, multimodal constructions combining text, visuals, and paralinguistic cues are the conventional standard, not the exception, rendering text-only communication a minority practice in many social media contexts. Second, a functional hierarchy exists among these modes, with visual elements often assuming primacy in conveying the core pragmatic intent in affectively charged contexts like humor and irony, effectively overriding or subverting the literal meaning of accompanying text. Third, paralinguistic cues, especially emojis, serve as essential and powerful tools of disambiguation, providing the interpretive key needed to understand illocutionary force and interpersonal tone for the entire multimodal message.

The central argument of this paper—that multimodality in internet discourse is transformative rather than merely additive—is robustly supported by the analysis. The meaning of a typical social media post is not a simple aggregate of its components but a composite meaning that could not exist without the specific interplay between them. This work contributes to the fields of communication studies and social semiotics by providing a detailed, empirical account of these integrative practices and by advocating for analytical frameworks that treat the multimodal ensemble as the primary unit of meaning. The findings underscore the sophistication of everyday digital literacy, revealing the complex semiotic calculations that users perform intuitively and instantaneously. As our social, professional, and civic lives become increasingly mediated by these multimodal platforms, a deep understanding of this new grammar of communication is not merely an academic exercise but a critical necessity for effective and empathetic participation in the digital age. Future research must continue to explore this evolving landscape, particularly by examining cross-platform and cross-cultural variations, to build a more comprehensive theory of digital multimodal communication.

References

Dresner, E., & Herring, S. C. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. Communication Theory, 20(3), 249–268. https://doi.org/10.1111/j.1468-2885.2010.01362.x

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. Communication Research and Practice, 2(1), 47-62. https://doi.org/10.1080/22041451.2016.1155332

Kress, G., & van Leeuwen, T. (2006). Reading images: The grammar of visual design (2nd ed.). Routledge.

Kress, G. (2010). Multimodality: A social semiotic approach to contemporary communication. Routledge.

Shifman, L. (2014). Memes in digital culture. The MIT Press.

Stöckl, H. (2004). In between modes: Language and image in printed media. In E. Ventola, C. Charles, & M. Kaltenbacher (Eds.), Perspectives on multimodality (pp. 9–30). John Benjamins.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. Communication Research, 19(1), 52–90. https://doi.org/10.1177/009365092019001003

Walther, J. B. (2011). Theories of computer-mediated communication and interpersonal relations. In M. L. Knapp & J. A. Daly (Eds.), The Sage handbook of interpersonal communication (4th ed., pp. 443–479). Sage Publications.