PERINATAL JOURNAL

Original Article

Perinatal Journal 2026; 34(1):1001-1014

https://doi.org/10.57239/prn.26.034100102

Communication alienation and value reconstruction of literary and artistic criticism in the era of short videos: Taking film and television review exts as an example

Miaozhu Zhang*

Graduate School, Jining Normal University, Ulanqab 012000, Inner Mongolia, China

Abstract

Existing research is largely limited to qualitative exploration of the impact of short videos on literary and artistic criticism, lacking quantitative empirical evidence and causal evidence for the core mechanism of "communication alienation". The purpose of this study is to provide a computational framework that can successfully dissect the mechanism of alienation through dynamic topic modeling and multimodal sentiment divergence method. This framework first measures the sentiment of review text using a BERT (Bidirectional Encoder Representations from Transformers)-based model and leverages pre-trained automatic audio and video analysis models, VGG16 and OpenSmile, to extract high-dimensional sentiment information from audio and video data. Next, a cross-modal attention mechanism is then used to calculate the relative weight of audiovisual sentiment to textual sentiment, quantifying the degree of sentiment divergence in the "form drives content" model. Furthermore, a dynamic topic model (DTM) is used to track the temporal decay of the topic depth in comment texts. Granger causality tests are then conducted with video communication metrics to empirically demonstrate the causal relationship between alienation features and traffic incentives. Experimental results show that the mean completion rate increases from 0.28 in the low-alienation range to 0.49 in the high-alienation range, confirming the driving effect of formal expression on communication effectiveness. Multimodal features indicate that the visual modality has a weight of 0.45 in predicting completion rate, significantly higher than the 0.25 weight of the text modality, reflecting the alienated nature of "form overwhelms content". This study provides quantifiable analysis and a solid theoretical basis for understanding the changing form of literary and art criticism in the algorithmic era, and also provides an operational governance path for platform algorithm optimization and value reconstruction of the comment ecosystem.

Keywords: Short video era, Literary and artistic criticism, Communication alienation, Multimodal sentiment analysis, Dynamic topic model

1.Introduction

The rise of short video platforms has completely reshaped the dissemination ecosystem of literary and artistic criticism. Critical practice, taking film and television criticism as an example, has rapidly shifted from the expression form of professional journals or long articles to fragmented, visual, and short video expression forms [1-2]. This media transformation has changed the way critical practice is disseminated, and has also deeply intervened in its production logic and reception mode, arousing people's deep concern about the intrinsic value and public significance of critical practice [3-4]. Against this backdrop, systematically analyzing the communication alienation of short video film and television reviews and exploring pathways for their value reconstruction has both urgent practical and theoretical significance for guiding the healthy development of the online literary and artistic ecosystem and safeguarding the rational spirit of criticism.

Current research explores a range of specific issues. The core representation of alienation, “form overwhelms content”, lacks practical quantitative metrics or definitions, which has kept related discussions almost entirely at the phenomenological level [5-6]. Within a large amount of short video comment data, the emotional factors that drive user engagement are coupled with factors related to audiovisual form, limiting the ability to isolate and verify their individual and interactive influences in research [7-8]. As an influencing variable of the environment, there is currently no empirical data on the causal relationship between algorithmic recommendation mechanisms and the alienation of comment content, relying heavily on logical deduction [9-10]. In the context of short videos, the “depth” of comment quality must be reconsidered, as even traditional metrics based on text complexity fail here [11-12, 31].

A range of studies have attempted to address these issues through various approaches. Content analysis,

Correspondence: Miaozhu Zhang, Graduate School, Jining Normal University, Ulanqab 012000, Inner Mongolia, China, **e-mail:** zhangmz8866@126.com,
Received: November 1, 2025 **Accepted:** December 29, 2025

while using manually coded video information to track surface features, may not provide insight into the audience's underlying dynamic communication mechanisms due to subjectivity and scale limitations [13-14]. Traditional sentiment analysis models tend to analyze the text within comments, failing to examine the potentially dominant emotional signals emitted simultaneously by audiovisual stimuli [15-16]. User studies using questionnaires or related methods may involve subjective perceptions, but they cannot objectively provide evidence of the full pattern of actual communication data in their metrics [17-18]. These methodological issues create an analytical perspective that is too weak to simultaneously examine and process multimodal data, nor can it generate internal deviation counts that can explain macro-level communication effects, and therefore cannot explain the mechanism of alienation.

To empirically quantify the "form drives content" alienation mechanism, a computational framework based on multimodal sentiment divergence analysis and dynamic topic modeling is proposed. Within this framework, sentiment vectors are generated for review text and audiovisual streams using a BERT-based text model and an audiovisual feature extraction model, respectively. A cross-modal attention network then calculates the attention weight of the audiovisual modality relative to the textual modality and considers this as a core indicator of sentiment divergence. A dynamic topic model is also used to track the temporal evolution of topic depth within review text sequences. Finally, Granger causality tests are performed on divergence, topic depth decay rate, and video diffusion metrics to identify potential driving relationships.

2.Alienation analysis model based on multimodal feature fusion

2.1 Data collection and preprocessing

The dataset for this research is designed to provide a thorough multifaceted account of short video film and television reviews. The dataset is built from three facets: video content, review text, and communication behavior. Data collection occurs through a focused crawl of the platform’s official open interface. The selected videos conform to an inclusion criterion in the form of short videos with one of the central tags “movie”, “TV series”, or “film review” and a length between 30 seconds to 5 minutes [19-20]. In total, the dataset of videos presented here contains 15,327 valid video instances, as well as re-used metadata. After performing the download, the original videos are decomposed into three individual data streams for later processing: video stream, audio stream, and text stream. The video stream has been uniformly sampled into an RGB image sequence with a frame rate of 25 frames-per-second, and the audio stream has been re-sampled into a 16 kHz mono waveform file to ensure the input format is consistent for feature extraction in a later phase. The text stream has video subtitles using automatic speech recognition, a video release title, and the 200 most-liked comments as texts.

Four fundamental metrics are generated from each video sample to measure its popularity: likes, shares, comments, and completion rate. Completion rate is defined as the ratio of the number of complete views of the video to the total number of impressions, and it reflects user retention of the content. Table 1 shows the characteristics of the dataset in a more detailed statistical manner. The table presents the metrics' central tendency and dispersion in a structured way, which is vital for future analysis.

Table 1. Statistical characteristics of the dataset

Category	Statistical Metric	Value
Overall Scale	Total Video Samples	15,327
	Data Collection Time Window	2023.01 - 2024.12
	Total Video Duration (Hours)	542.5
	Total Associated Comment Texts (×10^4)	482.1
Video Content Features	Average Duration (Seconds)	127.4
	Standard Deviation of Duration (Seconds)	58.7
	Min/Max Duration (Seconds)	31 / 298

Communication Impact	Average Number of Likes	15,642
	Median Number of Likes	8,795
	Average Number of Shares	1,258
	Average Play-through Rate	0.351
	Standard Deviation of Play-through Rate	0.127
User Interaction Depth	Average Number of Comments	314.5
	Average Number of Favorites per Video	2,451
	Comment Reply Rate (Proportion of Replied Comments)	0.186

The processing stage of textual data requires standard natural language processing practices - removal of irrelevant characters, word segmentation, and stop word filtering [21-22]. Given the specificity of online language, a context-specific stop word library that includes a list of high-frequency phrases that do not contribute to the essence of the review is created, i.e., "family" or "three-click combo". The text sequences are represented in word vector format derived from the BERT model in preparation for deep semantic analysis [23-24]. The extraction of audiovisual features is based on a single framework that specifies the mapping path from raw multimedia data to structured feature vectors. The complete architecture is shown in Figure 1.

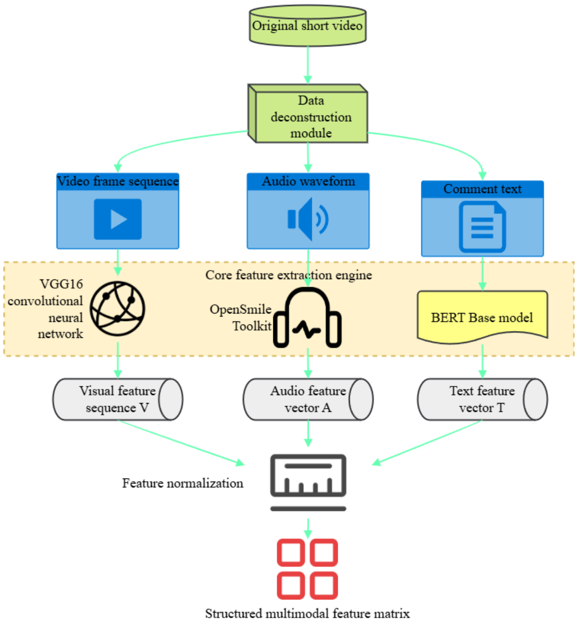

Figure 1. Multimodal data feature extraction framework

Figure 1 shows the full processing flow path from raw data to structured features. In the initial data deconstruction module, the original short video is divided into three separate data streams that function independently from one another: the video frame sequence, audio waveform, and comment text data stream. These data streams are fed in parallel into the core feature extraction engine. The visual features are extracted using a convolutional neural network, VGG16 (Visual Geometry Group 16), that processes the visual frames and outputs sequenced features. The audio features are extracted as high-dimensional acoustic vectors using the OpenSmile toolkit. Finally, the text features are encoded as semantic vectors from the comment stream of data using the BERT model. These three heterogeneous feature outputs undergo unified feature normalization and are ultimately integrated to construct a unified structured multimodal feature matrix. This design innovatively achieves end-to-end collaborative extraction and standardized integration of features from different modalities, providing a precise data foundation for the subsequent calculation of sentiment deviation and analysis of communication alienation mechanisms.

The video frame sequence is input into the pre-trained VGG-16 convolutional neural network, and the activation value before the last fully connected layer is extracted as the visual feature vector V_{frame}. For a video lasting t seconds, its features are aggregated into a visual feature sequence V=$\{V_1,V_2,...,V_{25t}\}$. The OpenSmile toolkit is used for audio feature extraction, and its acoustic feature set is configured. The set contains multi-dimensional feature vectors calculated by a variety of low-level descriptors and their temporal statistical functions [25-26]. Each audio is represented as an audio feature vector A$\in$R^{3427}. The text comments are processed by the BERT-base model, and the 768-dimensional vector corresponding to the [CLS] tag of the last hidden layer of each sentence is obtained as its semantic representation T. The Z-score method is used for normalization before multimodal feature fusion, and its calculation is shown in Formula (1):

$$X_{\text{norm}}=\frac{X\text{-}\mu}{\sigma}(1)$$

In Formula 1, X represents the original eigenvalue; μ is the mean of the feature across all samples; σ is the standard deviation. This operation unifies features of different dimensions and ranges to the same scale, eliminating the potential negative impact of numerical differences on model convergence and laying a solid foundation for subsequent cross-modal alignment and joint analysis.

2.2 Multimodal sentiment deviation calculation

Multimodal sentiment divergence calculation aims to quantify the degree of emotional dominance of the audiovisual modality over the textual modality. Its core lies in building a cross-modal interaction mechanism and defining a computable divergence metric [27-28]. This module takes as input the structured multimodal feature matrix generated in the previous stage, where the visual feature sequence $\mathrm{V}\in\mathrm{R}^{N_v\times d_v}$, audio feature vector $\mathrm{A}\in\mathrm{R}^{d_a}$, and text feature vector $\mathrm{T}\in\mathrm{R}^{d_t}$ represent the sentiment representations of each modality. To achieve semantic alignment between modalities, the heterogeneous features are first mapped to a unified sentiment semantic space through a fully connected layer:

$$\mathrm{H}_m=\mathrm{W}_m\mathrm{X}_m+\mathrm{b}_m(2)$$

In Formula 2, $m\in\{v,a,t\}$ identifies the modality type; W_m and b_m are trainable parameters; X_m corresponds to the original features. The projected features H_v, H_a, H_t have the same dimension d_{h}, forming a comparable,sentiment representation.

In order to capture the attention interference of audiovisual modalities on text emotions, this module designs a cross-modal attention mechanism. The text features H_t are used as query vectors and the audiovisual fusion features $[\mathrm{H}_v;\ \mathrm{H}_a]$ as key-value pairs, and the attention weights are calculated:

$$\omega_i=\frac{\exp\left(\mathrm{H}_t\cdot\mathrm{K}_i^{\mathrm{T}}\right)}{\sum_j\ \exp\left(\mathrm{H}_t\cdot\mathrm{K}_j^{\mathrm{T}}\right)}(3)$$

In Formula 3, K_i represents the ith audiovisual feature segment, and the attention distribution ω reflects the relative influence of different audiovisual segments on the text sentiment. The audiovisual context vector weighted by this distribution is:

$$\mathrm{C}_{va}=\sum_i\ \omega_i\mathrm{V}_i(4)$$

The sentiment deviation is ultimately defined as the semantic deviation measure between the audiovisual context vector and the text feature vector, calculated by cosine similarity:

$$\text{SentimentDev}=1\text{-}\cos(\mathrm{C}_{va},\mathrm{H}_t)(5)$$

The value range of this indicator is [0,2]. Larger values indicate a greater divergence between the visual and auditory emotional expression and the rational expression of the text. A value approaching 2 indicates that the video content is completely dominated by sensory stimulation. This quantitative result serves as a key independent variable in subsequent causal analysis to verify the alienating effect of formal expression on the essence of content.

2.3 Topic depth attenuation model

The topic deep decay model aims to quantify the deep degradation of film and television review content in the context of short video communication. The model takes the preprocessed set of all review texts as input. Its core assumption is that the value density of the review content shows a systematic decay as the communication heat increases. The model construction first performs time series modeling on the text corpus based on the dynamic topic model, and divides the review texts sorted by the video release time into continuous monthly time slices. The word frequency distribution in each time slice is generated through the Dirichlet prior to generate the topic distribution [29]. The generation process is shown in Formula (6):

$$\begin{aligned}&\phi_k\sim\text{Dirichlet}(\beta)\quad \text{for } k\in\{1,...,K\}\\&\theta_t\sim\text{Dirichlet}(\alpha)\quad \text{for } t\in\{1,...,T\}\qquad(6)\\&z_{t,n}\sim\text{Multinomial}(\theta_t)\end{aligned}$$

$$w_{t,n} \sim \text{Multinomial}\left(\phi_{z_{t,n}}\right)$$

In Formula 6, K represents the number of preset topics; T is the total number of time slices; ϕ_k is the term distribution of the kth topic; θ_t is the topic distribution of the tth time slice; $z_{t,n}$ and $w_{t,n}$ represent the topic assignment and observation value of the n-th word in the t-th time slice, respectively. This modeling approach can capture the temporal evolution of topic content. The parameter configuration of the dynamic topic model and the topic quality assessment results are shown in Table 2.

Table 2. Dynamic topic model parameters and topic consistency indicators

Parameter category	Parameter name	Parameter value
Prior parameters	Topic distribution prior α	0.1
	Term distribution prior β	0.01
Model structure	Number of topics K	50
	Time slice number T	24
Evaluation metrics	Thematic consistency score	0.68
	Perplexity	1,245

After obtaining a stable topic model, it is necessary to define a computable topic depth metric. This study constructs a depth evaluation framework based on external knowledge anchoring, quantifying the professionalism of each topic by calculating the semantic similarity between each topic and a professional film review dictionary D. Specifically, for the top-N representative terms set W_k={w_1,w_2,...,w_N} of topic k, its topic depth score is defined as shown in Formula (7):

$$\text{Depth}(k)=\frac{1}{|D|}\sum_{d\in D}\max_{w\in W_k}\text{Sim}(\text{E}(w),\text{E}(d))(7)$$

In Formula 7, E(·) represents the BERT-based word vector encoder, and Sim is the cosine similarity function. This metric measures the maximum alignment between topic terms and professional review terms in the semantic space.

To capture the decay of topic depth as virality increases, videos are divided into four topic depth quantile intervals, Q1 to Q4, based on their virality intensity. For each interval, the topic depth scores are averaged using weights based on the frequency of the topic in that respective interval. To visualize the output of the topic depth decay model and how to create the various core metrics, this concept is illustrated in Figure 2.

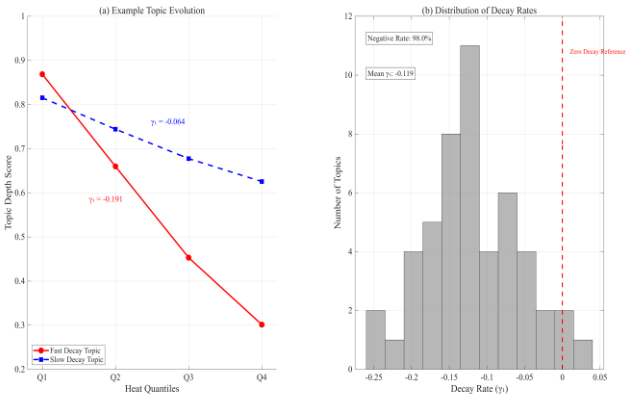

Figure 2. Subject depth decay pattern

The left side of Figure 2 illustrates the depth trajectories of two different representative topics. The depth score of the rapidly-decaying topic starts at 0.88 in Q1, dropping to 0.3 in Q4, with a decay rate of -0.191. The other representative topic has a slower decay with a score starting at 0.82 and is sustained to only 0.62 over the same duration, resulting in a decay rate of -0.064. The large contrast between the two trajectories illustrates the dynamic topic model's ability to capture heterogeneous decay. The other side presents a histogram of the decay rates for all 50 topics. The data shows that as many as 98 % of topics have negative decay rates, with an aggregate average decay rate of -0.119. This distribution pattern provides a statistical demonstration that deep decay is a prevalent, systematic phenomenon in the set of topics, as opposed to a singular occurrence of a few topics. The figure solidly portrays the logic of how the raw sequences of text are mapped into measurable decay with quality metrics, which subsequently leads to a subsequent causal analysis.

The quantitative calculation of the decay rate uses linear regression fitting to establish a regression model for each subject sequence $\{\text{Depth}(k,Q_i)\}_{i=1}^{4}$:

$$\text{Depth}(k,Q)=\gamma_0+\gamma_1\cdot Q+\epsilon(8)$$

The regression coefficient γ_1 in Formula 8 indicates the decay rate for the depth of the topic. A negative value implies that as popularity increases, the depth of the topic decreases. The decay rate distribution for all topics provides the primary input for future causal analysis and provides empirical evidence for confirming the connection between viral popularity and shallower content.

3.Empirical design of alienation mechanisms

3.1 Experimental data and settings

To test the explanatory power of multimodal sentiment divergence and topic depth decay with respect to the alienated communication phenomenon (across film and television reviews), this study employs a rigorous data partitioning and baseline comparison scheme. The experimental dataset includes 15,327 valid video samples, which are divided at random into training, validation, and test sets in a 7:2:1 sample ratio. The training set is solely utilized to perform parametric learning for the dynamic topic model and cross-modal attention network. The validation set is reserved for hyperparameter tuning and satisfaction of assumptions. Finally, the test set serves as the baseline and is executed for all measurements against final performance and causal tests. In order to preserve the integrity of our time series analysis, the dataset is randomized at the sample level for partitioning in order to avoid potential biases within the model evaluation associated with time leakage.

The experiment sets up three specific baseline methods for comparative study. Baseline 1 method uses the TextCNN convolutional neural network model, which uses 300-dimensional pre-trained word vectors as input, extracts text features through a one-dimensional convolution layer with convolution kernel sizes of 3, 4, and 5, and finally outputs a single-modal sentiment classification result based on the comment text. Baseline 2 method uses the Early Fusion multimodal fusion model, which splices visual features, audio features, and text features at the input layer, and realizes multimodal sentiment regression through a three-layer fully connected network. Baseline 3 method adopts the Tensor Fusion Network model, which explicitly models the high-order interactions between modalities through outer product operations. Its fusion tensor is input into the classifier after modality-specific factorization. All baseline models and the proposed cross-modal attention model are trained on the same training set, using the Adam optimizer with a cosine annealing learning rate schedule. Early stopping is triggered when validation set loss fails to improve for five consecutive epochs.

The construction of the baseline model provides a comparative benchmark for the computational reliability of key variables in subsequent causal analysis. Experiments are conducted on a workstation equipped with an NVIDIA RTX 3090 GPU. All models are trained and tested using the PyTorch 1.12 framework and Python 3.9. Key hyperparameters for all comparison models and the proposed model are determined through grid search, with the final configurations shown in Table 3.

Table 3. Model hyperparameter configuration table

Model component	Hyperparameter	Value
Cross-Modal attention network	Learning rate	5×10^{-4}
	Batch size	32
	Hidden dimension	512
	Number of attention heads	8
	Dropout rate	0.3
	Optimizer	Adam

Dynamic topic model (DTM)	Number of topics	50
	Topic distribution prior	0.1
	Word distribution prior	0.01
	Chunk size	200
	Number of iterations	1500
Vector autoregression (VAR) model	Lag order	3
	Trend specification	Constant
	Cointegration test	Johansen test
	Significance level	0.05

3.2 Causality test design

To empirically demonstrate the causal driving mechanism of multimodal sentiment deviation and topic depth decay rate on the video communication effect, this study uses the Granger causality test based on the vector autoregression model for quantitative analysis. The test constructs a three-variable VAR (Vector Autoregression) system, taking sentiment deviation, topic depth decay rate, and video completion rate as endogenous variables, while controlling for exogenous covariates such as video length and release time. The general form of the VAR model is shown in Formula (9):

$$Y_t=A_0+\sum_{i=1}^{p} A_i Y_{t-i}+BX_t+\epsilon_t(9)$$

Y_t is the endogenous variable vector; p represents the lag order; A_i is the coefficient matrix; X_t represents the exogenous control variable vector; ϵ_t is the white noise process. The core principle of Granger causality is that if the lagged value of a variable X can significantly improve the prediction accuracy of the variable Y, then X is said to be the Granger cause of Y.

Before conducting the test, variable stationarity must be ensured. ADF (Augmented Dickey-Fuller) unit root tests are performed on each variable series, and first-order differencing is performed on non-stationary series until stationarity is met. The optimal lag order for the VAR system is determined based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC). The minimum lag order that achieves the most agreement is selected. Granger causality is assessed using the Wald test. The null hypothesis is that "the coefficients of all lag terms are jointly zero". If the p-value is less than the significance level of 0.05, the null hypothesis is rejected, and the presence of a causal relationship is confirmed. Table 4 shows the definitions of all variables involved in the test, data sources, and preprocessing methods.

Table 4. Granger causality test variable definition and preprocessing

Variable category	**Variable name**	**Preprocessing method**	**Expected impact Direction**
Endogenous variables	Sentiment deviance	Z-score standardization	Positive
	Topic depth decay rate	First-Order differencing	Positive
	Video completion rate	Log transformation	-
Exogenous variables	Video duration	Binning discretization	Control
	Publishing time period	Quarterly dummy variables	Control

Table 4 presents the variable system for causal testing. Within the Granger causality testing framework, endogenous variables are used to test the impact of sentiment divergence and topic depth decay as potential "explanatory variables" on video completion rates, the "explained variable". Exogenous variables are used to control for known confounding factors. The differences in preprocessing methods reflect the data characteristics of each variable, and the expected direction of influence provides theoretical a priori information for subsequent interpretation of the results. The complete VAR model parameter configuration includes a range for selecting lag orders, stability testing requirements, and a cointegration treatment scheme. These technical details collectively ensure the statistical reliability of causal inference.

4.Results

4.1 Distribution and impact of sentiment deviation

To better understand the reasons for differences in short video review communication, the systematic effect of sentiment expression and rationality of content difference on communication outcomes is first investigated. This section uses quantification to analyze the distribution characteristics of multimodal sentiment divergence as well as its relationship to some key communication measures to demonstrate how sensory-stimulus-first communication logic reconfigures the review ecosystem. To calculate sentiment divergence, sentiment divergence is examined across a large sample of videos, and the data is separated into four continuous intervals to run between-groups comparisons. To look at its relationship with more broadly observed trend in user behavior results, it is also analyzed using nonparametric correlation. The findings are summarized in Figure 3.

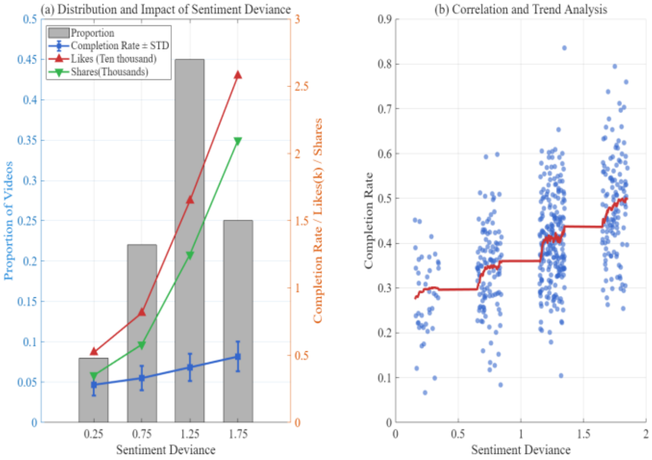

Figure 3. Distribution and correlation of sentiment alienation: (a) Distribution and impact of sentiment alienation; (b) Correlation and trend

Figure 3(a) shows the proportion and effects of sentiment divergence. The left axis is the proportion of videos; the right axis is the completion rate, likes, and shares. Videos with divergence from 1.0 to 1.5 make up the largest portion at 45%, meaning moderately high levels of alienation have become more mainstream in commenting content. The completion rate exhibits a monotonically increasing tendency as divergence increases, rising significantly from 0.28 for the low alienation range to 0.49 for the high alienation range, with a highly varied standard deviation from 0.08 to 0.11. This means high alienating content has more reach but has less variability. The conclusion is supported by the simultaneous increase in likes and shares from 5,200 and 350 to 25,800 and 2,100, respectively, which backs the hypothesis that the recommendation algorithm and stylized attention of users on short video platforms are both cumulative. In short, highly emotional and sensory content capture user attention quicker and instigate some level of interactivity during brief browsing events. The scatterplot and fitted curve in Figure 3(b) show the positive correlation between divergence and completion rate reinforces the reinforcing effect of formal expression on communication effectiveness. Overall, an increase in sentiment divergence is consistent showing a strong, significant relationship.

4.2 Verification of the universality of topic depth attenuation

Given that a quantitative evaluation method for topic depth decay has been established, the next step is to explore whether topic depth decay represents a systematic feature of the short video film and television review ecosystem, rather than just a random feature of individual content. In this section, the focus is on the overall pattern of content depth decay under the background of content popularity. By comprehensively analyzing the content depth trajectories of fifty themes, the core prediction of alienation theory about shallow content is verified at the macro level. The theme depth sequence output by the dynamic theme model is tested, and their group mean and statistical distribution in different heat intervals are calculated. The results are presented in the form of a comprehensive theme depth trajectory to visualize the contradictions they represent, as depicted in Figure 4.

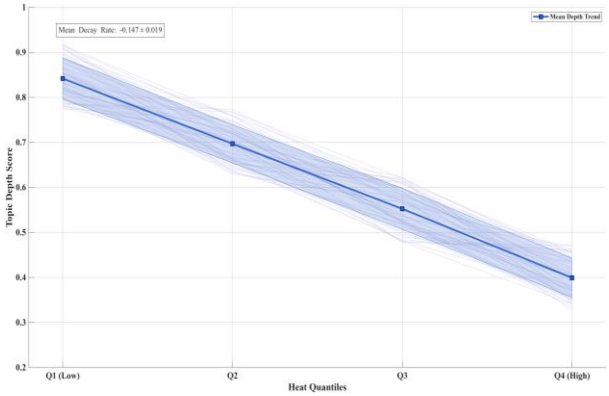

Figure 4. General verification of the subject depth decay pattern

Figure 4 depicts the systematization of decay, conceptualized in two dimensions: the popularity quantiles and the topic's depth score. The horizontal axis plots measure the four popularity levels, ranging from Q1 to Q4, and the vertical axis charts the standard depth score. Each of the fifty topics is shown at its individual level of evolution using semi-transparent lines, but the overall distribution shows a clear downward-sloping trend, with the average depth displaying a decrease from 0.85 in Q1 to 0.40 in Q4, sustaining a 52.9% cumulative decline in depth. Systematic decay is a natural outcome of the attention allocation functions of short video platforms, where algorithmic recommendation systems favor content that the algorithm knows content can quickly resonate with viewers. There is an inherent contradiction or conflict between the cognitive workload required for in-depth analysis and the demand for immediate feedback from short video platforms; creators are increasingly abandoning rational and critical narratives in favor of these popularity metrics. In fact, the distribution of rates of depth decay illustrates that the majority of topics show a negative topic trend and subjectively maintained an average of -0.147. Therefore, the dominance of cursory content has perpetuated itself as a universal structural feature of short video film and television reviews.

4.3 Causal relationship between alienation characteristics and communication effects

Once the systematic presence of sentiment alienation and subject depth attenuation has been established, the next step is an analysis of the internal mechanism between those alienated traits and communication effects and clarification of the causal relationship. To empirically substantiate the driving role of alienation on communication efficacy, a vector autoregression model is developed, and Granger causality test is conducted to examine the potential of reverse causation or spurious correlations. Utilizing time series data, this investigation iteratively models three principal variables - sentiment divergence, subject depth decay rate, and video completion rate - to test the lead-lag relationship between the three variables for statistical significance and showcase the results in a causal network diagram. The analysis is shown in Figure 5.

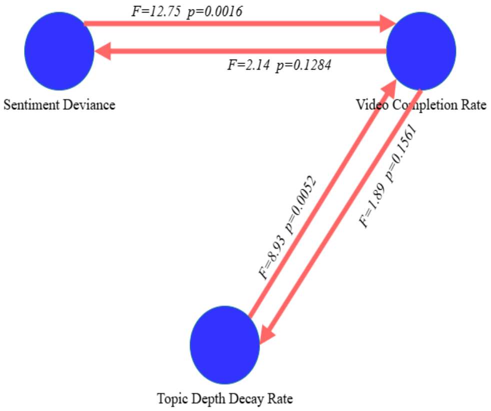

Figure 5. Causal relationship between alienation characteristics and communication effects

Figure 5 presents the empirical results of Granger causality tests in the form of a directed network, with nodes representing core variables and edges indicating statistically significant causal relationships. Sentimental divergence has the strongest explanatory power for the causal path to completion rate, with an F-statistic of 12.75 and a significance level of 0.0016, indicating that the dominant influence of sentiment expression on content can directly improve user retention. The topic depth decay rate also has a significant causal impact on completion rate, with an F-statistic of 8.93 and a p-value of 0.0052, indicating that strategies to reduce content depth generate positive returns in the context of algorithmic recommendation. Reverse causality tests fail to pass the significance threshold. The F-statistics for completion rate on sentimental divergence and depth decay rate are 2.14 and 1.89, respectively, with p-values greater than 0.05, ruling out the competing hypothesis that content characteristics are inversely shaped by diffusion effects. This one-way causal structure stems from the short video platform's ranking algorithm based on communication effectiveness metrics such as completion rate and number of likes. Through its traffic allocation mechanism, the system continuously rewards content that quickly elicits emotional resonance and shallow engagement. However, in-depth analytical content, which requires a higher cognitive investment and longer consumption time, is at a structural disadvantage in the communication logic of instant feedback.

The test results confirm that alienated characteristics are the cause, not the effect, of changes in communication effectiveness, establishing the causal driving mechanism of formal expression on the content ecosystem.

4.4 Importance of multimodal features

Once the causal association linking alienating features to communication effectiveness has been established, there arises the need for a further analysis of the relative contribution of content modalities to facilitate communication, in order to understand the functional mechanisms and how the alienating aspect of form overwhelms content. This section, through comparisons of multi-dimensional radar chart comparisons and quantitative assessments using ablation experiments, seeks to identify the degree to which textual, visual, and audio modalities contribute to a prediction of communication effectiveness and the distribution of weighting of sensory components within the content ecosystem.

Modal contributions are considered based on evaluations of static weighting distributions and dynamic removal impacts based on model performance considerations identified by feature importance scores. The findings are presented in Figure 6.

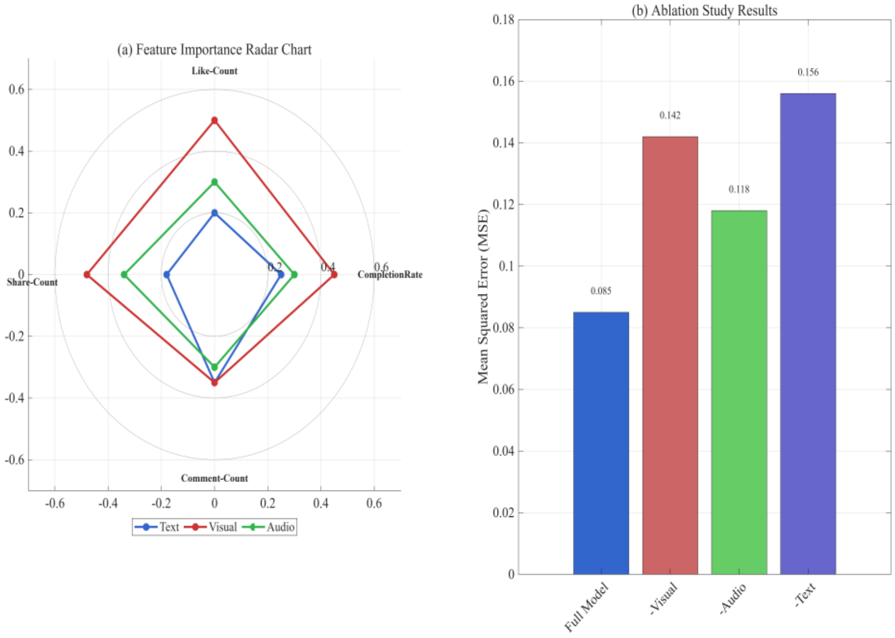

Figure 6. Multimodal feature contribution results: (a) Weight distribution of each modality in multiple prediction dimensions; (b) Ablation results

Figure 6(a) displays the element distribution of the three modalities across four prediction dimensions. The visual modality drives prediction of completion rate, likes, and shares with weights of 0.45, 0.50, and 0.48, respectively. The text modality demonstrates relatively high performance in prediction of the comment count dimension, with a weight of 0.35, but in the other three communication dimensions, the weights do not exceed 0.25. The audio modalities' weight remains stable at values between 0.30 and

0.34. This pattern of weights is a reflection of the short video platform's content consumption mechanism, which is focused on visual attention. Rapidly changing imagery and strong visual stimuli are much more likely to grab user attention and facilitate shallow interactive behaviors, especially in fragmented scenarios. However, text analysis carries a lot of information that, by its nature, requires that readers deeply process the information temperature which creates tension between the complex cognitive information and immediate feedback obligation.

The results of the ablation experiment, shown in Figure 6(b), support this conclusion. The mean squared error of the model increases from 0.085 to 0.142 relative to the original model (with visual featuring removal), a relative increase of 67.1%, and is higher than the mean squared error of 0.118 obtained with audio feature removal. As for comment context (i.e., textual feature removals), the mean squared error increases to an even higher level of error, as high as 0.156. The performance drop from visual feature removals supports the important role of visual contextual interactions for predicting interaction effectiveness. The mean squared error from comment interaction soars due to the fact the predictive power of comment interaction. The maximum importance weight for comments on predicting comment counts is 0.35. That means the ability to identify engaging, interactive, or deeply engaging content is undercut by simple comment context observations. This pattern in error distribution also points to an overall pattern of functional specialization among the modalities in the short video context. Visual context plays the most important role in gaining initial attention from viewers, while, limited textual (comment) context leads to engage deeply (interacting through liking, sharing, commenting etc.). Overall analysis suggests that when it comes to engagement in a short video context, the text (comment) context maintains a level of importance, engaging content in specific interactive contexts, with some degree of completion value. However, the visual context continues to dominate for the overall measure of communication effectiveness, which becomes a parameter for how formal expression keeps itself engaged with rational content in the short context of a video.

4.5 Cluster analysis of alienation types

After confirming the overall impact and causal mechanisms of alienation, it needs to further analyze the diverse structure within the alienation phenomenon and identify different types of communication strategies and their effectiveness characteristics. This section uses cluster analysis to identify patterns in video samples, revealing whether there are typical categories of alienation and the differential communication effects of each category. Based on two core alienation indicators—emotional deviation and topic depth decay rate—and combined with communication performance data, this study conducts unsupervised clustering and profiling analysis of video samples. The results are shown in Figure 7.

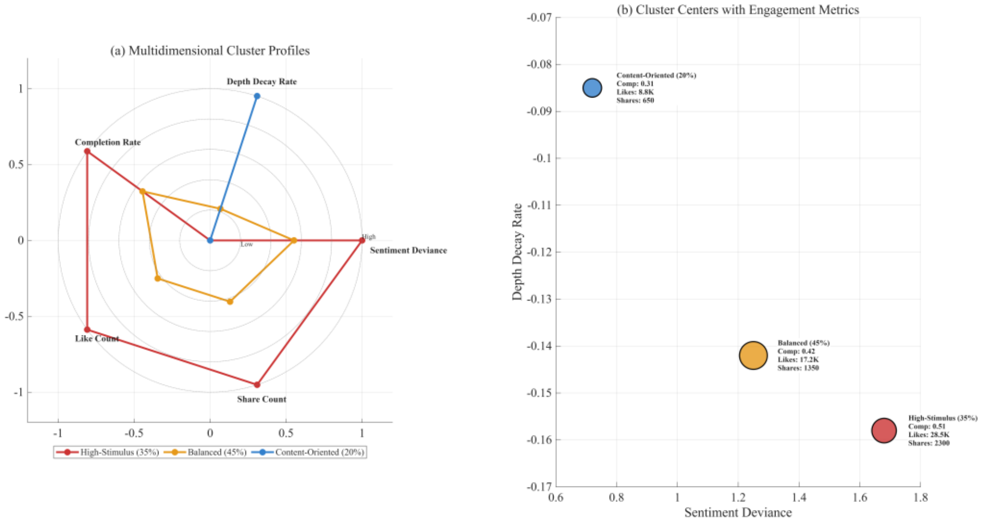

Figure 7. Cluster and profile analysis results

Figure 7 displays the normalized profile values and scatter plots for the three clusters across five dimensions. The high-stimulus cluster, depicted in the left subplot, is predominantly found in the peripheral region on each characteristic axis, with raw values of 1.68 and -0.158 for sentiment deviation and depth decay, respectively, with normalized values of 1. The balanced-dissimilar cluster is found to be intermediate across elements, and the content-oriented cluster is found closest to the innermost area of the radar chart. The pattern of distribution (overlying each other radially from high dissimilarity on the outside to low dissimilarity positioned in the middle) is indicative of a systematic relationship between the level of dissimilarity and effectiveness of the communication - with dissimilar content holding a holistic advantage in maintaining user attention and interacting with the content. The scatter plot, again in the right subplot, impels this classification. The clusters create a gradient related to sentiment deviation and depth decay space. The data reveals that the high-stimulus cluster has a completion rate of 0.51, 28.5 thousand likes, and 2.3 thousand shares, amply outpacing the content-centric cluster's completion rate of 0.31, 8.8 thousand likes, and 650 shares. This disparity in hierarchal organization arises from creators' differing interactions with algorithmic expectations. Platforms reward content that is increasingly shared through their traffic allocation, which in turn calls for creators (due to competitive forces) to slowly converge onto highly alienated strategies. Cluster analysis reveals that the short video comment ecosystem has thereby indexed a continuous spectrum based on alienation from content; with relative adherence to content on one end, formal performance on the other. Each category of strategy offers different potential returns in dissemination status, measure of value, and avenues for continuing to survive.

5.Conclusions

This research creates a computational framework that merges the analysis of multimodal sentiment divergence and dynamic topic modeling. By calculating the relative attention weights of audiovisual semiotic modalities on text sentiment, it identifies dissimilarities in communications and evaluates the chronology of content shallowness using the decay rate of topic depth. The empirical analysis has evidenced that the sentiment divergence positively predicts the video completion rate. For example, in the low dissimilarity range, the completion rate averages 0.28, whereas in the high dissimilarity range, the completion rate averages 0.49, thus verifying that formal expression continually drives communications effectiveness in that respect. Complementary Granger causality tests substantiate the unidirectionality of dissimilar features on communications effectiveness; in particular, the sentiment divergence significantly demonstrates a causal effect on the completion rate at the significance level. Finally, the multimodal feature importance evaluation indicates visual features are considerably important in the sample's communication mechanism with a 0.45 weight in predicting the completion rate, which is tremendously larger than the textual modality with 0.25; the disparity in values reflects the dissimilarity of "the form outweighs the content". Altogether, this research creates a computer analysis system that is based on a hybrid framework of multimodal sentiment divergence analysis and dynamic topic modeling, and potentially affords opportunities to develop a quantifiable approach in analyzing and synthesizing a video in literary and artistic reviews. The three types of alienation patterns diagnosed by it provide empirical basis for platform governance and content ecology optimization, and have methodological significance for constructing a value guidance path under the algorithm recommendation mechanism.

References

1.Li Man, Xie Lin. Research on short videos of film reviews under the background of new media. Journalism and Communications, 2022, 10: 224.

2.Liu Tianyi, Gao Shen. Research on the presentation characteristics and communication impact of short videos of film and television commentary: A case study of the Douyin platform. Journalism and Communications, 2025, 13: 619.

3.Gupta S, Deodhar S J, Tiwari A A, et al. How consumers evaluate movies on online platforms? Investigating the role of consumer engagement and external engagement. Journal of Business Research, 2024, 176: 114613.

4.Ameli S R, Farzaneh Siasi Rad F. Cultural Critique and Authority in the Digital Media Era: A Rhetorical Analysis of Instagram Vernacular Film Reviewers. Journal of Cyberspace Studies, 2025, 9(1): 107-125.

5.Javed M S, Safyan M, Manzoor A. Literature and Media: A Study of the Role of Literary Criticism in Shaping Communication in the Digital Age. AL-ĪMĀN Research Journal, 2025, 3(02): 16-36.

6.Lu S, Yu M, Wang H. What matters for short videos' user engagement: A multiblock model with variable screening. Expert Systems with Applications, 2023, 218: 119542.

7.Yang Q, Wang Y, Wang Q, et al. Harmonizing Sight and Sound: The Impact of Auditory Emotional Arousal, Visual Variation, and Their Congruence on Consumer Engagement in Short Video Marketing. Journal of Theoretical and Applied Electronic Commerce Research, 2025, 20(2): 69.

8.Wang C, Li Z. Unraveling the relationship between audience engagement and audiovisual characteristics of automotive green advertising on Chinese TikTok (Douyin). Plos one, 2024, 19(4): e0299496.

9.Yu X, Haroon M, Menchen-Trevino E, et al. Nudging recommendation algorithms increases news consumption and diversity on YouTube. PNAS nexus, 2024, 3(12): pgae518.

10.1Xu J. Analysis of social media algorithm recommendation system. Studies in Social Science & Humanities, 2022, 1(3): 57-63.

11.Wang L, Che G, Hu J, et al. Online review helpfulness and information overload: the roles of text, image, and video elements. Journal of Theoretical and Applied Electronic Commerce Research, 2024, 19(2): 1243-1266.

12.Zhu Z, Liu S, Zhang R. Examining the persuasive effects of health communication in short videos: systematic review. Journal of medical Internet research, 2023, 25: e48508.

13.1Alfayad K, Murray R L, Britton J, et al. Content analysis of Netflix and Amazon Prime Instant Video original films in the UK for alcohol, tobacco and junk food imagery. Journal of Public Health, 2022, 44(2): 302-309.

14.Fazeli S, Sabetti J, Ferrari M. Performing qualitative content analysis of video data in social sciences and medicine: The visual-verbal video analysis method. International Journal of Qualitative Methods, 2023, 22: 16094069231185452.

15.Caschera M C, Grifoni P, Ferri F. Emotion classification from speech and text in videos using a multimodal approach. Multimodal Technologies and Interaction, 2022, 6(4): 28.

16.Qiu K, Zhang Y, Zhao J, et al. A multimodal sentiment analysis approach based on a joint chained interactive attention mechanism. Electronics, 2024, 13(10): 1922.

17.Zhu H, Wei H, Wei J. Understanding users' information dissemination behaviors on Douyin, a short video mobile application in China. Multimedia Tools and Applications, 2024, 83(20): 58225-58243.

18.Zannettou S, Nemeth O N, Ayalon O, et al. Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations. arXiv e-prints, 2023, 23(01): 04945.

19.Sihag M, Shi Li Z, Dash A, et al. A Data-Driven Approach for Finding Requirements Relevant Feedback from TikTok and YouTube. arXiv e-prints, 2023, 23(05): 01796.

20.2Corso F, Pierri F, De Francisci Morales G. What we can learn from TikTok through its Research API. arXiv e-prints, 2024, 24(02): 13855.

21.Palomino M A, Aider F. Evaluating the effectiveness of text pre-processing in sentiment analysis. Applied Sciences, 2022, 12(17): 8765.

22.Dogra V, Verma S, Kavita, et al. A complete process of text classification system using state-of-the-art NLP models. Computational Intelligence and Neuroscience, 2022, 2022(1): 1883698.

23.Subakti A, Murfi H, Hariadi N. The performance of BERT as data representation of text clustering. Journal of big Data, 2022, 9(1): 15.

24.Hussain K, Ihsan I. Sentiment analysis in movie reviews using knowledge graph embeddings and deep learning classification. Journal of Computing & Biomedical Informatics, 2023, 6(01): 222-229.

25.Liu T, Yuan X. Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 23.

26.Chamishka S, Madhavi I, Nawaratne R, et al. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling. Multimedia Tools and Applications, 2022, 81(24): 35173-35194.

27.Vamsidhar D, Desai P, Shahade A K, et al. Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. Scientific Reports, 2025, 15(1): 25440.

28.Li Z, Liu P, Pan Y, et al. Text-dominant multimodal perception network for sentiment analysis based on cross-modal semantic enhancements. Applied Intelligence, 2025, 55(3): 188.

29.Diaf S, Fritsche U. Topic scaling: A joint document scaling–topic model approach to learn time-specific topics. Algorithms, 2022, 15(11): 430.

30.Miaozhu Zhang, born in 1975, holds a Master's degree in Literature and Art from Inner Mongolia Normal University and is a professor at Jining Normal University in Inner Mongolia. Her main research areas include literary theory research and work reviews, film and television art research and work reviews, and university discipline construction.

31. Jam, F. A., Ali, I., Albishri, N., Mammadov, A., & Mohapatra, A. K. (2025). How does the adoption of digital technologies in supply chain management enhance supply chain performance? A mediated and moderated model. Technological Forecasting and Social Change, 219, 124225.