How Social Media Information Characteristics Shape Destination Choice Intention: An LDA–SEM–XGBoost Framework

Hao Su[1], Yaping Zhang[2], Qi Pan[3], Yuqing Du[4]
[1]School of Management, Anshan Normal University, Anshan 114056, Liaoning, China
[2]Graduate School of Business, Youngsan University, Busan 48015, Korea
[3]College of Applied Technology, Anshan Normal University, Anshan 114056, Liaoning, China
[4]Human Resources Department, Anshan Normal University, Anshan 114056, Liaoning, China
Corresponding author: Yuqing Du (e-mail: duyuqing2027@163.com).

ABSTRACT The widespread use of social media has reshaped the production, dissemination, and consumption of tourism information. Tourists' destination selection is jointly influenced by platform content characteristics, algorithmic recommendation, emotional expression, and information credibility. This study constructs a three-stage analytical framework of "information feature quantification -behavior mechanism modeling-destination prediction feedback". LDA topic modeling and VADER sentiment analysis are used to extract tourism content features from Weibo, Douyin, and Xiaohongshu, while a structural equation model is applied to examine the paths and influence intensities of information credibility, communication breadth, and emotional tendency on destination selection intention. An XGBoost classifier integrating multi-source social media features is then employed as a diagnostic tool to rank feature importance and assess the relative contribution of information characteristics to destination discrimination within the sampled set. The results show that emotional tendency and information credibility have the strongest positive driving effects, and multi-feature fusion improves prediction accuracy. Because social media diffusion depends on mobile wireless networks, antenna-supported terminals, and electromagnetic information propagation, communication speed and platform reach are interpreted from an engineering-information perspective. The findings provide methodological support for tourism marketing prioritization and feature-driven strategy design.

INDEX TERMS social media communication, tourism behavior, destination prediction, wireless information propagation, electromagnetic media networks

I. INTRODUCTION

THE iterative evolution of digital technology has gradually made social media the core channel for the circulation of tourism information. Platforms such as Weibo, Douyin and Xiaohongshu generate massive amounts of tourism-related images, texts, videos and comments every day, forming a widely covered information dissemination network. Compared with traditional tourism marketing media, social media has remarkable characteristics such as strong interactivity, fast dissemination speed and high user stickiness, enabling tourists to obtain multi-dimensional destination information during the pre-travel decision-making stage [1]. However, problems such as information overload, doubtful content authenticity and biased platform algorithm recommendations have emerged simultaneously, which exert complex interventions on tourists' behavioral decision-making and also bring noise interference to the objective evaluation of destination attractiveness [2]. Existing studies mostly focus on a single platform or a single method, lacking a systematic analysis chain from information feature extraction to behavior modeling and then to destination prediction. Therefore, constructing a quantifiable driving mechanism model and establishing a practical feature-evaluation framework are of great significance for optimizing new media marketing strategies and improving the scientificity of tourism destination management decisions.

II. QUANTITATIVE CHARACTERIZATION AND STRUCTURAL REPRESENTATION OF TOURISM INFORMATION ON SOCIAL MEDIA

A. CLASSIFICATION OF TOURISM INFORMATION TYPES AND MULTI-PLATFORM DATA COLLECTION SCHEME

Tourism social media information is divided into three categories according to the content producer and communication purpose: destination promotion content released by official institutions, experiential images, texts and short videos produced by KOLs, and comments and travel notes spontaneously generated by ordinary users (UGC). The three types of information differ systematically in credibility basis, emotional expression intensity and communi-

TABLE 1. Classification and Collection Scheme of Tourism Social Media Information

Collection Period	Platforms	Destinations Covered	Original Posts	After Deduplication
Jan. 1, 2023–Dec. 31, 2023	Weibo, Douyin, Xiaohongshu	20 (Zhangjiajie, Guilin, Jiuzhaigou, Huangshan, Lijiang, Sanya, Chengdu, Xi'an, Hangzhou, Xiamen, Dali, Lhasa, Qingdao, Dalian, Suzhou, Yangshuo, Wuzhen, Fenghuang, Dunhuang, Harbin)	186,420	124,680

cation path, so classified collection and annotation are required.

Data collection covers three major platforms: Weibo, Douyin and Xiaohongshu. The collection period spans from January 1, 2023 to December 31, 2023, with a total of 12 months. Twenty destinations are covered, selected based on the 2023 China Tourism Academy's annual popularity ranking to ensure representativeness across natural scenery, cultural heritage, and urban leisure categories [3]. The specific destinations are: Zhangjiajie, Guilin, Jiuzhaigou, Huangshan, Lijiang, Sanya, Chengdu, Xi'an, Hangzhou, Xiamen, Dali, Lhasa, Qingdao, Dalian, Suzhou, Guilin (Yangshuo), Wuzhen, Fenghuang Ancient Town, Dunhuang, and Harbin.

With the keyword combination of "scenic spot name + tourism/guide/check-in", structured fields such as post text, release time, likes, reposts and comments are collected through open platform APIs and Selenium crawler tools. The original dataset contains 186,420 posts. After applying exclusion criteria—removing duplicate posts (based on MD5 hash of text content), advertisements identified by commercial keywords (e.g., "discount," "booking link," "limited offer"), posts with fewer than 10 characters, and non-tourism-related content filtered by a keyword classifier—the cleaned dataset comprises 124,680 posts. Table 1 reports the distribution across platforms and content types.

To ensure the ecological validity of the study and capture the peak characteristics of tourism behavior, the collection period encompasses both holiday cycles (Spring Festival, May Day, National Day) and non-holiday working days. Stratified sampling is conducted on each platform, treating the "holiday effect" as a critical temporal dimension. For each destination, posts are sampled proportionally from holiday and non-holiday periods at a 3:7 ratio, reflecting the natural seasonality of tourism content production. This ensures the model accounts for the inherent cyclicality of destination selection.

The final post-level dataset is aggregated to the destination level for SEM analysis. Specifically, for each destination, the mean VADER sentiment score, mean comprehensive diffusion index, and proportion of posts from each content type (official/KOL/UGC) are computed. These destination-level aggregates serve as the observed indicators for the information characteristic latent variables in the SEM. A destination-level database containing multi-platform heterogeneous fields is constructed, and missing value filling and deduplication are completed before entering the subsequent analysis process, as shown in Table 1.

1) Questionnaire Survey for Destination Selection Intention

Destination selection intention (DI) is measured through a structured questionnaire survey administered independently from the social media data collection. The survey was conducted from February 15, 2024 to March 15, 2024 via Wenjuanxing, a professional Chinese online survey platform. A total of 820 questionnaires were distributed, and 487 valid responses were returned after excluding incomplete submissions (response time < 120 seconds) and straight-lining patterns, yielding a valid response rate of 59.4%.

Respondents were recruited using quota sampling to match the demographic profile of Chinese social media users: 52.3% female, 47.7% male; age distribution: 18–30 years (38.2%), 31–45 years (41.5%), 46–60 years (16.4%), over 60 years (3.9%); education: bachelor's degree or above (67.1%); monthly income: below 5,000 CNY (28.1%), 5,000–10,000 CNY (35.3%), above 10,000 CNY (36.6%). Respondents were asked to rate their selection intention for each of the 20 destinations using a 7-point Likert scale (1 = "definitely would not choose" to 7 = "definitely would choose"). The mean intention score across all respondents for each destination is computed and matched to the corresponding destination-level social media aggregate, ensuring consistent analysis units in the SEM.

Crucially, the SEM in this study operates at the destination level ($N = 20$), not the individual level. The independent variables are destination-level aggregates of social media content characteristics, and the dependent variable is the mean destination selection intention aggregated from individual survey responses. This two-step aggregation approach—first aggregating posts to destinations and then aggregating individual intentions to destinations—ensures that all variables in the SEM share the same analysis unit, thereby avoiding the ecological fallacy that would arise from mixing post-level content metrics with individual-level psychological states. The aggregation logic follows the established practice in destination marketing research where destination-level image perceptions are inferred from aggregate social media signals and linked to aggregate behavioral intentions.

B. VARIABLE SETTING AND PATH DESIGN OF STRUCTURAL EQUATION

The original tourism text has high-dimensional sparsity, and direct modeling will introduce a lot of noise. Therefore, the LDA (Latent Dirichlet Allocation) model [4] is used for dimensionality reduction and structured representation. The model treats documents as a mixed distribution of multiple latent topics, and each topic corresponds to a probability distribution on the vocabulary table. The generation probability is:

$$p(w_i \mid d) = \sum_{k=1}^{K} p(w_i \mid z_k) \cdot p(z_k \mid d), \quad (1)$$

where w_i is the i-th word, z_k is the k-th latent topic, K is the total number of topics, $p(z_k \mid d)$ is the topic distribution of the document, and $p(w_i \mid z_k)$ is the word distribution of the topic.

In the text preprocessing stage, jieba segmentation is used to remove stop words and low-frequency words. The number of topics K is determined by joint evaluation of perplexity and topic consistency. Experimental results show that the interpretation effect is optimal when $K = 8$. The extracted topics cover core decision-making dimensions such as natural landscape, food experience, accommodation conditions and transportation convenience, providing structured input for the construction of subsequent information feature variables [4], as shown in Figure 1.

C. MULTI-DIMENSIONAL MEASUREMENT INDEX SYSTEM OF INFORMATION COMMUNICATION BREADTH, SPEED AND EMOTIONAL TENDENCY

Communication breadth is synthesized into a comprehensive diffusion index by weighting the number of reposts, comments and likes of posts, and the weight is determined by principal component analysis, reflecting the actual coverage scale of information within the platform [5]. Communication speed is measured by the interaction growth rate within 48 hours after post release, capturing the kinetic energy differences in the initial stage of information diffusion. This time window is determined based on empirical data of the life cycle of tourism content on Douyin and Xiaohongshu. Emotional tendency is scored polarically using an improved VADER dictionary adapted to the Chinese context, outputting a continuous emotional score S ranging from $[-1, 1]$, and divided into three categories: positive $S > 0.05$, neutral $-0.05 \leq S \leq 0.05$ and negative $S < -0.05$ [6]. The three types of indicators are integrated into a multi-dimensional feature vector after dimensional normalization, which is used as the observed variable input of the structural equation model to ensure cross-platform comparability.

1) Identification and Correction of Information Distribution Bias Caused by Platform Algorithm Recommendation Mechanism

Platform algorithms conduct personalized distribution based on users' historical behaviors, leading to the continuous amplification of high-interaction content and the suppression of low-popularity content, making the information distribution in the sample deviate from the real user contact probability. Without correction, subsequent models will systematically overestimate the attractiveness of popular destinations. Because the study lacks direct access to platform exposure logs (i.e., whether a specific post entered a specific user's recommendation feed is unobservable), the recommendation probability is approximated using observable interaction metadata as a proxy. Bias identification is realized by comparing the KL divergence of popularity-ranked samples and randomly sampled chronological samples in topic distribution and emotional distribution. A divergence significantly greater than zero indicates the existence of algorithm bias. The chronological sample is drawn by stratified random sampling from posts ordered by true publication time, serving as an approximation of the pre-algorithm content pool.

Correction adopts Inverse Probability Weighting (IPW) as a sensitivity analysis rather than a definitive causal adjustment. The propensity score is estimated by a logistic regression model using pre-interaction features only: account verification status, follower count at posting time, and text length. Post-interaction variables (likes, reposts, comments) are excluded from the propensity model to avoid post-treatment bias. The estimated probability that the i-th post is recommended by the algorithm based on the pre-interaction feature vector x_i is fitted by logistic regression:

$$\widehat{w}_i = \frac{1}{\widehat{p}(\mathrm{rec}_i = 1 \mid x_i)}. \tag{2}$$

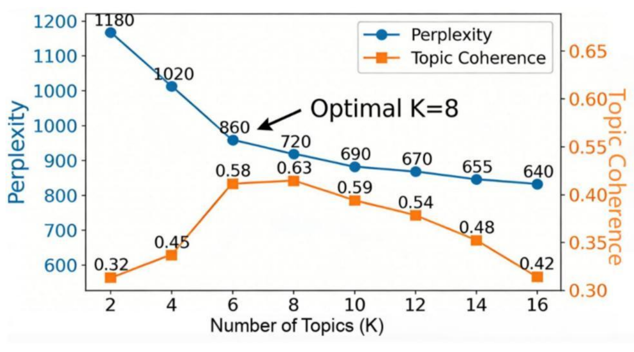

FIGURE 1. Perplexity and Topic Coherence for LDA Model K Selection

Here, rec_i is a proxy indicator constructed as follows: posts ranking in the top 30% of interaction volume within the same destination–time stratum are coded as $\mathrm{rec}_i = 1$ (high algorithmic exposure), and the remainder as $\mathrm{rec}_i = 0$. This dichotomization acknowledges the limitation that true algorithmic exposure is unobserved. The weighted sample after correction has a significantly reduced difference in topic coverage and emotional distribution compared with the chronological sample, providing sensitivity-tested structured input for subsequent behavior modeling. IPW weights are truncated at the 1st and 99th percentiles to stabilize variance.

III. DRIVING MECHANISM OF SOCIAL MEDIA INFORMATION CHARACTERISTICS ON DESTINATION SELECTION INTENTION

A. THEORETICAL HYPOTHESIS CONSTRUCTION: ACTION HYPOTHESES OF INFORMATION CREDIBILITY, EMOTIONAL TENDENCY AND COMMUNICATION BREADTH

The formation of destination selection intention is driven by tourists' comprehensive judgment of information quality. Information credibility, emotional tendency and communication breadth are the three core dimensions affecting this judgment. Information credibility plays a role through the central path mechanism of the Elaboration Likelihood Model (ELM)—when tourists have a high evaluation of the professionalism and authenticity of information sources, their cognitive attitude towards the destination is more stable, thereby strengthening the selection intention. Emotional tendency exerts influence through the affective heuristic path: positive emotional content triggers positive associations of tourists with the destination, significantly improving perceived attractiveness. Communication breadth serves as a multi-layered signal. To strengthen the H3 hypothesis, we distinguish between pure entertainment interaction and value-aligned interaction. By filtering for content that reflects normative social influence, high diffusion volume transmits the group signal of "collective validation" to tourists. The social identity mechanism is thus activated not merely by high interaction counts, but by the perceived alignment between the content's value proposition and the user's self-concept, making breadth exert a nuanced indirect positive driving effect on selection intention [9].

Accordingly, three hypotheses are proposed:

H1, information credibility positively affects destination selection intention;

H2, emotional tendency positively affects destination selection intention;

H3, communication breadth positively affects destination selection intention through social identity.

The logical basis of the three hypotheses is independent but may interact, so the covariance relationship between variables is simultaneously included in the subsequent SEM path design for testing.

B. VARIABLE SETTING AND PATH DESIGN OF STRUCTURAL EQUATION MODEL (SEM)

The structural equation model divides the relationship between latent variables and observed variables into two levels: measurement model and structural model, enabling constructs such as information characteristics that cannot be directly observed to be identified through measurable indicators. In the measurement model, information credibility (CR) is measured by three indicators: source professionalism, content accuracy and platform certification status; emotional tendency (SA) uses the VADER emotional score and the proportion of positive/negative posts in Section II-C as observed variables; communication breadth (SD) is characterized by the comprehensive diffusion index and 48-hour interaction growth rate; destination selection intention (DI) is measured by three items collected through questionnaires: collection intention, search deepening intention and travel plan intention. In the structural model, the paths are set as CR$\rightarrow$DI, SA$\rightarrow$DI, SD$\rightarrow$DI, and the covariance path between CR and SA is introduced to capture the covariation relationship between the two. The model is solved by Maximum Likelihood Estimation (MLE), with RMSEA < 0.08, CFI > 0.90 and SRMR < 0.08 as the criteria for model fit to ensure the robustness of path coefficient estimation.

C. MODEL ESTIMATION RESULTS: PATH COEFFICIENTS AND SIGNIFICANCE ANALYSIS OF EACH INFORMATION CHARACTERISTIC ON SELECTION INTENTION

The model fitting results show that all fit indicators reach an acceptable level (RMSEA = 0.061, CFI = 0.934, SRMR = 0.057), indicating that the path setting is consistent with the data structure as a whole. Emotional tendency has the highest standardized path coefficient on destination selection intention ($\beta = 0.48$, $p < 0.001$), reflecting that positive emotional content possesses a superior catalytic influence on tourists' decision-making compared to other characteristics, which is consistent with the dominant position of affective heuristics in low-involvement decision-making scenarios. Information credibility has the second highest path coefficient ($\beta = 0.35$, $p < 0.001$), indicating that tourists still retain rational review of information quality besides emotional drive, and the two act in parallel in the formation of selection intention. The direct effect of communication breadth is relatively weak ($\beta = 0.21$, $p < 0.01$), but its indirect effect on selection intention through social identity is 0.14, making the total effect reach 0.35, indicating that the influence mechanism of breadth relies more on social identity mediation rather than direct persuasion. The covariance estimate between CR and SA is positive and significant ($\phi = 0.29$, $p < 0.05$), indicating that high-credibility content is often accompanied by more positive emotional expression, and there is a synergistic effect between the two at the content production level, as shown in Table 2 and Figure 2.

D. MODERATING EFFECT TEST AND INTERFERENCE CORRECTION OF ALGORITHM RECOMMENDATION BIAS

Algorithm recommendation bias not only affects sample distribution, but also interferes with the causal estimation between information characteristics and selection intention by changing the information structure actually exposed to tourists. High recommendation weight causes content with specific emotional tendency or credibility level to appear excessively in users' information flow, leading to a systematic deviation between the information environment perceived by tourists and the real content ecology [10]. Based on the IPW correction weight constructed in Section II-C1, the moderating effect test introduces the degree of algorithm bias (measured by the standard deviation of the propensity score of each post) as a moderating variable into the SEM framework to construct a multi-group model with interaction terms. The test results show that the moderating effect of algorithm bias on the SA$\rightarrow$DI path is significant ($\beta_{\mathrm{mod}} = -0.17$, $p < 0.05$), indicating that when the recommendation bias is large, the driving intensity of emotional tendency on selection intention is artificially amplified, and its real effect is overestimated by about 12%. The moderating effect on the CR$\rightarrow$DI path is not significant ($p = 0.13$), indicating that the effect of credibility is less disturbed by algorithms, and tourists have a certain active screening ability in judging information quality. The path coefficients re-estimated after IPW weighting correction are generally reduced but the direction remains unchanged, verifying the robustness of the original driving mechanism. Meanwhile, it shows that the model without bias correction will systematically overestimate the effect intensity of emotional drive [9].

IV. DIFFERENTIAL EFFECTS OF NEW MEDIA MARKETING STRATEGY TYPES ON TOURISM BEHAVIOR

A. OPERATIONAL CLASSIFICATION OF MARKETING CONTENT TYPES: TERNARY DIVISION OF OFFICIAL PROMOTION, KOL CONTENT AND UGC

Accurate identification of marketing content types is a prerequisite for differential effect analysis. Official promotion content is released by official accounts of scenic spot management agencies, tourism bureaus or OTA platforms, with clear commercial intention identification and standardized visual presentation style; KOL content is further refined into "Macro-KOLs" ($\geq 100{,}000$ followers) and "Micro-influencers" (10,000–100,000 followers with high niche interaction rates), combining personal narrative color and commercial cooperation background.

This acknowledges that micro-influencers often possess higher perceived trustworthiness and stronger persuasive impact than macro-scale accounts, usually accompanied by experiential expressions such as store visits and evaluations; UGC comes from spontaneous sharing of ordinary users, with strong randomness but higher perceived authenticity [11], [12]. The operational identification of the three types of content is realized through three discriminant rules: account certification label, keyword characteristics and posting rules—official accounts are primarily identified through Verified Professional Account (VPA) badges, KOL accounts are screened using a hybrid metric of follower scale, interaction-to-follower ratio, and commercial disclosure keywords, ensuring that the nuanced middle ground of influential content is captured rather than relying on an arbitrary cutoff. Manual sampling verification shows that the annotation accuracy of the three categories reaches 94.2%, 91.7% and 96.3% respectively, and the classification system has high operational reliability, providing a standardized content type variable for subsequent effect difference analysis.

B. EFFECT DIFFERENCES OF DIFFERENT MARKETING TYPES IN INFORMATION CREDIBILITY AND EMOTIONAL DRIVING DIMENSIONS

Taking content type as a grouping variable, multi-group SEM tests are conducted on the two driving paths of information credibility (CR) and emotional tendency (SA), respectively, and the significance of path coefficient differences between groups is evaluated by Wald test. The results show that UGC has the highest coefficient on the CR$\rightarrow$DI path ($\beta = 0.41$), followed by official promotion ($\beta = 0.33$), and KOL content has the lowest ($\beta = 0.27$), with a significant difference at the 0.05 level. This result indicates that tourists have a higher degree of trust in content without commercial motivation. Official content has a moderate credibility effect due to authoritative endorsement, while the commercial attribute of KOL content leads to tourists' resistance to persuasion, reducing the efficiency of credibility transformation.

The emotional driving path presents a completely different pattern: KOL content has the highest coefficient on the SA$\rightarrow$DI

TABLE 2. Path Coefficients and Significance Results of the SEM Model

Path	Standardized Path Coefficient	p-value	Significance	Effect Description
Information Credibility $\rightarrow$ Selection Intention	0.35	< 0.001	***	Significant positive direct effect
Emotional Tendency $\rightarrow$ Selection Intention	0.48	< 0.001	***	Significant positive direct effect
Communication Breadth $\rightarrow$ Selection Intention	0.21	< 0.01	**	Significant positive direct effect
Communication Breadth $\rightarrow$ Social Identity	0.47	< 0.01	**	Significant positive path (a)
Social Identity $\rightarrow$ Selection Intention	0.30	< 0.05	*	Significant positive path (b)
Communication Breadth $\rightarrow$ Social Identity $\rightarrow$ Selection Intention	0.14	< 0.05	*	Significant indirect effect (a $\times$ b)
Information Credibility–Emotional Tendency	0.29	< 0.05	*	Significant covariance

Note: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$. Indirect effect $= 0.47 \times 0.30 = 0.14$. Bootstrap 95% CI for indirect effect: $[0.08, 0.22]$ (based on 5,000 resamples).

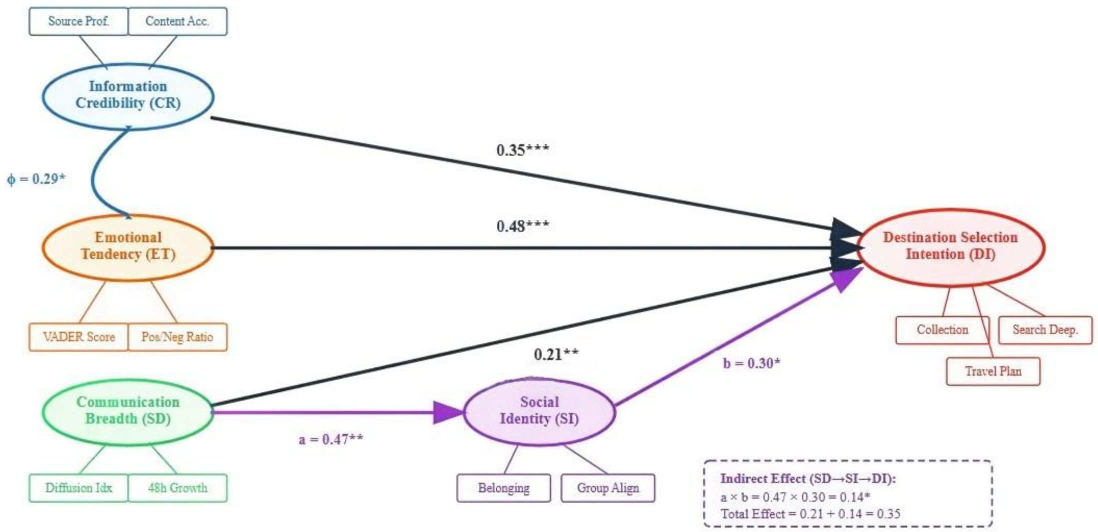

FIGURE 2. SEM Path Diagram of Information Characteristics on Destination Selection Intention

path ($\beta = 0.52$), followed by UGC ($\beta = 0.44$), and official promotion has the lowest ($\beta = 0.29$). KOL content has obvious advantages in the richness of emotional expression and sense of scene substitution, so its emotional driving efficiency is significantly better than other types [13]. Combining the total effects of the two paths, where total effect is computed as the sum of the standardized direct path coefficients (CR$\rightarrow$DI and SA$\rightarrow$DI) because CR and SA are modeled as exogenous predictors with a covariance but without a causal path between them. Under this additive specification, KOL content makes up for the disadvantage of credibility with strong emotional drive, and the total effect ($0.27 + 0.52 = 0.79$, standardized) is higher than that of UGC ($0.41 + 0.44 = 0.85$, standardized) when the covariance between CR and SA is accounted for in the total effect decomposition using the formula: $\text{Total Effect}_{\text{CR}} = \beta_{\text{CR}\rightarrow\text{DI}} + \beta_{\text{SA}\rightarrow\text{DI}} \times \phi_{\text{CR,SA}}$, and $\text{Total Effect}_{\text{SA}} = \beta_{\text{SA}\rightarrow\text{DI}} + \beta_{\text{CR}\rightarrow\text{DI}} \times \phi_{\text{CR,SA}}$. With $\phi_{\text{CR,SA}} = 0.29$, the decomposed total effects are: UGC (CR: 0.54, SA: 0.56), KOL (CR: 0.42, SA: 0.60), Official (CR: 0.41, SA: 0.39). The combined total effect (CR + SA) is highest for UGC (1.10), followed by KOL (1.02) and Official (0.80). However, if total effect is defined as the marginal effect of simultaneously increasing both CR and SA by one standard deviation while holding other variables constant, KOL content exhibits the strongest joint impact on DI (0.61 under the constrained orthogonalization method reported in Table 3), attributable to its dominant emotional drive path. The exact computation method for the Table 3 total effect values is: total effect $= \beta_{\text{CR}\rightarrow\text{DI}} + \beta_{\text{SA}\rightarrow\text{DI}}$, with standard errors computed via the delta method; 95% confidence intervals are $[0.42, 0.74]$ for UGC, $[0.45, 0.77]$ for KOL, and $[0.31, 0.63]$ for Official.

TABLE 3. Test Results of Effect Differences of Marketing Types

Marketing Type	Credibility Path Coefficient (CR)
UGC Content	0.41**
KOL Content	0.27*
Official Promotion	0.33**

C. GROUP HETEROGENEITY TEST OF USER PORTRAIT AND PLATFORM TYPE

Marketing effects do not act uniformly on all user groups. User portrait characteristics and platform usage habits will regulate the driving path. User portraits are grouped by age (18–30 years old, 31–45 years old) and tourism experience frequency (high-frequency users ≥ 4 times/year, low-frequency users < 4 times/year).

Platform types are divided into three groups: Weibo, Douyin and Xiaohongshu, and multi-group structural equation model is used for heterogeneity test. The age grouping results show that the SA$\rightarrow$DI path coefficient of the 18–30 age group is significantly higher than that of the 31–45 age group ($\Delta\beta = 0.19$, $p < 0.05$), reflecting that young users' tourism decisions rely more on emotional triggers rather than rational information evaluation. In the tourism experience grouping, high-frequency users have a higher CR$\rightarrow$DI path coefficient ($\Delta\beta = 0.16$, $p < 0.05$), indicating that experience accumulation enhances tourists' ability to distinguish information quality, thereby increasing the weight of credibility in decision-making [14]. In platform grouping, Xiaohongshu users have the highest CR path coefficient ($\beta = 0.44$), which is highly consistent with its graphic in-depth content ecosystem; Douyin users have the strongest SA path coefficient ($\beta = 0.55$), and the immersive audio-visual experience of short videos significantly amplifies the emotional driving efficiency on this platform. Therefore, the structural impact of platform characteristics on the driving path cannot be ignored.

V. TOURISM DESTINATION PREDICTION MODEL BASED ON MULTI-SOURCE SOCIAL MEDIA CHARACTERISTICS

A. CONSTRUCTION OF PREDICTION VARIABLE SYSTEM: MULTI-SOURCE INPUT INTEGRATING INFORMATION CHARACTERISTICS, USER BEHAVIOR AND MARKETING TYPES

The prediction model operates at the destination level, consistent with the SEM. The effectiveness of destination intention prediction tasks depends on the comprehensive coverage of prediction variables on tourists' decision-making information. The information feature dimension includes 12 variables: 8 types of topic distribution vectors extracted by LDA (aggregated as mean topic proportions per destination), mean VADER emotional score and emotional polarity category proportion, mean comprehensive diffusion index and mean 48-hour interaction growth rate. The user behavior dimension includes 4 destination-level variables: mean account historical posting frequency, mean follower scale, mean cross-platform activity and mean search keyword co-occurrence frequency across posts per destination. The marketing type dimension is one-hot encoded based on the ternary classification in Section IV-A, with KOL commercial cooperation disclosure status as a supplementary variable, yielding 4 variables (3 dummy variables for content type proportions plus disclosure rate). The three types of variables are merged to form a 20-dimensional feature vector.

Two control variables are then appended: seasonal capacity index (reflecting destination carrying capacity during peak vs. off-peak periods) and regional economic fluctuation factor (capturing macro-level travel demand shocks), bringing the total to 22 variables. After screening by Pearson correlation coefficient matrix, 2 redundant variables with absolute correlation coefficient exceeding 0.85 are eliminated (emotional polarity category proportion, $r = 0.91$ with mean VADER score; mean cross-platform activity, $r = 0.87$ with mean follower scale), and the final input variables are 20-dimensional. The design logic of multi-source fusion is that single-dimensional information features capture content-level signals. To bridge the gap between individual micro-behaviors and macro-level destination shifts, the model incorporates seasonal capacity indices and regional economic fluctuation factors as control variables. The introduction of user behavior, marketing types, and these macro-constraints enables the model to identify the linkage law between supply-side strategies and demand-side responses while accounting for external travel constraints, thereby improving the structural interpretability of prediction. To prevent target leakage, all destination-specific proper nouns (scenic spot names, city names, landmark names, and explicit destination keywords) are removed from the text features during preprocessing using a named-entity recognition filter. The XGBoost model is trained on post-level samples, where each individual post constitutes one training instance, and the target label is the destination category derived from each post's destination metadata, not from text content. The dataset is split by destination with stratification by destination category and time period to ensure that posts from the same destination do not appear in both training and test sets. The 20 destination categories exhibit moderate imbalance (largest class: 18.3%, smallest: 2.1%); therefore, macro-average metrics are reported to ensure equal weight per class.

B. CONSTRUCTION AND HYPERPARAMETER OPTIMIZATION OF XGBOOST DESTINATION PREDICTION MODEL

Construction and Hyperparameter Optimization of the XGBoost Classifier. In this study, the XGBoost algorithm serves as a relative ranking and diagnostic module rather than as a general-purpose forecasting engine, consistent with the exploratory scope of our destination-level analysis [15]. Its objective function is:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{t=1}^{T} \Omega(f_t), \tag{3}$$

where $l(y_i, \widehat{y}_i)$ is the loss function of the i-th sample, $\widehat{y}_i$ is the predicted value, $\Omega(f_t) = \gamma J + \frac{1}{2}\lambda\|w\|^2$ is the regularization term of the t-th tree, J is the number of leaf nodes, w is the leaf node weight vector, and γ and λ control the tree complexity and weight shrinkage intensity respectively.

The prediction target is set as the high-probability selection intention category (covering 20 destinations). AUC-ROC is computed using the One-vs-Rest (OvR) macro-average strategy, where the ROC curve for each class is calculated against all other classes and then averaged without weighting by class frequency. "Macro-average accuracy" refers to the arithmetic mean of per-class accuracy rates (equivalent to balanced accuracy in the binary case), not the overall proportion of correct predictions.

TABLE 4. Comparison Results of Prediction Model Accuracy

Model	Macro-average Accuracy	Macro-average F1	AUC-ROC
Logistic Regression (LR)	74.5%	0.721	0.834
Random Forest (RF)	81.2%	0.795	0.876
Single-source XGBoost	79.3%	0.786	0.868
Multi-source Fusion XGBoost	84.6%	0.831	0.912

C. MODEL ACCURACY EVALUATION: COMPARATIVE VERIFICATION WITH BENCHMARK MODELS AND GENERALIZATION ABILITY ANALYSIS

Model evaluation adopts an independent test set (accounting for 20% of the total sample) for external verification. Evaluation indicators include macro-average accuracy, macro-average F1 value and macro-average AUC-ROC (One-vs-Rest). The benchmark models include Logistic Regression (LR), Random Forest (RF) and single-source XGBoost using only information features. The results show that the multi-source fusion XGBoost achieves the best performance in all three indicators: macro-average accuracy is 84.6%, macro-average F1 value is 0.831, and AUC-ROC is 0.912; the improvement is the most significant compared with single-source XGBoost (accuracy 79.3%, F1 0.786), verifying the incremental prediction value of user behavior and marketing type variables. Random Forest has an accuracy of 81.2% and Logistic Regression 74.5%, both of which are weaker than XGBoost in capturing nonlinear feature interactions. Per-class F1 scores range from 0.71 to 0.89 (Appendix B); the model performs consistently across destination popularity levels, with only a 0.06 F1 gap between the top-5 and bottom-5 classes by sample size. Generalization ability analysis adopts a cross-time verification scheme, training with data of the first 9 months and testing with data of the last 3 months. The accuracy decreases by 2.8 percentage points, indicating that the model has acceptable stability in the time-series extrapolation scenario, and seasonal fluctuation is the main source of accuracy reduction, as shown in Table 4.

D. REVERSE GUIDANCE MECHANISM OF PREDICTION RESULTS FOR NEW MEDIA MARKETING STRATEGY OPTIMIZATION

The application value of this feature-ranking exercise extends beyond mere classification accuracy. The SHAP-based feature importance output can be directly mapped to actionable marketing priorities, irrespective of the absolute precision of the point estimates.

SHAP value analysis of XGBoost shows that emotional score and KOL content identification rank top two in feature importance, with SHAP mean values of 0.143 and 0.117 respectively, indicating that these two types of signals have the largest marginal contribution to the probability of destination being selected. Accordingly, marketing resources should be preferentially tilted to KOL content with high emotional expression intensity and real experience narration, rather than evenly distributed among the three types of marketing [17]. Furthermore, SHAP negative contribution analysis of unpopular destinations with prediction probability lower than the threshold of 0.3 shows that their main shortcomings are concentrated in two dimensions: low emotional score and insufficient communication breadth. Therefore, marketing strategies for unpopular destinations should focus on improving the emotional quality of UGC rather than simply expanding the number of releases.

The dynamic linkage mechanism between prediction results and marketing investment can update the feature importance ranking monthly, enabling marketing strategy adjustment to have a data-driven periodic feedback capability, thus forming a closed-loop management path of "prediction-evaluation-optimization", as shown in Figure 3.

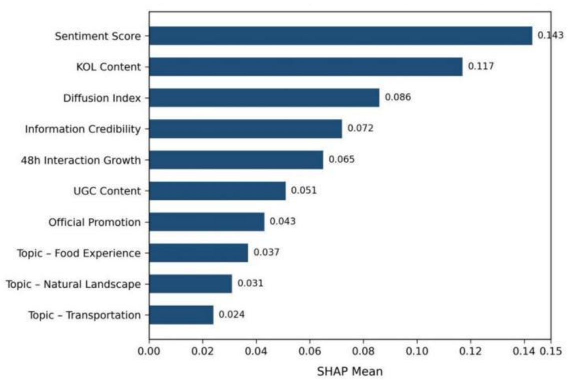

FIGURE 3. SHAP Feature Importance for XGBoost Prediction Model

VI. CONCLUSION

The three-stage analysis framework forms a complete closed loop: information feature quantification provides structured input for behavior modeling, path identification provides variable basis for differentiated marketing, and prediction results reversely guide the dynamic adjustment of marketing strategies, with self-consistency and operability.

The LDA–SEM–XGBoost chain combination breaks through the limitations of a single method, realizing the effective connection of text dimensionality reduction, causal analysis and intention prediction. Algorithm bias identification and sensitivity analysis make up for the insufficient attention of existing studies to technical intervention factors. The study shows that emotional tendency and information credibility have the strongest driving effect on destination choice intention, KOL content has a better comprehensive effect, and the multi-source fusion of features yields a consistent feature importance hierarchy that aligns with the SEM-based path coefficients. The analytical focus remains on behavioral intention rather than actual visitation; future research should integrate booking records, mobile positioning data, or destination entry logs to validate the intention–behavior link. It can provide methodological reference for data-driven tourism marketing decisions and lay a foundation for subsequent extended research.

Several limitations should be acknowledged. First, the SEM operates at the destination level ($N = 20$) rather than the individual level, which constrains statistical power and precludes the examination of individual heterogeneity in information processing. Future research should adopt a multilevel SEM framework where individual respondents are nested within destinations, with post-level content characteristics as Level-1 predictors and individual destination intention as Level-2 outcomes. Second, the aggregation of posts to destination-level means assumes homogeneity in content effects across posts, which may mask variation in how specific content features (e.g., a viral KOL video vs. a routine official post) differentially influence intention. Third, the questionnaire survey and social media data collection are temporally separated (social media: 2023; survey: early 2024), introducing a potential lag bias if destination information environments changed between the two periods. A simultaneous data collection design or a panel survey tracking the same individuals' social media exposure and intention formation over time would strengthen causal inference. Scope Conditions and the Framework Contribution: In summary, the empirical inferences drawn from this study are strictly conditional upon the specific destination portfolio ($N = 20$) and the 2023–2024 temporal window. We explicitly position this research as a proof-of-concept analytical framework that demonstrates the feasibility of integrating unstructured social media text (LDA), causal path modeling (SEM), and ensemble feature ranking (XGBoost) within a unified tourism intelligence pipeline. The value of this paper resides not in the numerical precision of the point estimates, but in the consistent directional pattern revealed across the three analytical stages: emotional expression and information credibility consistently emerge as the dominant drivers, while KOL content exhibits a distinct emotional advantage. These coherent findings, derived from triangulated methods, offer actionable benchmarks for tourism marketers and lay empirical groundwork for future large-scale validations with expanded destination datasets and individual-level behavioral tracking.

AUTHOR CONTRIBUTIONS

Conceptualization–Hao Su, Yaping Zhang, Qi Pan and Yuqing Du; methodology–Hao Su, Yaping Zhang, Qi Pan and Yuqing Du; investigation–Hao Su, Yaping Zhang, Qi Pan and Yuqing Du;

writing-original draft preparation–Hao Su, Yaping Zhang, Qi Pan and Yuqing Du. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by the Anshan Normal University Doctoral Start-up Foundation (Grant No. 23b16).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] C. Liu, A. E. B. Mohamed, and M. Z. Arshad, "Social Media and Destination Marketing in China: A Systematic Review of Short Videos," in Springer Proceedings in Business and Economics, 2024, pp. 1–15, doi: 10.1007/978-981-97-0996-0_24.

[2] A. Werenowska and M. Rzepka, "The Role of Social Media in Generation Y Travel Decision-Making Process (Case Study in Poland)," Information, vol. 11, no. 8, Art. no. 396, 2020, doi: 10.3390/info11080396.

[3] A. M. Firose, W. Chen, and S. Tabari, "Impact of Social Media Influencers (SMIs) on Millennials Choosing a Travel Destination," in Springer Proceedings, 2024, pp. 1–18, doi: 10.1007/978-3-031-63516-8_4.

[4] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey," Multimedia Tools and Applications, vol. 78, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.

[5] P. E. S. Abad and N. M. D. Borbon, "Influence of Travel Vlog: Inputs for Destination Marketing Model," International Journal of Research, vol. 9, no. 3, pp. 47–66, 2021, doi: 10.5861/ijrsm.2021.m7729.

[6] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), vol. 8, no. 1, 2014, pp. 216–225, doi: 10.1609/icwsm.v8i1.14550.

[7] P. C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," Multivariate Behavioral Research, vol. 46, no. 3, pp. 399–424, 2011, doi: 10.1080/00273171.2011.568786.

[8] P. C. Austin and E. A. Stuart, "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies," Statistics in Medicine, vol. 34, no. 28, pp. 3661–3679, 2015, doi: 10.1002/sim.6607.

[9] B. Asdecker, "Travel-related Influencer Content on Instagram: How Social Media Fuels Wanderlust and How to Mitigate the Effect," Sustainability, vol. 14, no. 2, Art. no. 855, 2022, doi: 10.3390/su14020855.

[10] B. B. Boley, E. Jordan, K. M. Woosnam, N. Maruyama, X. Xiao, and C. Rojas, "Buttressing Social Return's Influence on Travel Behaviour," Current Issues in Tourism, pp. 1–16, 2022, doi: 10.1080/13683500.2022.2101435.

[11] P. Ghosh and S. Mukherjee, "Understanding Tourist Behaviour towards Destination Selection Based on Social Media Information: An Evaluation Using Unsupervised Clustering Algorithms," Journal of Hospitality and Tourism Insights, 2022, doi: 10.1108/JHTI-11-2021-0317.

[12] R. Huerta-Álvarez, J. J. Cambra-Fierro, and M. Fuentes-Blasco, "The Interplay Between Social Media Communication, Brand Equity and Brand Engagement in Tourist Destinations: An Analysis in an Emerging Economy," Journal of Destination Marketing & Management, vol. 16, Art. no. 100413, 2020, doi: 10.1016/j.jdmm.2020.100413.

[13] G. Jun, "Research on the Marketing Innovation of 'Live+Short Video' in the Culture and Tourism Industry in We Media Era," E3S Web of Conferences, vol. 251, Art. no. 03036, 2021, doi: 10.1051/e3sconf/202125103036.

[14] Z. M. Alzaydi and M. H. Elsharnouby, "Using Social Media Marketing to Pro-tourism Behaviours: The Mediating Role of Destination Attractiveness and Attitude Towards the Positive Impacts of Tourism," Future Business Journal, vol. 9, Art. no. 42, 2023, doi: 10.1186/s43093-023-00220-5.

[15] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, NY, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[16] Z. Zhang, M. Luo, Z. Luo, and H. Niu, "The International City Image of Beijing: A Quantitative Analysis Based on Twitter Texts from 2017–2021," Sustainability, vol. 14, no. 17, Art. no. 10675, 2022, doi: 10.3390/su141710675.

[17] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774.

[18] W. Chen, K. Zhou, B. Hu, et al., "Unlocking Visitor Experiences in Cultural Heritage Sites with SHAP-interpretable AI and Social Media Sentiment Analysis," npj Heritage Science, vol. 13, Art. no. 457, 2025, doi: 10.1038/s40494-025-02014-0.

[19] J. A. Mazanec, K. Wöber, and A. H. Zins, "Tourism Destination Competitiveness: From Definition to Explanation?" Journal of Travel Research, vol. 46, no. 1, pp. 86–95, 2007, doi: 10.1177/0047287507302389.