Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36245

Research on Emotional Feature Recognition of Streamer Discourse and Conversion Effect Prediction in Live Streaming E-commerce

Yixuan Li

Yonyou Digital Intelligence College, Nantong Institute of Technology, Nantong, China
liyixuan02418@outlook.com

Abstract. This study develops a window-level framework for recognizing emotional features in streamer discourse and predicting conversion effects in live streaming e-commerce. The research examines how emotional intensity, persuasive tone, affective fluctuation, discourse rhythm, and acoustic arousal are associated with product click, add-to-cart behavior, and completed purchase. The dataset consists of 120 live-streaming sessions collected from the beauty and personal-care category over 60 days. It includes live audio, speech transcripts, product exposure records, interaction logs, and order-conversion logs. All records are synchronized into 30-second discourse windows and matched with 3-minute post-exposure conversion outcomes. Whisper-large-v3 is used for transcription, BERT-Base Chinese and an emotion lexicon are used for textual emotion extraction, openSMILE 3.0 is used for acoustic features, and XGBoost 2.0 is used for multi-label conversion prediction. The results show that multimodal emotional discourse features outperform single-modality models, with purchase-conversion AUC reaching 0.873±0.012* and macro-F1 reaching 0.782±0.016*.

Keywords: Live streaming e-commerce, streamer discourse, emotional feature recognition, conversion prediction, multimodal analysis

1. Introduction

Live streaming e-commerce has become a highly interactive retail environment in which product explanation, emotional performance, real-time interaction, and purchase guidance are compressed into short promotional moments. Unlike conventional e-commerce pages, live commerce depends on the streamer's ability to convert attention into action through spoken discourse. Streamers introduce product functions, compare prices, answer questions, create urgency, reinforce trust, and repeatedly guide viewers toward clicking, adding products to cart, or completing payment. Prior research has shown that online celebrities in live-streaming commerce influence purchase intention through emotional contagion [1]. Customer engagement in live-streaming settings is also associated with purchase intention and customer acquisition from the e-tailer's perspective [2]. Streamer-related research further identifies streamers as product demonstrators, emotional persuaders, trust builders, and interactive coordinators [3]. Platform affordance studies show that visibility, interactivity, and social presence shape engagement in live-streaming commerce [4]. More recent multimodal analytics work has begun to predict sales outcomes from anchor reputation and multimodal behavioral signals [5]. Research on real-time interaction and sentiment also indicates that live-streaming comments and affective communication can influence online sales [6]. At the method level, speech emotion recognition and multimodal sentiment analysis provide tools for extracting emotional cues from language and acoustic signals [7, 8]. Cross-modal prediction research in live-streaming e-commerce further supports the use of audio, visual, and linguistic features for business outcome prediction [9]. BERT-based text classification studies also show that live-commerce textual content can be transformed into structured predictive features [10]. Based on these developments, this study constructs a synchronized discourse-window framework to predict interaction, cart, and purchase conversion from streamer emotional discourse.

2. Literature review

2.1. Streamer discourse and emotional persuasion in live commerce

Streamer discourse in live commerce combines product description, emotional encouragement, scarcity prompting, trust signaling, and immediate response to audience interaction. It differs from ordinary advertising because persuasion occurs continuously and is adjusted in real time. Emotional persuasion appears through positive evaluation, urgency words, intensified adjectives, reassurance statements, repeated recommendations, and rhythmic calls to action. In short product-promotion windows, the streamer may rapidly shift from factual explanation to affective stimulation. This makes streamer discourse a dynamic emotional signal rather than a stable speaking style. Existing live-commerce research indicates that streamer emotion, credibility, interaction, and role performance can shape viewer engagement and purchase response [1-4].

2.2. Emotional feature recognition in speech and commerce communication

Emotional feature recognition in speech-driven commerce requires both textual and acoustic analysis. Textual emotion includes sentiment polarity, affective intensity, evaluative language, trust-building expressions, urgency cues, and repeated emotional wording. Acoustic emotion includes pitch variation, loudness, speaking rate, pause ratio, jitter, shimmer, and energy fluctuation. Speech emotion recognition research shows that vocal patterns can reveal arousal and affective states under complex noise conditions [7]. Multimodal sentiment analysis further demonstrates that integrating text and audio can improve affective recognition compared with relying on one modality alone [8]. In live commerce, this is necessary because the same textual message may have different persuasive effects when spoken with different rhythm, pitch, and intensity.

2.3. Conversion effect prediction in live streaming environments

Conversion effect in live streaming is reflected in product click, coupon claim, add-to-cart action, and completed purchase after exposure to a promotion segment. Accurate prediction requires temporal alignment among discourse windows, product-display periods, and behavioral response logs. Recent multimodal studies have shown that streamer behavior can be used to predict downstream commercial outcomes in live-streaming e-commerce [5-9]. Real-time sentiment and comment-based research also suggests that conversion is shaped by affective and interactive signals during live sessions [6]. Language-model-based classification studies further support the extraction

DOI: 10.54254/2755-2721/2026.36245

of structured information from live-commerce text [10]. Therefore, a window-level prediction framework is suitable for linking emotional discourse features with short-term conversion outcomes.

3. Research methodology

3.1. Live streaming dataset construction and observation unit definition

The dataset is constructed from 120 live-streaming sessions collected from one fixed e-commerce live platform over 60 days. The sample includes 15 streamers from the beauty and personal-care category because this product category relies strongly on explanation, demonstration, trust building, and emotional persuasion. Each live session contains synchronized audio recordings, product display timelines, viewer interaction logs, and order-conversion records. All raw records are stored in MySQL 8.0 and indexed by session_id, streamer_id, product_id, timestamp, discourse_window_id, and conversion_window_id. This relational structure allows discourse features, product exposure, and behavioral conversion to be joined without changing the original time order.

The observation unit is defined as one fixed 30-second discourse window matched with one 3-minute post-exposure conversion window. The 30-second window is used because it is long enough to contain a complete persuasive expression but short enough to preserve real-time discourse variation. The 3-minute response window is used to capture immediate consumer actions after product exposure. If one discourse window overlaps with multiple product displays, the product occupying the longest exposure time is retained. Windows with less than 70% valid product exposure are removed. After deleting incomplete, silent, invalid, and low-traffic records, the final dataset contains 18,000 discourse-conversion windows. Each observation contains streamer speech, text transcript, acoustic features, discourse structure, product information, audience size, and three conversion labels.

3.2. Emotional feature extraction and variable construction

Speech transcription is completed with Whisper-large-v3 in Python 3.11. The output transcript is manually checked at the punctuation and sentence-boundary level to reduce automatic segmentation errors. Textual emotional features are extracted using BERT-Base Chinese and a fixed emotion lexicon. These features include emotional polarity score, affective intensity score, urgency-expression frequency, trust-expression density, positive evaluative density, emotional repetition rate, and call-to-action frequency. Acoustic emotional features are extracted from the original audio using openSMILE 3.0 with the eGeMAPS configuration. The acoustic variables include pitch mean, pitch variance, intensity mean, speaking rate, jitter, shimmer, pause ratio, and energy fluctuation. Discourse-structure variables include sentence-length variance, imperative-sentence ratio, interrogative prompting ratio, and exclamation density.

The discourse-window feature vector is defined as shown in Equation (1):

$$X_{s,p,w} = [EPol_w, AInt_w, Urg_w, Trust_w, Eval_w, Rep_w, \ \frac{1}{T}\sum\nolimits_{t=1}^{T} F0_t, \tag{1}$$

$$Var(F0_t), \frac{WordCount_w}{Duration_w}, PauseRatio_w, CTA_w, Control_{s,p,w}]$$

DOI: 10.54254/2755-2721/2026.36245

where $X_{s,p,w}$ denotes the emotional discourse feature vector for streamer s, product p, and discourse window w. $EPol_w$ is emotional polarity, $AInt_w$ is affective intensity, Urg_w is urgency frequency, Trust_w is trust-expression density, $Eval_w$ is positive evaluative density, Rep_w is emotional repetition, $F0_t$ is pitch, and CTA_w is call-to-action frequency. Control variables include live audience size, product discount rate, product price band, product category sub-type, and session stage.

3.3. Prediction model and interpretation configuration

The conversion prediction model is implemented in XGBoost 2.0 using a fixed multi-label classification structure. The three target labels are interaction conversion, cart conversion, and purchase conversion. Interaction conversion is coded as 1 when the discourse window triggers product click behavior in the following 3-minute response window. Cart conversion is coded as 1 when an add-to-cart action occurs. Purchase conversion is coded as 1 when a completed order is recorded. Continuous variables are standardized with StandardScaler, and categorical variables are one-hot encoded in scikit-learn. The hyperparameters are fixed as max_depth = 6, learning_rate = 0.05, n_estimators = 300, subsample = 0.8, colsample_bytree = 0.8, and random_state = 42.

The prediction objective is specified as shown in Equation (2):

$$\mathscr{L} = -\sum_{i=1}^{N} \sum_{k=1}^{3} [y_{ik} \log(\widehat{p}_{ik}) + (1 - yik) \log(1 - \widehat{p}_{ik})] \tag{2}$$

$$+ \sum m = 1^{M} [\gamma K_m + \tfrac{1}{2} \lambda \sum_{l=1}^{K_m} v_{ml}^2]$$

where y_{ik} is the true label of conversion target k, $\widehat{p}_{ik}$ is the predicted probability, K_m is the number of leaves in tree m, vml is the leaf weight, and γ and λ control model complexity. SHAP TreeExplainer is used to calculate global feature importance, target-specific feature contribution, and discourse-window-level local explanations. The output table stores feature name, mean absolute SHAP value, class-level contribution, and local SHAP value.

4. Experimental design

4.1. Data cleaning, timeline alignment, and sample splitting

The raw live-stream records are cleaned before feature construction. Duplicate timestamps, invalid product IDs, inactive product exposures, silent audio intervals longer than 10 seconds, and conversion logs unrelated to the displayed product are removed. Audio is segmented into fixed 30-second windows according to the live-session timeline. Each window is then matched with the product promoted during that interval. If product exposure changes within the window, the product with the longer exposure duration is retained. Windows with ambiguous product exposure are excluded to ensure that each discourse segment corresponds to one product.

The subsequent 3-minute behavioral logs are linked to the same product ID and discourse window. Product clicks, add-to-cart actions, and completed purchases are counted only when they occur after discourse exposure and before the response window closes. Textual, acoustic, discourse-structure, product, and session-control variables are merged by session_id, product_id, and discourse_window_id. The dataset is split into 70% training set, 15% validation set, and 15% test set

DOI: 10.54254/2755-2721/2026.36245

by session rather than by individual windows. This prevents windows from the same live session from appearing across different subsets. All preprocessing parameters, including standardization and encoding rules, are fitted on the training set and applied unchanged to validation and test sets.

4.2. Model training, validation, and prediction output generation

The XGBoost model is trained separately for interaction, cart, and purchase conversion using the same feature system and hyperparameter settings. The training set supports parameter learning, while the validation set monitors AUC, macro-F1, and log loss. Early stopping is applied when validation log loss does not improve for 25 rounds, and the final models are frozen before test evaluation. Each discourse window receives three conversion probabilities. Binary labels are assigned at a 0.50 threshold, while probabilities between 0.45 and 0.55 are marked as low-confidence for later interpretation and diagnostic analysis.

4.3. Emotional pattern analysis and streamer-level comparative output

Emotional pattern analysis is conducted after prediction evaluation. First, SHAP values are calculated on the full test set to identify the most influential emotional discourse features for interaction, cart, and purchase conversion. The top 15 features are ranked by mean absolute SHAP value. Second, local SHAP explanations are generated for high-conversion and low-conversion windows. This comparison shows whether successful discourse windows depend more on affective wording, urgency language, trust expression, pitch fluctuation, speaking speed, or call-to-action frequency. Third, textual and acoustic features are interpreted both separately and jointly to evaluate whether conversion is driven by wording alone or by combined verbal and vocal expression.

Streamer-level comparison is conducted by aggregating emotional discourse profiles across all sessions for each streamer. The aggregated indicators include average emotional intensity, urgency-expression density, trust-expression density, pitch fluctuation, speaking rate, pause ratio, exclamation density, and purchase-conversion probability. Streamers are then classified into high-conversion and low-conversion groups according to the median purchase-conversion probability. Final statistical summaries are generated in R 4.4.

5. Experimental results

5.1. Descriptive statistics of live streaming discourse windows, emotional features, and conversion labels

The final dataset contains 18,000 valid discourse-conversion windows. The average emotional polarity score is 0.624±0.142, affective intensity is 0.571±0.168, and urgency-expression frequency is 2.416±1.305 occurrences per window. Acoustic indicators also show clear dispersion, with pitch variance of 34.782±11.964 and intensity mean of 67.436±5.218 dB. High-conversion windows show stronger trust-expression density, higher call-to-action frequency, and lower pause ratio than low-conversion windows (see Table 1).

Table 1. Descriptive statistics of emotional discourse features and conversion labels

Variable	Mean±SD	P25	Median	P75	Low-conversion mean	High-conversion mean
Emotional polarity	0.624±0.142	0.518	0.631	0.728	0.587±0.136	0.691±0.129*
Affective intensity	0.571±0.168	0.442	0.566	0.704	0.523±0.159	0.653±0.151*

DOI: 10.54254/2755-2721/2026.36245

Table 1. (continued)

Urgency frequency	2.416±1.305	1	2	3	2.103±1.226	3.047±1.271*
Trust-expression density	0.184±0.073	0.126	0.176	0.231	0.161±0.066	0.224±0.071*
Pitch variance	34.782±11.964	26.418	33.607	41.925	31.906±10.842	39.684±12.117*
Speaking rate	5.318±1.026	4.572	5.284	6.013	5.012±0.984	5.871±1.004*
Pause ratio	0.142±0.061	0.096	0.133	0.181	0.157±0.063	0.119±0.052*
Purchase-conversion label	0.216±0.000*					

5.2. Prediction performance of interaction, cart, and purchase conversion under the emotional discourse model

The multimodal emotional model achieves stable prediction results across three conversion targets. Interaction conversion performs best, with AUC of 0.901±0.010* and macro-F1 of 0.824±0.013*. Cart conversion reaches AUC of 0.884±0.011* and macro-F1 of 0.801±0.015*. Purchase conversion is more difficult but remains predictable, with AUC of 0.873±0.012* and macro-F1 of 0.782±0.016*. Multimodal emotional features outperform text-only, acoustic-only, and discourse-structure models (see Figure 1).

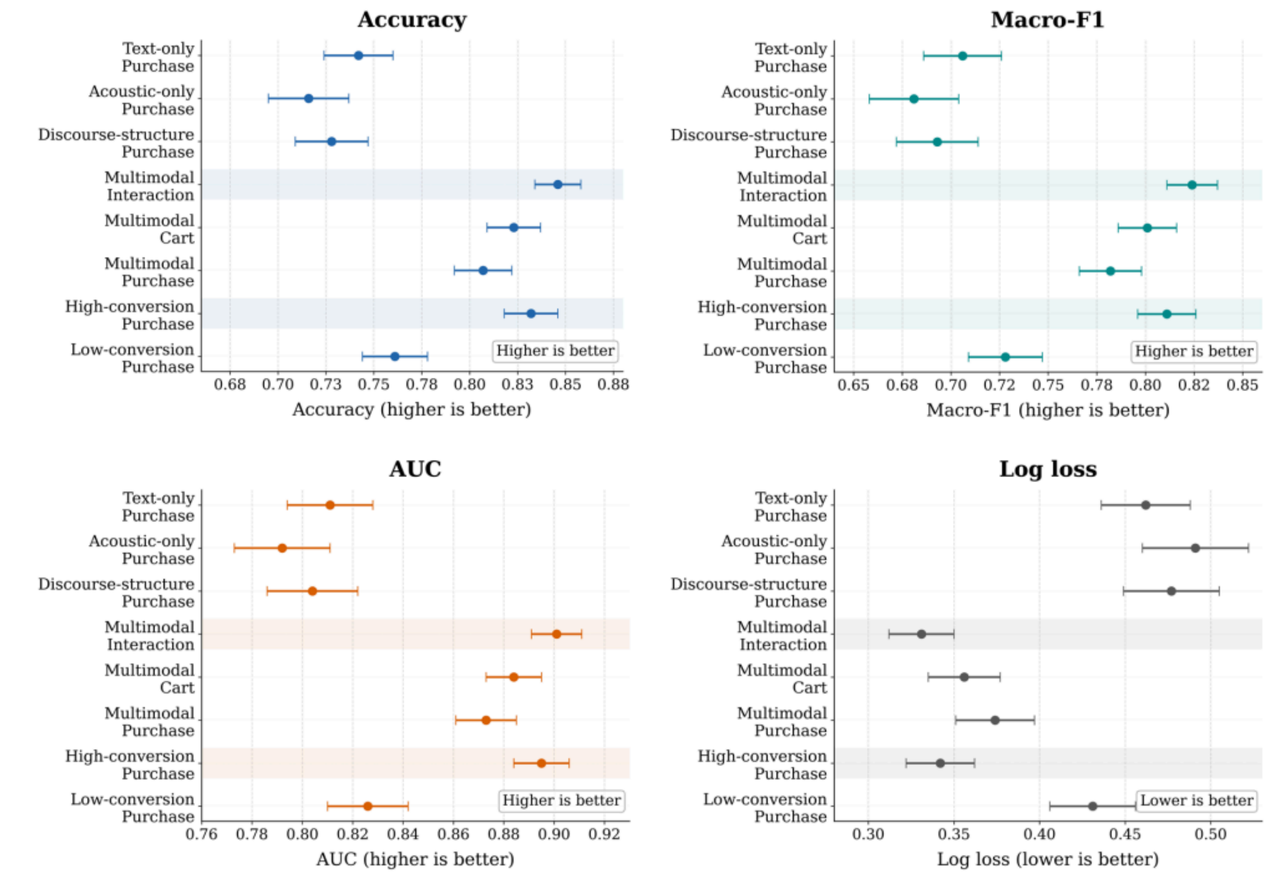

Figure 1. Prediction performance and modality comparison

DOI: 10.54254/2755-2721/2026.36245

5.3. SHAP-based emotional feature attribution and streamer-level conversion pattern differences

SHAP results show that affective intensity has the highest mean absolute contribution for interaction conversion, reaching 0.137±0.019*. Urgency-expression frequency contributes more strongly to cart conversion, with 0.124±0.017*. Purchase conversion is most strongly associated with trust-expression density, positive evaluative density, and lower pause ratio, with SHAP values of 0.142±0.021*, 0.118±0.016*, and 0.097±0.014*.

6. Conclusion

This study constructs a window-level framework for emotional feature recognition of streamer discourse and conversion effect prediction in live streaming e-commerce. By integrating audio, transcript text, product exposure timelines, interaction logs, and order-conversion records, the study models how streamer emotional discourse is associated with product click, add-to-cart behavior, and completed purchase. The results show that multimodal emotional features outperform single-modality models, especially in purchase-conversion prediction. SHAP analysis further indicates that affective intensity, urgency frequency, trust-expression density, call-to-action frequency, pitch variance, and pause ratio are important predictors. The findings suggest that live-commerce conversion is shaped by coordinated emotional discourse patterns rather than product information alone.

References

[1] Wongkitrungrueng, A., Dehouche, N., & Assarut, N. (2020). Live streaming commerce from the sellers' perspective: Implications for online relationship marketing. *Journal of Marketing Management, 36*(5–6), 488–518.
[2] Sokolova, K., & Kefi, H. (2020). Instagram and YouTube bloggers promote it, why should I buy? How credibility and parasocial interaction influence purchase intentions. *Journal of Retailing and Consumer Services, 53*, 101742.
[3] Hilvert-Bruce, Z., Neill, J. T., Sjöblom, M., & Hamari, J. (2021). Social motivations of live-streaming viewer engagement on Twitch. *Computers in Human Behavior, 84*, 58–67.
[4] 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.
[5] Ma, Y., Sun, S., & Kekre, S. (2022). Live streaming commerce: The impact of streamer characteristics and consumer trust on purchase intention. *Electronic Commerce Research and Applications, 54*, 101174.
[6] Wulfert, T., Schmitt, M., & Leimeister, J. M. (2023). Understanding customer engagement in livestream shopping: The role of social presence and interaction quality. *Electronic Markets, 33*(1), 1–18.
[7] George, S. M., & Ilyas, P. M. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing, 568*, 127015.
[8] Poria, S., Majumder, N., Hazarika, D., Mihalcea, R., & Gelbukh, A. (2023). Multimodal sentiment analysis: State of the art and future directions. *Information Fusion, 91*, 424–444.
[9] Chicaiza, J., Corrales, J. C., & Luján-Mora, S. (2024). Machine learning approaches for e-commerce conversion prediction: A systematic literature review. *Expert Systems with Applications, 237*, 121465.
[10] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2023). Text classification algorithms: A survey. *Information, 14*(1), 43.