December 2024

Affective and Cognitive: Exploring the Multi-modal data for Predicting User Engagement Behavior in Live Streaming Commerce

Guang Xu
Renmin University of China, 2020000919@ruc.edu.cn

Ming Ren
Renmin University of China, renm@ruc.edu.cn

Qun Zhou
Renmin University of China, zhouq1105@ruc.edu.cn

Recommended Citation

Xu, Guang; Ren, Ming; and Zhou, Qun, "Affective and Cognitive: Exploring the Multi-modal data for Predicting User Engagement Behavior in Live Streaming Commerce" (2024). *ICIS 2024 Proceedings*. 12.
https://aisel.aisnet.org/icis2024/data_soc/data_soc/12

Affective and Cognitive: Exploring the Multimodal data for Predicting User Engagement Behavior in Live Streaming Commerce

Completed Research Paper

Guang Xu
Renmin University of China
Beijing, China
2020000919@ruc.edu.cn

Ming Ren
Renmin University of China
Beijing, China
renm@ruc.edu.cn

Qun Zhou
Renmin University of China
Beijing, China
2019000875@ruc.edu.cn

Abstract

The boom of the live streaming commerce provides a wealth of multimodal information, which provide more possibilities for predicting user engagement. Existing studies usually employ a unified framework to process and fuse multimodal information, which fails to understand user engagement behaviors deeply. This paper proposes to handle the multimodal information in live streaming commerce from affective and cognitive perspectives. An Elm-based Multimodal Analysis Framework (EMAF) is presented, which extracts features from multimodal information from affective and cognitive perspectives respectively and predicts user engagement behavior in real-time in live streaming commerce. A module named MD-Transformer is designed to integrate product details more effectively. Experiments have been conducted on a real-world dataset, and the results demonstrate the advantages of our framework against the state-of-the-art multimodal fusion methods.

Keywords: live streaming commerce, user engagement behavior, ELM, affective-related information, cognitive-related information, multimodal data

Introduction

In recent years, advancements in information and communication technology have fostered the development of live streaming commerce. This emerging business model combines e-commerce and live streaming technology, enabling users to interact with anchors in real-time while purchasing products (Zhang et al., 2022) and providing a brand-new social shopping experience (Xin et al., 2023). Compared with the traditional e-commerce, live streaming commerce provides users with more modalities of information (Shin et al., 2023), not limited to numerical data, text and images, but also including video and audio. At the same time, the interactive data and the wealth of multimodal data from in live streaming scenarios provide more possibilities for researchers and practitioners to study various tasks.

Capturing user engagement is essential for the prosperity of live streaming commerce, which can be viewed as an efficient approach for anchors or merchants to acquire long-term competitive advantages (Yan et al., 2023; Zhang et al., 2024). Amid a landscape of many platforms and minimal barriers to entry, enhancing user engagement significantly aids in retaining users (Kang et al., 2021; Wongkitrungrueng & Assarut, 2020). A preliminary analysis of our live streaming dataset reveals a positive correlation between user engagement and live streaming sales, as shown in Figure 1. This correlation suggests that increased engagement during live streams likely enhances revenue. Throughout the live streaming process, anchors can strategically boost engagement by employing multimodal content (Luo et al., 2024a). Consequently, investigating user engagement through multimodal data is of paramount importance.

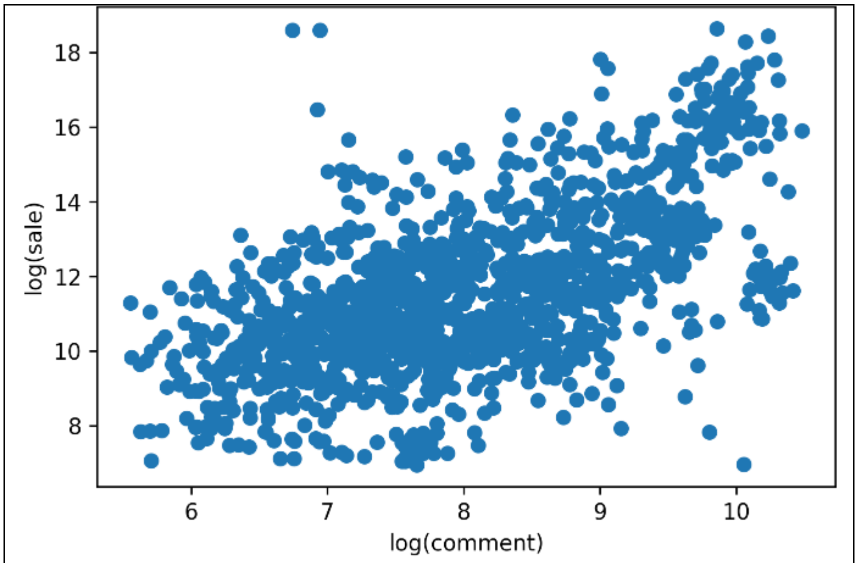

Figure 1. The Relationship between Sale and Comment

To use multimodal data, most existing studies input all modal data into deep models to automatically extract feature (Xu et al., 2023), which might be effective for subsequent tasks but are incomprehensible to humans. Some studies extract features based on human-defined feature (Chen et al., 2023), but these methods are limited by human understanding and do not fully utilize the potential of multimodal data. Further research is needed on the use of multimodal data to predict user engagement behavior in live streaming commerce.

According to the elaboration likelihood model (ELM) theory (Petty et al., 1983), the information promotes engagement by influencing users' affective route and cognitive route (Lo et al., 2022), which appears to provide a potential framework for predicting user engagement behavior through multimodal data in live streaming commerce. From the cognitive route perspective, the multimodal information in live streaming offers a richer and more dimensional product presentation. Users can see not only static information of products, but also understand practical applications and detailed demonstrations by the anchor through videos. This comprehensive information display significantly enhances users' cognition about the product, and its perceptual and persuasive advantages have been proven in previous user behavior research. From the affective route perspective, the information in live streaming can trigger stronger emotional resonance among users. Anchor's influence, commentary style, and background of live room are key factors in stimulating users' affective route. The anchor provides users with a sense of belonging in the live streaming through this information, thereby significantly increasing their engagement.

Existing studies usually employ a unified framework to process and fuse multimodal information, which fails to adequately distinguish the differences in information at affective and cognitive levels. Therefore, there is an urgent need to develop a new method to handle the differences in multimodal information from affective and cognitive perspectives in live streaming commerce. Such a framework should not only be able to handle information from different modalities, but also be flexible enough to capture the differences of the information in predicting user engagement behavior.

This paper proposes an **E**lm-based **M**ultimodal **A**nalysis **F**ramework (**EMAF**) for predicting user engagement behavior in real-time in live streaming commerce. The framework incorporates more modal information, including numerical, textual, image, and real-time modalities such as video and audio. Firstly, the framework draws inspiration from ELM theory and extracts features from multimodal information from affective and cognitive perspectives. Affective-related information can also be referred to as external clues, which may include the influence of the anchor, background of live room and visual effect, etc. This information influences user engagement behavior through their affective route. Cognitive-related information involves user's in-depth processing and cognitive thought of information, usually including detailed information about the product in live streaming commerce. Such information, which will be accepted and deeply processed by users, influences their engagement behavior through cognitive route. The framework integrates multimodal information during the live streaming and predicts user engagement behavior in real-time.

The contributions of this study are threefold. Firstly, we draw on the ELM theory and expand its application scenarios, modeling the multimodal information into as affective-related and cognitive-related information, which will help to deeply understand the impact of different information in predicting user engagement behavior in live commerce. Secondly, we propose EMAF, which, to our knowledge, is the first framework to predict user engagement behavior in real-time based on multimodal data fusion from both cognitive and affective perspectives in live streaming commerce. Thirdly, we introduce the innovative MD-Transformer module within our EMAF framework, designed to manage multidimensional product details effectively. This model fully considers that the connections between different modalities within the same product are much stronger than those between the same modality across different products, thus significantly increasing the generalizability of EMAF and enabling its application across various tasks and scenarios in live streaming commerce.

The remainder of this paper is structured as follows. Section 2 reviews related work on user engagement behavior in live streaming commerce, Elaboration Likelihood Model, and multimodal data for predictive models. Section 3 presents the problem formulation, and Section 4 describes the proposed framework and the computational methodologies employed. Section 5 presents the experiment settings, results, and discussion. Section 6 concludes the research and discusses the future work.

Related work

User engagement behavior in live streaming commerce

Live streaming commerce is an emerging business model that combines live streaming technology with e-commerce, allowing users to watch and purchase products or services through real-time video streaming (Huang & Ma, 2024). In this model, user engagement refers to the interaction between users and products or anchors within the live streaming commerce environment, surpassing transactions and immediate purchasing intentions. The modes of user engagement include but are not limited to watching, comment, and like (Addo et al., 2021; Zheng et al., 2022). At the same time, researchers are increasingly focusing on the dimensions of user engagement, as shown in Table 1. Although there is no universal type of user engagement in live streaming commerce and user engagement in different studies is expressed in various dimensions, but overall, live streaming integrates user engagement through interactions within the service system (Zheng et al., 2022).

In current research, researchers primarily explore user engagement through two paths. The first is to investigate the influencing factors of user engagement, such as Xue et al. (2020) who found that real-time interaction influences perceived usefulness, perceived risk, and psychological distance, thereby promoting user engagement behavior. The second is to explore how user engagement behavior during live streaming affects final decision-making behavior, such as purchase intention, customer acquisition, and influence willingness (Gao et al., 2021; Zheng et al., 2022). Zheng et al. (2022) explore the relationship between user engagement behavior, purchase intention, and customer acquisition by collecting visits, likes, and comments in the live streaming environment. However, most studies are conducted through hypothesis testing and use the final static data at the end of the live streaming as quantitative indicators when collecting data related to user engagement behavior. Currently, there is still limited research on how to use the multimodal information provided by the platforms to predict user engagement behavior and its influencing factors during a live streaming in real-time. Our research will delve into this aspect, utilizing the multimodal information in conjunction with ELM to conduct our exploration.

Authors	Dimension	Research context
Addo et al. (2021)	Likes, visits, chats and exposure time	Study how user engagement affects purchase intentions in live-streaming
Zheng et al. (2022)	Visits, likes, comments	Explore the relationship between customer engagement behavior, purchase intention, and customer acquisition
Lin et al. (2021)	Viewer emotion, viewer tips, likes, comments, et al.	Explore the role of emotions in interactive and dynamic business environments, such as live streaming.
Kang et al. (2021)	Likes, gifts, comments	Study the dynamic impact of interactivity on user engagement behavior through the strength of connections in live streaming commerce
Yan et al. (2023)	Visits, shares, exposure time, et al.	Explore how swift guanxi and perceived enjoyment influence customer engagement through in live streaming by IT affordances.
Wongkitrungrueng and Assarut (2020)	Visits, shares, revisits, recommends, et al.	Propose a framework to study the relationship between user perceived value, user trust, and participation in live streaming.
Xue et al. (2020)	Shares, recommends, visits, et al.	Explore the moderating effect of real-time interaction on social commerce participation based on the S-O-R paradigm and susceptibility to information impact
Gao et al. (2021)	Likes, gifts, comments	Explores the audience's decision-making process through two different routes by ELM, consisting of a central and a peripheral route.

Table 1. Dimension of user engagement in different research

Elaboration Likelihood Model (ELM)

The Elaboration Likelihood Model is a persuasion and information theory that explains how people process persuasive information and form attitudes (Yang et al., 2021). It explicates how the dual routes of affective and cognitive responses affect an individual's attitude and subsequent behavior (Bhattacherjee & Sanford, 2006). The cognitive route involves a high level of cognitive effort and information processing, typically occurring when an individual is highly engaged, interested in the topic, or has high cognitive ability. On this route, individual carefully evaluate the quality and reasonableness of the information (Chou et al., 2022). By comparison, the affective route requires less cognitive effort. In this case, individual may rely on simple or external cues, such as the persuader's credibility or the attractiveness of the information presentation (Chou et al., 2022). ELM has been widely used in persuasion scenarios in live streaming commerce to explain changes in user attitudes or behaviors, such as user engagement, impulse buying, purchase intention and response intention (Gao et al., 2021; Luo et al., 2024b; Xiao et al., 2023).

In live streaming scenarios, users can access product-centered information as well as a wealth of external cues from the environment (Gao et al., 2021). Specifically, before making their final purchasing decisions in the live streaming, users carefully consider information related to the products (Liu et al., 2023). Given that the information provided in live streaming is multimodal, product-related information mainly comes from text description, numerical description, anchors' commentary, and image-based explanations, which help enhance users' cognitive level regarding the products. At the same time, users' positive or negative emotions in live streaming commerce may be influenced by external cues from the live streaming environment, including the anchor's influence, commentary style and background of live room, which in turn affect the user's attitude and perception towards the product (Li et al., 2024; Lin et al., 2021; Xu et al., 2023). Therefore, both two different information processing routes ultimately affect user behavior in live streaming.

Multimodal data for predictive models

Multimodal data refers to a collection of data in different forms, which includes text, images, numerical data, etc. By fusing the multimodal data, the information among various modality can be complementary, thus providing more effective output for downstream tasks than that of a single modality information (Gan et al., 2024; Xu et al., 2024a). Early research has explored various strategies for multimodal fusion (Baltrušaitis et al., 2018), including kernel-based methods, graphical models, and neural networks. With the advent of attention mechanisms, their application in multimodal fusion has become increasingly prevalent. These models have powerful representation capabilities and achieve more effective information fusion through Self-attention (Cui et al., 2020) or cross-modal attention (Yang et al., 2022).

Live streaming commerce, as an emerging multimodal information application scenario, has also attracted widespread attention in multimodal fusion. Live streaming platform utilizes various media forms, such as text, audio, and video, to provide channels for anchors to showcase products and interact with users (Wang et al., 2024). This multimodal presentation and interactive method not only enrich the user experience but also add new dimensions to the effectiveness of live streaming commerce (Kang et al., 2021). Predictive models have always played an important role in data analysis and business applications. With the increasing complexity of multimodal data, the prediction model has gradually developed from a single modality to a complex system that can integrate multiple modal information. In live streaming commerce, multimodal fusion provides richer input data for prediction models, significantly improving their performance in sales prediction (Xu et al., 2024b), live streaming recommendation (Chen & Liu, 2024), traffic forecasting (Lin et al., 2023), and product return prediction (Xu et al., 2023). These models integrate multimodal information such as text, images, and audio to better capture user's interests and needs, providing data support for optimizing live streaming strategies. The combination of multimodal fusion and prediction models has continuously made progress in the research and application of live streaming commerce. Although this paradigm has achieved significant results and proven its effectiveness in live streaming commerce, it still lacks a deep exploration of how different types of information affect predicting user engagement behavior.

In summary, our study is based on ELM and further explores the impact of different multimodal information on predicting user engagement behavior in live streaming commerce. According to ELM theory, the influence of information can be unfolded through two routes. The cognitive route mainly involves the processing of high-quality and thought-provoking information, while the affective route relies on some simpler heuristic clues or shortcuts. In live streaming commerce, different modalities of information may convey both cognitive-related and affective-related information simultaneously. For example, the audio modality from the anchor not only conveys detailed product information, but also promotes user engagement behavior through affective expression. Therefore, we have summarized the research on ELM in live streaming commerce, as shown in Table 2, and use corresponding processing methods to distinguish multimodal information into these two categories. Cognitive-related information is centered around the product, including product detail information composed of text, image, and numerical data, audio modality information for introducing the product, and video modality information for displaying the product; Affective-related information is heuristic clues in the live streaming environment, including anchor influence, anchor commentary style (audio), visual effects (video), and the background (atmosphere) of the live streaming room.

Information	Affective	Cognitive	Authors
Anchor influence	√		Gao et al. (2021); Xiao et al. (2023)
Anchor commentary style	√		Luo et al. (2024a)
Visual effects	√		Xin et al. (2024)
Background	√		Tong et al. (2023)
Product detail information		√	Gao et al. (2021); Luo et al. (2024a); Xin et al. (2024)

Table 2. Affective-related and Cognitive-related information in ELM

Problem formulation

In this paper, the user engagement behavior U focuses primarily on comment. Our goal is to predict the total amount of user comment for the period *T+1*, denoted as U^{T+1}, based on information from the period *T*. From the multimodal information during live streaming, we extract affective-related information A and cognitive-related information C. The affective-related features include the anchor's influence AI, the background of the live room BG, the anchor's commentary style AA, and visual effects of the live streaming AV, while the cognitive-related features consist of product details feature CP, the anchor's commentary CA, and visual information of the live streaming CV, which consist of numerical data, text, and image·

Certain information, such as AI and BG, remain static during the live streaming. They present to users as numerical data, text, and image. Other information is dynamic, and will change with the progress of the live streaming. Since the period T consists of n smaller and continuous time segments, i.e., T= $\{T_1, T_2, ..., T_n\}$, accordingly, the affective-related information for T consists of a collection of smaller time-scale segments. Specifically, we have AA^T as $\{AA_1^T, AA_2^T, ..., AA_n^T\}$, and AV^T as $\{AV_1^T, AV_2^T, ..., AV_n^T\}$. The cognitive-related information for T also comprises a collection of smaller segments. Specifically, we have CP^T as $\{CP_1^T, CP_2^T, ..., CP_m^T\}$, CA^T as $\{CA_1^T, CA_2^T, ..., CA_n^T\}$, and CL^T as $\{CL_1^T, CL_2^T, ..., CL_n^T\}$.

Utilizing the multimodal data as input, and user comment behavior data within the period *T+1* as output, we aim to train a model predicting user engagement behavior in real-time, by extracting affective-related information and cognitive-related information, as shown in Eq. (1).

$$U^{T+1} = f_\theta\left(f_\rho(AI, BG, AA^T, AV^T), f_\tau(CP^T, CA^T, CL^T)\right), \tag{1}$$

where f_ρ is a module with a set of trainable parameters ρ, used for aggregating factors affecting users' affective route. f_τ is a module with a set of trainable parameters τ, used for aggregating factors affecting users' cognitive route. f_θ is a prediction model with a set of trainable parameters θ.

The EMAF framework

This section presents the EMAF framework for predicting user behavior in live streaming commerce. As depicted in Figure 2, EMAF consists of three components, including information representation, information fusion, and user engagement behavior prediction.

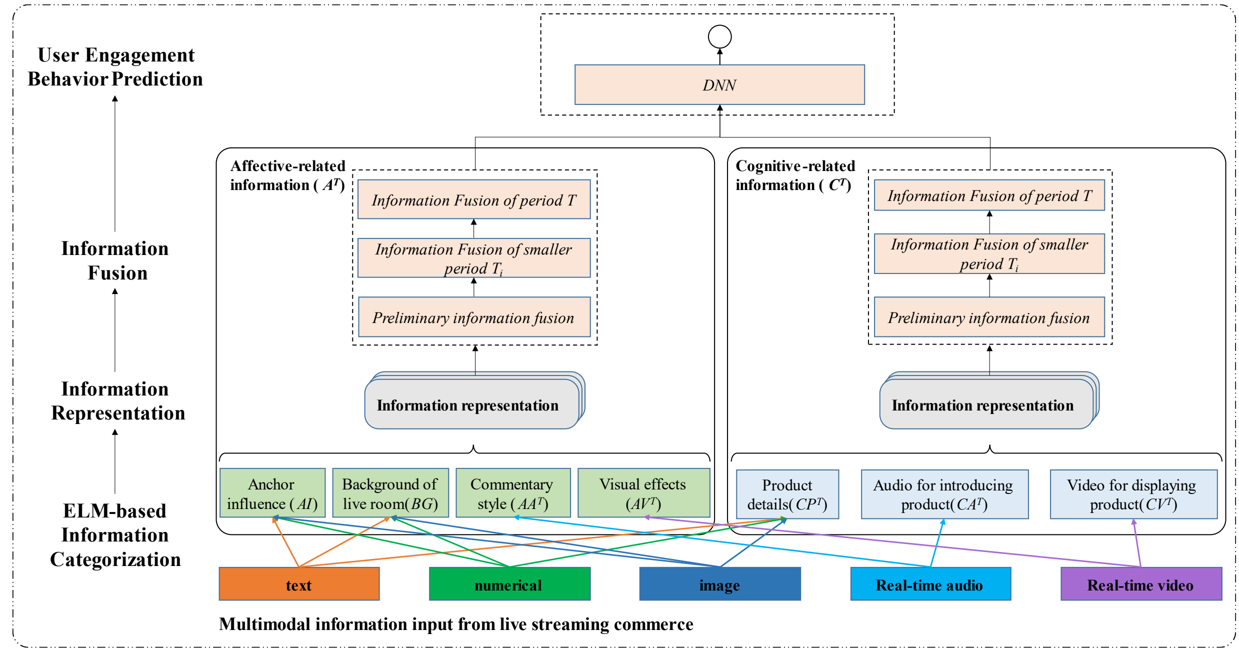
User Engagement Behavior Prediction
Information Fusion
Information Representation
ELM-based Information Categorization
DNN
Affective-related information (A^T)
Information Fusion of period T
Information Fusion of smaller period T_i
Preliminary information fusion
Information representation
Anchor influence (AI)
Background of live room(BG)
Commentary style (AA^T)
Visual effects (AV^T)
Cognitive-related information (C^T)
Information Fusion of period T
Information Fusion of smaller period T_i
Preliminary information fusion
Information representation
Product details(CP^T)
Audio for introducing product(CA^T)
Video for displaying product(CV^T)
text
numerical
image
Real-time audio
Real-time video
Multimodal information input from live streaming commerce

Figure 2. EMAF Model framework

Information Representation

This module aims to map multimodal data to specific vector spaces and obtain its feature representations through traditional methods and deep neural networks.

For audio in the period T composed of smaller time segments $\{T_1, T_2, ..., T_n\}$, we use the openSMILE (Eyben et al., 2013) toolkit to extract its affective and cognitive factors. Previous studies have demonstrated the effectiveness of openSMILE in feature extraction (Rahmani et al., 2023). In this study, on the one hand, by extracting feature such as rhythm and intensity by openSMILE, we can identify anchor's commentary style, which a kind of affective- related information, represented as $\{aa_1^T, aa_2^T, ..., aa_n^T\}$. On the other hand, features extracted by openSMILE, such as pitch, accentuation, speech rate, and voice quality, can determine the clarity and quality of audio, which directly impact the users' cognition and understanding of the live content and are a kind of cognitive-related information, represented as $\{ca_1^T, ca_2^T, ..., ca_n^T\}$.

For real-time video information, live streaming visuals not only convey specific product information to the user, but also influence users' emotion and experience through the content of the images. Like audio, many factors affect the presentation of live streaming, such as the layout of the live room and the way products are displayed, which can be predetermined by design. To analyze the real-time video information of live streaming, this study adopts a method of extracting key frames from videos, which allows for capturing important moments of the live streaming and analyzing the impact of visual elements. Considering that the *ResNet* model (He et al., 2016) can effectively recognize and analyze important information from the key frames, we use a pre-trained *ResNet-50* model to extract visual information features as $\{cl_1^T, cl_2^T, ..., cl_n^T\}$, which can represent the cognitive-related information conveyed to users. At the same time, to evaluate the impact of the key frames on users' affective route, we construct a visual emotion recognition model based on the *EmoSet* dataset (Yang et al., 2023) to extract the visual effects from them, represented as $\{av_1^T, av_2^T, ..., av_n^T\}$.

For unstructured text, it provides users with detailed introductions, such as product title and live room title. High-quality titles enable users to quickly judge their relevance. We employ the pre-trained *BERT* model (Devlin et al., 2019) to capture the feature of unstructured text and fine-tune it for downstream tasks. For images, the content conveyed by them is much richer than text, and creating attractive photos is crucial, especially in live streaming commerce, since users have relatively less time to think and purchase (Xu et al., 2023). To accurately capture complex information in pictures, we use the pre-trained *ResNet-50* model to extract features. For numerical data, which often involves product prices, discounts, etc., we use a multilayer perceptron (Rynkiewicz, 2019) to process them, which can effectively extract non-linear features.

Information Fusion

When watching a live streaming, user not only receives detailed content about product-centered information, but also gains rich external cues from the environment, such as the influence of anchor and the background of live room. All the information has an important impact on how users feel about the live streaming, and then ultimately affect user engagement behavior in live streaming (Zhu et al., 2017). For example, after an anchor introduces a product, users will ask the anchor questions through comments. At the same time, the anchor actively responds to the questions and urge the users to place an order with reasons such as large discounts and limited quantities. These all have a promoting effect on user engagement behavior. Therefore, we consider integrating information from both affective and cognitive routes, as shown in Figure 2. The Information Fusion module includes three fusion processes, which will be introduced one by one.

Preliminary Information Fusion

We notice that, for AI, BG, and CP^T, all the information exhibits multimodal characteristics. For example, CP^T may include text description of title, numerical description of price, and image description. Therefore, before integrating information from affective and cognitive routes, it is necessary to perform preliminary information fusion on these elements. Here, given that the Transformer structure can consider the interrelationship and dependency among different features, we employ it to generate a more powerful and comprehensive fused feature representation. Each module in the Transformer structure contains Multi-

Head Self-Attention mechanism, feed-forward neural network, residual connectivity and layer normalization.

Unlike the other two types of information, CP^T in a live streaming not only has multimodal characteristic, but also an important quantitative dimension, namely, the number of products sold. However, traditional Transformer structure can only accept inputs of two-dimensional tensor and cannot handle three-dimensional tensor. Therefore, we design a novel transformer structure for product details in live streaming commerce, called MD-Transformer. The design of this structure is based on the consideration that the connections between different modalities within the same product are much stronger than those between the same modality across different products (Xu et al., 2024a). Therefore, the structure first fuses the features of the product at the multimodal dimension, and then further fuses them at the quantitative dimension to obtain the final product features cp^T, as shown in Eq. (2) to (6).

$$Self\_Attention(Q,K,V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2}$$

$$y_i = Average\left(LayerNorm\left(X_i + Self\_Attention\left(X_iW_q, X_iW_k, X_iW_v\right)\right)\right) \tag{3}$$

$$Y = [y_1, y_2, \ldots, y_m]^{\top} \tag{4}$$

$$V = LayerNorm\left(Y + Self\_Attention\left(YW_q, YW_k, YW_v\right)\right) \tag{5}$$

$$Output = LayerNorm\big(V + FeedForward(V)\big), \tag{6}$$

where X represents the feature of product details and $X \in R^{m*l*d}$; m is the number of products; l is the number of all modalities for each product detail; and d is the dimension of each modality's representation; X_i represents the matrix of all modal features for the i-th product and $X_i \in R^{l*d}$; $Average(\cdot)$ denotes the averaging operation and transforms the dimensions from R^{l*d} to R^d ; $LayerNorm(\cdot)$ denotes layer normalization; $FeedForward(\cdot)$ represents the feed-forward neural network; $Y \in R^{m*d}$; $W_q, W_k, W_v \in R^{d*d}$ are trainable parameters.

As for *AI* and *BG*, we use traditional transformer structure to integrate and obtain the final features ai and bg, just as shown in Eq. (2), (5), and (6).

Information Fusion of smaller period *T_i*

As mentioned before, the reasons that these users enter the live streaming are greatly related to the anchors' influence and the background of the live room. Because these are all affective-related information that triggers users to enter a live streaming, and they are also prerequisites for users to have some engagement behaviors, such as comment.

Therefore, for each smaller period T_i within period T, we define the affective-related features as composed of the anchor's influence feature ai, the background of the live room feature bg, the anchor's commentary style feature aa_i^T, and visual effects of the live streaming feature av_i^T, while cognitive-related features consist of product details feature cp^T, the anchor's commentary feature ca_i^T, and visual information of the live streaming feature cv_i^T . As mentioned before, the Transformer architecture can consider the relationships between different features to generate a more powerful feature representation. Thus, it continues to be used to merge the affective-related information and cognitive-related information, to achieve the final information representation for each time segment T_i. In order to adapt different features to the Transformer structure, linear transformation is used to unify their dimensions in advance, and finally, the affective-related representation a_i^T and cognitive-related representation c_i^T at time T_i are obtained through this structure. The aforementioned steps are shown in Eq. (7), (8), (2), (5), (6), taking affective-related information as an example.

$$ai, bg, aa_i^T, av_i^T = ai \cdot W^{ai}, bg \cdot W^{bg}, aa_i^T \cdot W^{aa}, av_i^T \cdot W^{av} \tag{7}$$

$$Y = [ai, bg, aa_i^T, av_i^T]^{\top} \tag{8}$$

Information Fusion of period *T*

Each period T consists of n smaller time segments, which means that these are generated over time in sequence. Each smaller time segment T_i not only contains the information affecting user engagement behavior at that moment, but also implies a temporal sequence relationship. In this context, if we want to obtain the final representation of affective-related feature ha_n^T and cognitive-related feature hc_n^T for the period T, it becomes particularly important to consider the temporal dependencies and sequential relationships between data points. Long Short-Term Memory Network (LSTM) are a type of recurrent neural network, suitable for processing time series data. Because it can consider the information from each smaller period T_i as well as the information from all previous periods. By maintaining a long-term internal state, LSTM can capture dynamic features in the time series and manage the flow of information effectively through a gating mechanism, thus ultimately providing a deep understanding and representation of each larger period T. Therefore, for period T, we use LSTM to further process these time series data, as shown in Eq. (9) to (13), taking affective-related feature ha_i^T as an example.

$$f_i^T = \sigma\left(W_f[ha_{i-1}^T, a_i^T] + b_f\right) \tag{9}$$

$$r_i^T = \sigma(W_r[ha_{i-1}^T, a_i^T] + b_r) \tag{10}$$

$$o_i^T = \sigma(W_o[ha_{i-1}^T, a_i^T] + b_o) \tag{11}$$

$$Cell_i^T = f_i^T * Cell_{i-1}^T + r_i^T * tanh(W_c[ha_{i-1}^T, a_i^T] + b_c) \tag{12}$$

$$ha_i^T = o_i^T * tanh(Cell_i^T) \tag{13}$$

User Engagement Behavior Prediction

Our model uses the multimodal information of live streaming to predict user engagement behavior in real-time from the perspectives of affective-related and cognitive-related information for the next period. Therefore, it is important to integrate the final affective-related representation ha_n^T and cognitive-related representation hc_n^T over period T to construct an accurate predictor. Extensive research shows that Deep Neural Network (DNN) exhibits superior performance across various fields. Inspired by this, we also choose DNN as the core predictor and expect to produce accurate prediction. The input to the predictor is the ha_n^T and hc_n^T generated from the information fusion module, and the output is the number of user comments in the next period *T+1*, as shown in Eq. (14) to (15). In the construction of the DNN, we control the amount of information from different factors by using α (default value of 0.5). Besides, Sigmoid is used as the primary activation function to achieve nonlinear mapping, and Dropout is used to prevent overfitting.

$$X^T = (1 - \alpha) * ha_n^T + \alpha * hc_n^T \tag{14}$$

$$U^{T+1} = W_{fc2} \cdot \sigma\left(W_{fc1} \cdot X^T + b_{fc1}\right) + b_{fc2} \tag{15}$$

Experiments

Data description

Our study utilizes a real-world dataset from Taobao Live, which is a leading live streaming shopping platform in China and occupies a significant market share particularly in clothing. The data collection period is from October 1 to 31, 2021. Through the analysis of the Taobao Live platform and data availability, the types of data collected are as shown in Table 3. Additionally, we carefully exclude incomplete data points to ensure the fairness and accuracy of the experiments. Consequently, we have 1437 complete live streaming data points.

	Description	Modality
Product	Name of a product	Text
	Discount description of a product	Text
	Image of a product	Image

	Discounted price of a product	Numerical
	Price of a product	Numerical
Live room	Tags associated with the live room on the platform	Text
	Name of a live room	Text
	Image of a live room	Image
	Duration of a live streaming session	Numerical
Anchor	Compilation of fields for all an anchor' live rooms in the platform	Text
	Pre-defined category of an anchor on the platform	Text
	Tags associated with an anchor on the platform	Text
	A score rating the ability of an anchor by the platform	Numerical
	Number of views during live streaming in the past 30 days	Numerical
	Average number of comments across all live streaming sessions	Numerical
	Average number of likes in all live streaming sessions	Numerical
	Total number of comments in all live streaming sessions	Numerical
	Count of fans who favorite an anchor	Numerical
	Total number of fans of an anchor	Numerical
	Total likes across all live streaming sessions	Numerical
Video	A series of videos featuring a streaming session	Video
Audio	A series of audios featuring a streaming session	Audio
Comment	A series of user behavior data	Numerical

Table 3. Description of collected data types

Further, we note that the platform counts the total user engagement behavior for each live streaming every 30 minutes, which provides feasibility to explore user engagement behavior in real-time from the perspectives of affective-related and cognitive-related information. Therefore, we segment the collected live streaming data points into individual points every 30 minutes. As for product details, we record the listing time of each product during the live streaming and categorize them by time segment. And then we use a cumulative aggregation method, which means that the product details from one time segment is also included in the next. To analyze visual effects of the live streaming, visual information of the live streaming, anchor's commentary and its style, we randomly sample each 30-minute segment. Specifically, we extract a 10 to 20-second clip every 5 minutes, and process these clips by using the method described in Information Representation to obtain corresponding data feature representations. The anchor's influence and the background of the live room are key factors attracting user and are prerequisites for users to have some engagement behaviors in the live streaming. Additionally, we exclude the first and last segments of a live streaming due to unstable user engagement behavior. Ultimately, we have 11,949 data points used for the experiment. The statistics of the dataset are presented in Table 4.

	Dataset of the live streaming commerce
Record	11949
Anchor	106
Product	116399
Live room	1437
Audio clips	11949
Video clips	11949
Comment(logarithm)	Average = 5.4744; variance = 1.0997

Table 4. Descriptive Statistic of the dataset

Experiment settings

All experiments are conducted on a Linux server using Python, with an Intel(R) Xeon(R) CPU E5-2637 v4 @ 3.50GHz CPU and an NVIDIA GeForce RTX 2080 Ti GPU. Our deep learning framework is based on PyTorch. To evaluate model performance and optimize hyperparameters, we randomly divide the dataset into a training set, validation set and test set, with the training set comprising 70% of the data, the validation set 15%, and the test set the remaining 15%. The best model is selected via the validation set, and the final results on the test set are used to demonstrate the model's performance. Finally, since our model is built on Transformer and LSTM structure, to better train the model and achieve optimal performance, we use SGD as the optimizer and a cosine annealing learning rate scheduling strategy to dynamically adjust the learning rate. For comparative experiments, we use the same strategy to individually fine-tune the parameters of each model.

Evaluation metrics

As this study addresses a regression task, we employ several evaluation metrics to assess our model's comprehensive performance. **MAE** calculates the mean absolute error between actual and predicted values. **MSE** represents the mean squared error between actual and predicted values. **MAPE** indicates the mean absolute percentage error, showing differences between actual and predicted values as a percentage of actual values. **R-squared Score (R^2)**, the coefficient of determination, measures how well the model fits the data. These metrics facilitate a comprehensive evaluation of the model's accuracy and robustness.

Experimental results

Comparison of performances of the predictive methods

To verify the effectiveness of our proposed model in fusing multimodal information, we compare it with several popular multimodal fusion methods, including MFB (Yu et al., 2017), TFN (Zadeh et al., 2017), CMMT (Yang et al., 2022), VTFSA (Cui et al., 2020), two multimodal analytics methods in live streaming commerce (Xu et al., 2023; Xu et al., 2024b). MFB and TFN capture the interaction between modalities and hidden representations through tensor decomposition. CMMT and VTFSA capture interactions between different modalities based on transformer structure and attention mechanism. Xu et al. (2023) and Xu et al. (2024b) are both the state-of-the-art predictive models in live streaming commerce.

Model	Validation				Test			
	MAE	MSE	MAPE	R^2	MAE	MSE	MAPE	R^2
MFB	0.5050	0.4400	0.0984	0.6467	0.5207	0.4590	0.1013	0.6357
TFN	0.5016	0.4283	0.0965	0.6561	0.5075	0.4376	0.0975	0.6526
CMMT	0.4914	0.4253	0.0964	0.6585	0.4858	0.4165	0.0954	0.6694
VTFSA	0.4593	0.3790	0.0901	0.6957	0.4694	0.3930	0.0917	0.6881
Xu et al. (2023)	0.4577	0.3788	0.0842	0.6510	0.4731	0.4014	0.0872	0.6422
Xu et al. (2024b)	0.4673	0.3663	0.0894	0.6625	0.4779	0.3835	0.0920	0.6581
EMAF	**0.4335**	**0.3574**	**0.0845**	**0.7130**	**0.4472**	**0.3673**	**0.0869**	**0.7084**

Table 5. Results of different predictive methods

As seen in Table 5, compared to other baseline models, our proposed model demonstrates superior performance. Compared to the second-best performing model on the test dataset, our EMAF model reduces MAE, MSE, and MAPE errors by 4.7%, 6.5%, and 5.2% respectively, and enhances the R^2 fitting degree by 3.0%. The main reason is that our model draws inspiration from the ELM, recombining and fusing multimodal information. Specifically, we process affective-related information and cognitive-related information separately to obtain their respective features. There is no interaction between the two features, enabling our framework to more effectively capture the differences between various types of information. Finally, through further integration using DNN, we obtain feature that encompass both diversity and complementarity. This fusion method is more in line with the way human processes information, and therefore stands out among various multimodal fusion models.

Effectiveness evaluation of the MD-Transformer

To evaluate the multimodal MD-Transformer's effectiveness at integrating product details, we compare it with several prevalent deep learning methods, including **Fully Connected layer** (**FC**), **Convolutional Neural Networks** (**CNN**), **Long Short-Term Memory networks** (**LSTM**), and the standard **Transformer**. Table 6 demonstrates that MD-Transformer surpasses all alternative models across all metrics. Compared to the standard Transformer, our model shows a reduction of 6.3% in MAE, 10.7% in MSE, and 6.6% in MAPE. Additionally, it improves the R^2 by 5.2%.

Model	Validation				Test			
	MAE	MSE	MAPE	R^2	MAE	MSE	MAPE	R^2
FC	0.4637	0.3984	0.0911	0.6801	0.4780	0.4225	0.0937	0.6646
CNN	0.4727	0.4163	0.0914	0.6658	0.4821	0.4221	0.0930	0.6649
LSTM	0.4869	0.4299	0.0953	0.6548	0.4929	0.4349	0.0965	0.6548
Transformer	0.4576	0.3855	0.0897	0.6905	0.4774	0.4112	0.0930	0.6736
MD-Transformer	**0.4335**	**0.3574**	**0.0845**	**0.7130**	**0.4472**	**0.3673**	**0.0869**	**0.7084**

Table 6. Results of different fusion methods on product details

The results indicate that although Transformers perform well in handling multimodal data, the standard Transformer structure needs to be further optimized to adapt to this change when dealing with more dimensional feature. MD-Transformer is specifically designed to handle more dimensional feature of product details. It can consider both the relationships between different modalities and the relationships between different products, which can better represent the feature of product details. The effectiveness of MD-Transformer has been demonstrated through the results, providing a novel strategy for Transformer to handle more dimensional feature.

Performance comparisons of different information

To explore the impact of all the different affective-related and cognitive-related information, we design a series of experiments by individually removing a kind of information and compare the results with the complete set of information, as shown in Table 7. The experimental results indicate that both affective-related information and cognitive-related information significantly influence on predicting user comment behavior. Regarding cognitive-related information, product details have always been a focal point of user attention and an important driving force for commenting, indicating that product details should be emphasized when building models to enhance the predictive ability of user engagement behavior. Additionally, for affective-related information, it is surprising that the influence of anchor and the background of the live room have a greater impact compared to visual effects. We infer that the reasons users enter a live streaming are closely tied to the anchor's influence and the background of the live room, which are prerequisites for user engagement behavior.

Moreover, the experimental results also show that cognitive-related information has a more substantial impact than affective-related information. The reason may be that users will carefully think before commenting, which mainly reflects their cognitive route when processing information. These findings provide anchors or merchants with a data-driven, operational priority perspective when analyzing their live streaming data, thus helping them to more accurately understand the behavior patterns and preferences of their audience.

model		Validation				Test			
		MAE	MSE	MAPE	R^2	MAE	MSE	MAPE	R^2
C^T	w/o CP^T	0.4860	0.4103	0.0975	0.6706	0.5024	0.4341	0.1005	0.6555
	w/o CA^T	0.4683	0.3939	0.0911	0.6837	0.4764	0.4051	0.0928	0.6785
	w/o CL^T	0.4824	0.4163	0.0928	0.6658	0.4925	0.4236	0.0942	0.6638
A^T	w/o AI	0.4802	0.4051	0.0955	0.6748	0.4931	0.4157	0.0979	0.6700
	w/o BG	0.4611	0.3774	0.0897	0.6970	0.4811	0.4008	0.0934	0.6819
	w/o AA^T	0.4737	0.4181	0.0905	0.6643	0.4757	0.4160	0.0906	0.6698
	w/o AL^T	0.4661	0.3810	0.0917	0.6941	0.4805	0.3967	0.0942	0.6851
EMAF		**0.4335**	**0.3574**	**0.0845**	**0.7130**	**0.4472**	**0.3673**	**0.0869**	**0.7084**

Table 7. Impact of different information

Ablation study and sensitivity test

Our research model employs a strategy of multimodal fusion for affective-related information and cognitive-related information separately. To validate the efficacy of each component of the model, we conduct a series of ablation experiments, as shown in Table 8. Analysis of Table 8 clearly shows that cognitive-related information has a significantly greater impact on predicting user engagement behavior than affective-related information. This finding further supports our initial inference that users carefully think before commenting, and their in-depth thinking process relies more on cognitive route.

Model	Validation				Test			
	MAE	MSE	MAPE	R^2	MAE	MSE	MAPE	R^2
w/o C^T	0.4989	0.4431	0.0965	0.6442	0.4973	0.4437	0.0962	0.6478
w/o A^T	0.4928	0.4281	0.0956	0.6563	0.4948	0.4236	0.0960	0.6638
EMAF	**0.4335**	**0.3574**	**0.0845**	**0.7130**	**0.4472**	**0.3673**	**0.0869**	**0.7084**

Table 8. Results of ablation experiments

Furthermore, we use DNN to predict user engagement behavior, adjusting the fusion ratio of affective-related and cognitive-related features via the parameter α. During this process, we incrementally increase the value of α from 0 to 1 by 0.01 each time to observe the prediction at different fusion ratios, as shown in Figure 3. In this figure, the blue line represents the mean of the model's predicted results, while the green line indicates the mean of all actual comments in the test dataset. Results from Figure 3 reveal that as more cognitive-related information is integrated, the EMAF shows an upward trend in predicting the number of user comment. This trend also confirms that user comment behavior is largely influenced by the cognitive route, which they utilize when processing information. Therefore, anchors who wish to increase user interaction should enhance elements related to cognitive route in their live streaming content to encourage more user comments.

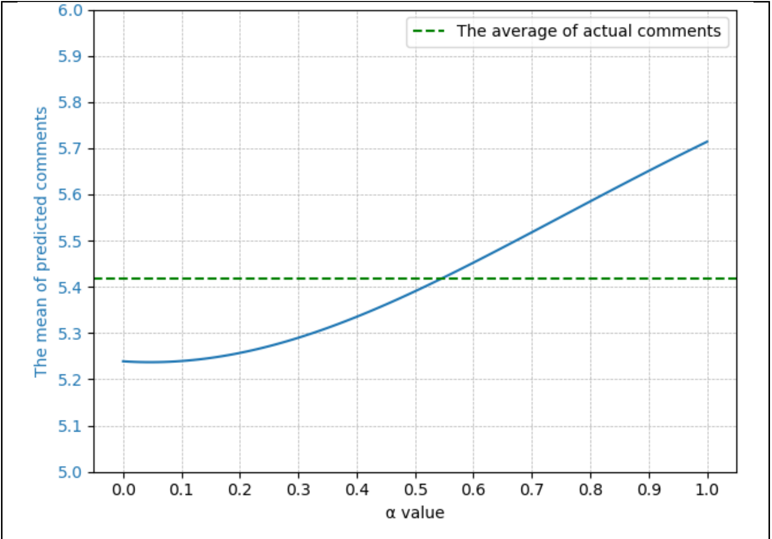

Figure 3. The sensitivity analysis of parameter α

Discussion

In this study, we propose a multimodal analysis framework to predict user engagement behavior in real-time in live streaming commerce. This framework innovatively adopts the ELM theory, integrating affective-related and cognitive-related information from the multimodal data. Table 7 shows that all the information in live streaming influences the prediction of user engagement behavior, with product details having the greatest impact. To better integrate product details, we design a module named MD-transformer to extract their feature. This model considers both the relationships between different modalities and the relationships between different products. The experimental results in Table 6 reflect the effectiveness of the MD-transformer in handling multidimensional features of product details. Furthermore, our ablation experiments and sensitivity test on α reveal that affective-related information in live streaming commerce more easily influences user's comment, which require contemplation.

Methodologically, we innovatively apply this theory to multimodal fusion, predicting user engagement behavior in real-time from the perspectives of affective-related and cognitive-related information in live streaming commerce. Moreover, our research also extends the application scenarios of the transformer model, providing a new approach for handling more dimensional feature.

From a managerial perspective, our research provides significant insights for practitioners in live streaming commerce. The substantial information transmitted through live streaming platforms can help predict user engagement behavior by affective and cognitive routes. Our experimental results demonstrate that multimodal information in live streaming commerce has different effects on the prediction of user engagement. This finding can help live streaming platforms and anchors more accurately adjust content and interaction strategies to maximize user engagement and commercial success. In addition, this study also provides data-driven decision support tools for live streaming commerce, offering an operational priority perspective for analyzing different modalities.

From a practical perspective, our research provides operational recommendations for anchors and platforms in live streaming commerce. First, it is crucial for anchors to deepen the integration of cognitive content with affective interactions. During live streaming, providing detailed product information can meet users' cognitive needs. Concurrently, enhancing the user experience is essential, and anchors can achieve this by developing a unique commentary style and creating an interactive live environment to boost affective engagement. In addition, our research sheds light on the relationship between the number of comments and multimodal information, enabling anchors and platforms to adjust content in time to enhance user engagement behavior. For example, when anchors want to increase interaction with users, they should strengthen the information elements related to cognitive route in the content to promote more user's comment behavior.

Conclusions

This study proposes a multimodal analysis framework to predict user engagement behavior in real-time in live streaming commerce. Based on a real-world dataset from Taobao Live, our model achieves better results compared to existing multimodal fusion models. This is attributed to the recombination and fusion of multimodal information that affects user affective and cognitive routes. Because multimodal information from different routes can more fully consider the relationships between information on the same route, thereby better representing its features. A key innovation of our framework is not just the multimodal fusion framework, but also handling the differences in multimodal information from emotional and cognitive perspectives based on ELM theory, which extends its applications.

This research acknowledges certain limitations. First, while video information is used in our model, we haven't fully explored aspects like the anchor's gestures and facial expressions, which could offer additional insights into the exploration of user engagement behavior. Second, although we have verified that different information may have different impacts on predicting user engagement behavior, the interpretability of the proposed framework is somewhat limited, because it cannot adequately explain how the information influences the prediction. Third, our examination of user engagement has primarily relied on user comments, with less emphasis on a broader spectrum of user engagement behaviors such as watching and liking. Consequently, future research will focus on improving the model's interpretability and deeply exploring the specific roles of different information. We also aim to explore more user engagement behaviors with more refined information and extending the applicability of the model to a wider range of scenarios, to uncover more valuable insights in the field.

References

Addo, P. C., Fang, J., Asare, A. O., & Kulbo, N. B. (2021). Customer engagement and purchase intention in live-streaming digital marketing platforms. The Service Industries Journal, 41(11-12), 767-786.

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423-443.

Bhattacherjee, A., & Sanford, C. (2006). Influence processes for information technology acceptance: An elaboration likelihood model. MIS quarterly, 805-825.

Chen, J., & Liu, H. (2024). Modeling interaction behavior and preference decline for live stream recommendation. Decision Support Systems, 179, 114146.

Chen, X., Ji, L., Jiang, L., & Huang, J. T. (2023). The bright side of emotional extremity: evidence from tipping in live streaming platform. Information & Management, 60(1), 103726.

Chou, Y. C., Chuang, H. H. C., & Liang, T. P. (2022). Elaboration likelihood model, endogenous quality indicators, and online review helpfulness. Decision Support Systems, 153, 113683.

Cui, S., Wang, R., Wei, J., Hu, J., & Wang, S. (2020). Self-attention based visual-tactile fusion learning for predicting grasp outcomes. IEEE Robotics and Automation Letters, 5(4), 5827-5834.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013, October). Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (pp. 835-838).

Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y., & Zhu, Y. (2024). A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. Expert Systems with Applications, 242, 122731.

Gao, X., Xu, X. Y., Tayyab, S. M. U., & Li, Q. (2021). How the live streaming commerce viewers process the persuasive message: An ELM perspective and the moderating effect of mindfulness. Electronic Commerce Research and Applications, 49, 101087.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Huang, L., & Ma, L. (2024). A protective buffer or a double-edged sword? Investigating the effect of "parasocial guanxi" on consumers' complaint intention in live streaming commerce. Computers in Human Behavior, 151, 108022.

Kang, K., Lu, J., Guo, L., & Li, W. (2021). The dynamic effect of interactivity on customer engagement behavior through tie strength: Evidence from live streaming commerce platforms. International Journal of Information Management, 56, 102251.

Li, L., Chen, X., & Zhu, P. (2024). How do e-commerce anchors' characteristics influence consumers' impulse buying? An emotional contagion perspective. Journal of Retailing and Consumer Services, 76, 103587.

Lin, Q., Jia, N., Chen, L., Zhong, S., Yang, Y., & Gao, T. (2023). A two-stage prediction model based on behavior mining in livestream e-commerce. Decision Support Systems, 174, 114013.

Lin, Y., Yao, D., & Chen, X. (2021). Happiness begets money: Emotion and engagement in live streaming. Journal of Marketing Research, 58(3), 417-438.

Liu, Z., Li, J., Wang, X., & Guo, Y. (2023). How search and evaluation cues influence consumers' continuous watching and purchase intentions: An investigation of live-stream shopping from an information foraging perspective. Journal of Business Research, 168, 114233.

Lo, P. S., Dwivedi, Y. K., Tan, G. W. H., Ooi, K. B., Aw, E. C. X., & Metri, B. (2022). Why do consumers buy impulsively during live streaming? A deep learning-based dual-stage SEM-ANN analysis. Journal of Business Research, 147, 325-337.

Luo, L., Xu, M., & Zheng, Y. (2024a). Informative or affective? Exploring the effects of streamers' topic types on user engagement in live streaming commerce. Journal of Retailing and Consumer Services, 79, 103799.

Luo, X., Cheah, J. H., Hollebeek, L. D., & Lim, X. J. (2024b). Boosting customers' impulsive buying tendency in live-streaming commerce: The role of customer engagement and deal proneness. Journal of Retailing and Consumer Services, 77, 103644.

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. Journal of consumer research, 10(2), 135-146.

Rahmani, S., Hosseini, S., Zall, R., Kangavari, M. R., Kamran, S., & Hua, W. (2023). Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects. Knowledge-Based Systems, 261, 110219.

Rynkiewicz, J. (2019). Asymptotic statistics for multilayer perceptron with ReLU hidden units. Neurocomputing, 342, 16-23

Shin, H., Oh, C., Kim, N. Y., Choi, H., Kim, B., & Ji, Y. G. (2023). Evaluating and eliciting design requirements for an improved user experience in live-streaming commerce interfaces. Computers in Human Behavior, 107990.

Wang, H., Li, G., Xie, X., & Wu, S. (2024). An empirical analysis of the impacts of live chat social interactions in live streaming commerce: A topic modeling approach. Electronic Commerce Research and Applications, 101397.

Wongkitrungrueng, A., & Assarut, N. (2020). The role of live streaming in building consumer trust and engagement with social commerce sellers. Journal of business research, 117, 543-556.

Xiao, L., Lin, X., Mi, C., & Akter, S. (2023). The effect of dynamic information cues on sales performance in live streaming e-commerce: an IFT and ELM perspective. Electronic Commerce Research, 1-30.

Xin, B., Hao, Y., & Xie, L. (2023). Strategic product showcasing mode of E-commerce live streaming. Journal of Retailing and Consumer Services, 73, 103360.

Xin, M., Liu, W., & Jian, L. (2024). Live streaming product display or social interaction: How do they influence consumer intention and behavior? A heuristic-systematic perspective. Electronic Commerce Research and Applications, 67, 101437.

Xu, G., Ren, M., Wang, Z., & Li, G. (2024a). MEMF: Multi-entity multimodal fusion framework for sales prediction in live streaming commerce. Decision Support Systems, 184, 114277.

Xu, W., Cao, Y., & Chen, R. (2024b). A multimodal analytics framework for product sales prediction with the reputation of anchors in live streaming e-commerce. Decision Support Systems, 177, 114104.

Xu, W., Zhang, X., Chen, R., & Yang, Z. (2023). How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce. Decision Support Systems, 113984.

Xue, J., Liang, X., Xie, T., & Wang, H. (2020). See now, act now: How to interact with customers to enhance social commerce engagement? Information & Management, 57(6), 103324.

Yan, Y., Chen, H., Shao, B., & Lei, Y. (2023). How IT affordances influence customer engagement in live streaming commerce? A dual-stage analysis of PLS-SEM and fsQCA. Journal of Retailing and Consumer Services, 74, 103390.

Yang, J., Huang, Q., Ding, T., Lischinski, D., Cohen-Or, D., & Huang, H. (2023). EmoSet: A large-scale visual emotion dataset with rich attributes. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20383-20394).

Yang, L., Na, J. C., & Yu, J. (2022). Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. Information Processing & Management, 59(5), 103038.

Yang, S., Zhou, C., & Chen, Y. (2021). Do topic consistency and linguistic style similarity affect online review helpfulness? An elaboration likelihood model perspective. Information Processing & Management, 58(3), 102521.

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In Proceedings of the IEEE international conference on computer vision (pp. 1821-1830).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Zhang, M., Liu, Y., Wang, Y., & Zhao, L. (2022). How to retain customers: Understanding the role of trust in live streaming commerce with a socio-technical perspective. Computers in Human Behavior, 127, 107052.

Zhang, Y., & Xu, Q. (2024). Consumer engagement in live streaming commerce: Value co-creation and incentive mechanisms. Journal of Retailing and Consumer Services, 81, 103987.

Zheng, R., Li, Z., & Na, S. (2022). How customer engagement in the live-streaming affects purchase intention and customer acquisition, E-tailer's perspective. Journal of Retailing and Consumer Services, 68, 103015.

Zhu, Z., Yang, Z., & Dai, Y. (2017). Understanding the gift-sending interaction on live-streaming video websites. In Social Computing and Social Media. Human Behavior: 9th International Conference, SCSM 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 9 (pp. 274-285). Springer International Publishing.