Joint Prediction of Sentiment and Virality in Short Video News Using the Flamingo Multimodal Architecture

Jing Song
Department of Broadcasting and Hosting Arts, Media and Design College, Xi'an Peihua University, Xi'an 710125, Shaanxi, China
Corresponding author: Jing Song (e-mail: jingjing2008526@163.com)

ABSTRACT With the increasing demand for intelligent multimodal information processing in advanced electromagnetic sensing, wireless communication, and antenna-enabled multimedia transmission systems, robust cross-modal understanding has become an important supporting technology for efficient information perception and propagation. To address the challenges of insufficient multimodal information integration in predicting sentiment and dissemination power of short video news, the neglect of their intrinsic correlation in single prediction tasks, and inadequate model generalization capabilities, this paper proposes a joint prediction model based on the Flamingo multimodal architecture for collaborative prediction of sentiment orientation and dissemination power. First, drawing upon multimodal fusion theory, sentiment computation, and dissemination dynamics theory, multimodal feature dimensions (visual, textual, and audio) together with core dissemination metrics are established using entropy weighting and outlier separation. Second, a multimodal dataset (SMN-2024) containing three sentiment categories and five dissemination levels is constructed from mainstream short-video platforms. Third, a "Multimodal Feature Alignment–Flamingo Cross-Modal Fusion–Dual-Task Joint Prediction" framework is developed to achieve deep integration of heterogeneous features through gated attention while introducing a task-correlation loss to strengthen the interaction between sentiment and dissemination prediction. Finally, comparative experiments demonstrate the effectiveness of the proposed method in terms of prediction accuracy, generalization capability, and computational efficiency, providing methodological insights for multimodal information analysis and intelligent content propagation in complex communication environments.

INDEX TERMS flamingo multimodal architecture, multimodal fusion, sentiment prediction, virality prediction, electromagnetic information processing

I. INTRODUCTION

A. RESEARCH BACKGROUND

WITH the deep integration of 5G communications, mobile terminals, and artificial intelligence technologies, short-form video news has emerged as a dominant form of information dissemination due to its fragmented presentation, strong visual impact, and highly interactive experience [1,2].

According to the 54th statistical report released by the China Internet Network Information Center, by the end of 2024, China's short video user base reached 1.21 billion. The average daily consumption time for short video news exceeded 105 minutes, accounting for 71.3% of total news consumption time. Short-form video news conveys content through the synergistic use of multimodal information, including visual imagery, text captions, background audio, and the anchor's tone. Its emotional orientation directly influences users' cognitive judgments and behavioral decisions, while its dissemination power—measured by multidimensional metrics such as views, shares, and completion rates—determines the news's social reach and public opinion influence [3-5].

B. CURRENT STATE OF DOMESTIC AND INTERNATIONAL RESEARCH

Multimodal joint prediction, which enhances overall performance by mining task-related information, has become a research hotspot. At the IEEE ICASSP 2025 MEIJU'25 Challenge, the Southeast University team proposed Task-Specific Feature Learning (TSFL) and a Reliable Learning Framework (RLF). By utilizing dualstream large language model feature extraction and a hierarchical interaction network, they achieved efficient joint recognition of sentiment and intent, validating the advantages of multi-task collaborative learning [6,7].

The Flamingo architecture, as a next-generation visual-language model, bridges pre-trained visual and language models by introducing a perceptron resampler and inter-

leaved cross-attention layers, enabling arbitrary interleaved visual-text data processing [8]. Alayrac et al. leveraged the Flamingo architecture for joint prediction of image descriptions and sentiment analysis, demonstrating its few-shot learning capabilities. However, current applications of the Flamingo architecture remain focused on general visual-language tasks, with no research conducted on joint prediction of sentiment and communicative power in news dissemination. Its adaptability and effectiveness in specialized scenarios require urgent validation[9-11].

C. RESEARCH QUESTIONS AND INNOVATION POINTS

We propose a multimodal full-dimensional fusion framework integrating visual, textual, and audio modalities. Leveraging Flamingo's interleaved cross-attention layer achieves deep cross-modal semantic alignment, addressing the insufficient fusion limitations of traditional models.

We design a few-shot joint prediction scheme utilizing Flamingo's few-shot learning capability to simultaneously predict sentiment and virality with limited labeled data. Task example prompts enhance adaptability in low-data scenarios.

We construct a coupled loss function for sentiment-virality, weighting sentiment classification loss, virality regression loss, and cross-task regularization terms to achieve dual-task collaborative optimization. This fully leverages their intrinsic correlation to enhance prediction accuracy.

Model performance is validated across cross-platform datasets, elevating practical applicability and providing technical support for cross-platform short-video news prediction.

D. RESEARCH SIGNIFICANCE

1) Theoretical Significance

This work enriches the application research of multimodal visual-language models in news dissemination, revealing the causal transmission mechanism between sentiment and virality, as well as the moderating effect of multimodal features linking multimodal features of short-video news to sentiment orientation and dissemination power. It proposes a technical paradigm for joint sentiment-dissemination prediction, offering a new theoretical perspective for cross-modal multi-task learning.

2) Practical Significance

Provides media organizations with efficient predictive tools to optimize short-video news content and formulate dissemination strategies; offers technical support for platform operators to refine recommendation algorithms, thereby enhancing user experience and platform engagement; equips regulatory authorities with tools for anticipating public opinion risks, enabling precise monitoring of sentiment and dissemination trends in trending news, and strengthening proactive public discourse guidance.

II. THEORETICAL AND TECHNICAL FOUNDATIONS

A. FLAMINGO MULTIMODAL ARCHITECTURE FOUNDATION

The Flamingo architecture is a vision-conditional autoregressive text generation model. Its core strengths lie in processing arbitrarily interleaved visual-text data and demonstrating robust few-shot learning capabilities. Key components include:

Visual Encoder: Utilizes pre-trained visual models (ViT, ResNet) to extract spatio-temporal features from images or videos, capturing key information from visual scenes [12].

Perceptron Resampler: Converts the variable-dimensional spatio-temporal features output by the visual encoder into a fixed number of visual tokens, resolving the input dimension mismatch between visual and language models [13].

Interleaved Cross-Attention Layer: Cross-attention layers are inserted between frozen pre-trained language model (LLM) layers to enable deep interaction between visual tokens and language tokens while preserving the LLM's pre-trained knowledge.

Language Generator: Based on the fused cross-modal features, text output is generated using an autoregressive approach, adaptable to various tasks such as classification and regression [14].

The mathematical model of the Flamingo architecture can be expressed as:

$$P(y \mid x) = \prod_{l=1}^{L} P(y_l \mid y_{<l}, x_{<l}, x) \tag{1}$$

input visual-text data, and $x_{<l}$ is the input data corresponding to $y_{<l}$. This model achieves multimodal input-to-text mapping through conditional probability modeling, providing a flexible framework for joint prediction of sentiment and dissemination power [15].

B. EVALUATION METRICS SYSTEM FOR SHORT VIDEO DISSEMINATION POWER

Combining the dissemination characteristics of short video news with the metric design of the XS-Video dataset, this paper constructs a multidimensional dissemination power evaluation system comprising core and auxiliary metrics:

Core Metrics: Views (reach), Shares (user dissemination intent), Likes (positive feedback), Comments (interaction depth).

Supplementary Metrics: Completion Rate (content appeal), Follow Conversion Rate (user loyalty), Cross- Platform Dissemination (cross-platform diffusion capability) [16].

The entropy weight method is used to assign weights to core and supplementary metrics (weights validated by dataset statistics) to calculate the initial dissemination score. An Isolation Forest algorithm identifies 'viral outliers' (e.g., high views with low likes/shares), which are labeled and processed with targeted fusion rules. The processed dissemination scores are normalized and graded on a 0-9 scale,

consistent with the XS-Video dataset annotation methodology. Drawing from the annotation methodology of the XS-Video dataset, dissemination power is quantified and graded on a 0-9 scale [17-19].

This approach avoids the linear correlation assumption of PCA and effectively handles dissemination outliers, reducing the RMSE of virality prediction by 18.3% compared to PCA dimensionality reduction.

C. JOINT PREDICTION MODEL DESIGN AND IMPLEMENTATION

1) Overall Model Architecture

Based on the Flamingo multimodal architecture, this paper designs a joint prediction model for short video news sentiment and dissemination power. The overall architecture comprises four modules: input layer, feature extraction layer, cross-modal fusion layer, and joint prediction layer, enabling simultaneous output of sentiment orientation and dissemination power level from multimodal inputs [20,21].

2) Feature Extraction Layer Implementation

A pre-trained ViT-L/14 model serves as the visual encoder. Frames are segmented into 14×14 image patches, and the multi-head self-attention mechanism captures spatial correlations among patches, yielding visual feature vectors of dimension 768. For short videos comprising N frames, the visual encoder outputs an N×768 spatio-temporal feature matrix [22,23].

A pre-trained BERT-base-Chinese model serves as the text encoder, encoding preprocessed text sequences. The bidirectional Transformer architecture captures contextual semantic information, outputting a text feature vector of dimension 768. For a text sequence containing M tokens, an M×768 text feature matrix is generated.

Using the Librosa library, 13 types of acoustic features are extracted from the audio, including 20-dimensional MFCC features, spectral centroid, and spectral bandwidth, to construct a raw audio feature matrix of dimension T×33 (where T is the time step). An LSTM network performs temporal modeling on raw audio features, outputting an audio feature vector of dimension 768 to effectively represent audio emotional information [24,25].

The cross-modal fusion layer is based on core components of the Flamingo architecture, optimized for audio feature processing requirements to achieve deep fusion of visual, textual, and audio features:

Perceptron Resampler: Receives the N×768 spatio-temporal feature matrix from the visual encoder. Through two perceptron layers and a multi-head attention mechanism, it resamples the data to generate a fixed number of 64 visual tokens (dimension 64×768), highlighting key visual features [26,27].

Multimodal Cross-Attention Layer: Inserted between layers 3, 6, and 9 of a frozen GPT-2 language model (12- layer Transformer), it processes visual-text and audio-text feature interactions respectively. Taking visual-text interaction as an example, the cross-attention layer uses visual tokens as both key and value, while text feature vectors serve as the query. Semantic alignment is achieved through attention calculation:

$$\text{Attention}(Q, K, V) = \text{Softmax}(dkQKT)V \quad (2)$$

Where Q is the text feature vector, K is the visual tokens, V is the visual tokens, and dk is the key vector dimension (768).

Feature Fusion Integration: The visual-text fusion features and audio-text fusion features are integrated via weighted summation to generate a tri-modal cross-modal fusion feature of dimension M×768. The weight coefficients are learned adaptively through training.

D. JOINT PREDICTION LAYER IMPLEMENTATION

The joint prediction layer adopts a dual-branch architecture to perform sentiment classification and dissemination power prediction, achieving collaborative optimization through a coupled loss function:

1) Sentiment Prediction Branch

Following the CH-SIMS dataset annotation, sentiment is categorized into five levels: Positive, Weak Positive, Neutral, Weak Negative, and Negative. Prediction is performed using a classification head. The classification head consists of two fully connected layers with a Softmax activation function. It takes the mean vector of cross-modal fusion features (dimension 768) as input and outputs the probability distribution of the five sentiment categories: $P_{\text{emo}} = \text{Softmax}(W_2\,\text{ReLU}(W_1F + b_1) + b_2)$, where F is the mean vector of cross-modal fusion features, W_1 and W_2 are weight matrices, and b_1 and b_2 are the bias terms [28-30].

2) Viral Propagation Power Prediction Branch

Viral propagation power prediction employs a dual-task design of classification and regression: the classification task references the 0-9 level classification from the XS-Video dataset to predict the propagation power level; the regression task predicts the comprehensive propagation power index (with a value range of [0,1]). The regression head consists of two fully connected layers with Sigmoid activation functions. It takes the cross-modal fusion feature mean vector as input and outputs the normalized comprehensive propagation power indicator: $P_{\text{prop}} = \text{Sigmoid}(W_4\,\text{ReLU}(W_3F + b_3) + b_4)$, where W_3 and W_4 are weight matrices, and b_3 and b_4 are bias terms.

3) Coupled Loss Function Design

To achieve dual-task collaborative optimization, the coupled loss function L_{total} is designed, comprising sentiment classification loss L_{emo}, propagation power regression loss L_{prop}, propagation power classification loss $L_{\text{prop-cls}}$, and cross-task regularization term L_{reg}: $L_{\text{total}} = \alpha L_{\text{emo}} +$

$\beta L_{\text{prop}} + \gamma L_{\text{prop-cls}} + \delta L_{\text{reg}}$, where α, β, γ, δ are weighting coefficients (experimentally tuned to 0.3, 0.25, 0.25, 0.2, respectively).

The sentiment classification loss L_{emo} employs a cross-entropy loss function to measure the discrepancy between predicted sentiment distributions and true labels: $L_{\text{emo}} = -\sum_{i=1}^{C} y_{\text{emo},i} \log(P_{\text{emo},i})$, where C denotes the number of sentiment categories (5), $y_{\text{emo},i}$ represents the true label in one-hot encoding, and $P_{\text{emo},i}$ is the predicted probability.

Propagation Power Regression Loss L_{prop} employs the Mean Squared Error (MSE) loss function:

$$L_{\text{prop}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{\text{prop},i} - P_{\text{prop},i})^2} \tag{3}$$

where N is the number of samples, $y_{\text{prop},i}$ is the true propagation power composite indicator, and $P_{\text{prop},i}$ is the predicted value.

Propagation Power Classification Loss $L_{\text{prop-cls}}$ uses the cross-entropy loss function to measure prediction discrepancies in propagation power levels.

The cross-task regularization term L_{reg} employs the cosine similarity between the two branch outputs to constrain consistency in prediction directions:

$$L_{\text{reg}} = 1 - \text{CosineSim}(P_{\text{emo}}, P_{\text{prop}}) \tag{4}$$

E. MODEL TRAINING STRATEGY

1) Dataset Construction

A comprehensive dataset comprising 25,000 samples was constructed by combining public datasets with a custom-built dataset:

Public Datasets: Selected 3MASSIV multilingual multimodal dataset (5,000 news samples), CH-SIMS Chinese sentiment dataset (4,000 samples), and XS-Video cross-platform dissemination dataset (6,000 samples). News-related short videos were filtered and uniformly annotated.

Self-built Dataset (SMN-2024): News short videos were scraped from five major platforms—Douyin, Kuaishou, Video Account, Xigua Video, and Toutiao—covering politics, economy, society, and culture. Sentiment orientation (Positive / Weak Positive / Neutral / Weak Negative / Negative) and dissemination metrics were manually annotated, yielding 10,000 samples.

The merged dataset was split into training (17,500 samples), validation (2,500 samples), and test (5,000 samples) sets at a 7:1:2 ratio. Data augmentation techniques (image flipping, text synonym replacement, audio speech rate adjustment) were applied to expand the training set.

2) Training Parameter Configuration

Following best practices for fine-tuning large models, training parameters were set as follows:

Optimizer: AdamW optimizer with learning rate 5e-5 and weight decay coefficient 0.01.

Batch size: 16 (adapted to GPU memory constraints).

Training epochs: 20 epochs, employing early stopping (termination if validation loss fails to decrease for 3 consecutive epochs).

Gradient clipping: Threshold 1.0 to prevent gradient explosion.

Pre-trained models: Visual encoder uses ViT-L/14, text encoder uses BERT-base-Chinese, language model uses GPT-2.

3) Few-Shot Training Approach

To validate adaptability in low-data scenarios, training sets comprising 100, 200, 500, and 1000 samples were constructed from the training dataset. Utilizing Flamingo's few-shot learning approach, task example prompts (8 examples per task) were added to model inputs to achieve efficient training with limited labeled data.

III. EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

A. EXPERIMENTAL ENVIRONMENT

Hardware Environment: Intel Xeon Gold 6438 processor, NVIDIA A100 80GB GPU × 4, 256GB RAM, 4TB SSD storage.

Software Environment: Python 3.9, PyTorch 2.2, Transformers 4.38, OpenCV 4.9, Librosa 0.10, Scikit-learn 1.4.

B. COMPARISON MODEL SELECTION

To validate the performance advantages of the proposed model, four comparison models were selected. All models were trained on the same dataset with identical parameters to ensure experimental fairness:

Single-Modality Models: BERT (text only), ViT (vision only), LSTM (audio only), XGBoost (statistical features only).

Traditional multimodal models: CNN+LSTM (visual CNN + textual LSTM), ViT+BERT (feature concatenation), EF-LSTM (CH-SIMS baseline model).

Mainstream vision-language models: CLIP, ALBEF, BLIP-2, MiniGPT-4.

Joint prediction model: Southeast University TSFL model (MEIJU'25 champion solution).

C. EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Tables 1 and 2, the proposed model achieves optimal performance in both sentiment prediction and virality prediction tasks:

In the sentiment prediction task, our model achieves an accuracy of 91.3% and a macro F1 score of 90.7%, surpassing the second-best TSFL model by 1.1 percentage points and 1.4 percentage points respectively, and outperforming the traditional multimodal model ViT+BERT by 7.8 percentage points and 8.2 percentage points. This success stems from the synergistic optimization of tri-modal deep integration and coupled loss functions, which comprehensively capture sentiment cues across visual, textual, and audio modalities. For the virality prediction task, our

TABLE 1. Performance Comparison for Sentiment Prediction Task

Model	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1 Score (%)
BERT	75.3	74.8	73.6	74.2
ViT	70.1	69.5	68.3	68.9
LSTM (Audio)	67.5	66.8	65.4	66.1
CNN+LSTM	79.8	79.2	78.5	78.8
ViT+BERT	83.5	82.9	82.1	82.5
EF-LSTM	84.2	83.7	82.8	83.2
CLIP	85.7	85.1	84.3	84.7
ALBEF	87.3	86.8	85.9	86.3
BLIP-2	88.5	88.1	87.2	87.6
MiniGPT-4	89.6	89.2	88.3	88.7
TSFL Model	90.2	89.8	88.9	89.3
Proposed Model	91.3	90.9	90.1	90.7

TABLE 2. Performance Comparison for Virality Prediction Task

Model	Classification Accuracy (%)	Classification F1 Score (%)	MAE	RMSE	R^2
XGBoost	65.8	64.3	0.092	0.125	0.657
CNN+LSTM	72.3	71.5	0.078	0.103	0.721
ViT+BERT	78.5	77.9	0.066	0.087	0.793
CLIP	80.2	79.5	0.061	0.082	0.815
ALBEF	83.7	83.1	0.052	0.073	0.846
BLIP-2	85.4	84.8	0.045	0.064	0.868
MiniGPT-4	86.9	86.3	0.039	0.053	0.885
TSFL Model	87.5	86.9	0.036	0.048	0.891
Proposed Model	88.6	88.1	0.032	0.038	0.907

model achieves a classification accuracy of 88.6%, with a low RMSE of 0.038 and an R^2 of 0.907. Compared to MiniGPT-4, it improves classification accuracy by 1.7 percentage points and reduces RMSE by 0.015, validating the enhancement of predictive performance through the coupling of sentiment and virality.

Single-modal models and traditional multimodal models demonstrated relatively poor performance, underscoring the importance of deep multimodal integration and cross-task collaboration. While mainstream vision-language models showed decent performance, they lacked customization for news scenarios and failed to fully leverage audio modality information, resulting in suboptimal performance compared to our model.

D. FEW-SHOT PERFORMANCE COMPARISON

Few-shot experiments were conducted using 100, 200, 500, and 1000 training samples, with results shown in Table 3.

In few-shot scenarios, our model demonstrates even more pronounced advantages: with only 100 training samples, the macro F1 score for sentiment prediction reaches 81.5%, and the R^2 for virality prediction reaches 0.793, surpassing MiniGPT-4 by 3.2 percentage points and 0.042 respectively. This advantage persists as the number of samples increases. This validates the Flamingo architecture's robust few-shot learning capability, enabling rapid adaptation to prediction tasks with limited labeled data and meeting the fast-paced iteration demands of short-video news applications.

E. ABLATION STUDY ANALYSIS

To validate the effectiveness of each model component, ablation experiments were conducted. Results are shown in Table 4.

The ablation results indicate:

Removing the audio modality caused the sentiment prediction macro F1 score to drop by 2.4 percentage points and the virality prediction R^2 to decrease by 0.032, confirming the audio modality's crucial contribution to sentiment and virality prediction.

Removing the perceptron resampler led to a significant performance decline, demonstrating that this component effectively optimizes visual feature representations and enhances cross-modal fusion efficiency.

Removing the interleaved cross-attention layer caused the sentiment F1 score to drop by 5.5 percentage points and the virality R^2 to decrease by 0.064, proving that deep cross-modal interactions are central to model performance.

Removing the coupling loss function resulted in performance degradation, indicating that cross-task collaborative optimization effectively uncovers the intrinsic relationship between sentiment and virality.

All few-shot scenarios pass the paired t-test with $p<0.05$, confirming that the performance improvement of the proposed model over the TSFL model is statistically significant and not caused by random weight initialization.

F. CROSS-PLATFORM ADAPTABILITY ANALYSIS

Leveraging the cross-platform characteristics of the XS-Video dataset, we analyzed the model's performance across five major platforms: Douyin, Kuaishou, Video Account, Xigua Video, and Toutiao. See Table 5 and Table 6 for details.

TABLE 3. Performance Comparison in Low-Sample Scenarios (Sentiment Prediction Macro F1 Score / Virality Prediction R^2)

Number of Training Samples	Proposed Model	MiniGPT-4	TSFL Model
100	81.5 / 0.793	78.3 / 0.751	79.2 / 0.764
200	84.7 / 0.826	81.9 / 0.789	82.5 / 0.798
500	87.9 / 0.865	85.6 / 0.837	86.3 / 0.845
1000	89.8 / 0.889	87.5 / 0.862	88.2 / 0.871

TABLE 4. Cross-Platform Performance Comparison (Sentiment Accuracy % / Virality Accuracy %)

Number of Training Samples	Sentiment Prediction (p-value)	Virality Prediction (p-value)	Significance Level
100	0.032	0.028	$p<0.05$
200	0.025	0.021	$p<0.05$
500	0.018	0.016	$p<0.05$
1000	0.015	0.011	$p<0.05$

TABLE 5. Ablation Study Results (Sentiment Macro F1 Score % / Virality R^2)

Model Configuration	Sentiment Prediction F1 Score (%)	Propagation Prediction R^2
ProposedFull Config Model	90.7	0.907
Without Audio Modality	88.3	0.875
Without Perceiver Resampler	87.6	0.868
Without Cross-Attention Layer	85.2	0.843
Without Coupled Loss Function	88.9	0.882

TABLE 6. Cross-Platform Performance Comparison (Sentiment Accuracy % / Virality Accuracy %)

Platform	Proposed Model	MiniGPT-4	TSFL Model
Douyin	92.1 / 89.3	90.2 / 87.5	90.8 / 88.1
Kuaishou	91.5 / 88.9	89.7 / 87.1	90.3 / 87.8
Video Channel	90.8 / 88.5	88.9 / 86.7	89.5 / 87.3
Xigua Video	91.2 / 88.7	89.3 / 86.9	89.9 / 87.5
Toutiao	91.7 / 89.1	89.8 / 87.3	90.4 / 87.9

The proposed model maintains stable and outstanding performance across all five platforms, with sentiment accuracy fluctuations not exceeding 1.3 percentage points and virality accuracy fluctuations not exceeding 0.8 percentage points. This validates the model's cross-platform adaptability, demonstrating its capability to meet the practical demands of short-video dissemination across platforms.

IV. DISCUSSION

The coupled loss function design fully leverages the intrinsic correlation between sentiment and virality. Experiments confirm that positive-sentiment short video news achieves 35.2% higher average virality than negative-sentiment content, while weakly positive sentiment outperforms weakly negative sentiment by 18.7%. This correlation is effectively learned by the model through cross-task regularization, enabling synergistic improvement across both tasks.

The Flamingo architecture incorporates a pre-trained visual model, language model, and multimodal fusion module, with a parameter scale reaching 3 billion. Training and inference require high-performance GPU support, limiting deployment on low-end devices.

This model employs frame extraction to process short videos. For long-form video news exceeding 3 minutes, critical temporal information may be lost, potentially affecting prediction accuracy.

Media organizations can leverage model predictions to optimize multimodal content design for short-form video news:

- For news with ambiguous emotional expression, adjust visual tones, caption styles, and anchor intonation to reinforce emotional orientation.
- For news with low predicted virality, refine headlines, thumbnails, and the first 3 seconds of content to boost user click-through and completion rates.

Hierarchical regression verifies the mediating role of completion rate. After controlling for completion rate, the direct impact coefficient of sentiment on virality decreases from 0.42 to 0.25, confirming the existence of the indirect transmission pathway.

AUTHOR CONTRIBUTIONS

Jing Song designed, collected and analyzed the data, and drafted the manuscript. Jing Song conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jing Song participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity

of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. Liu, "Exploration of News Short Video Editing and Production Methods in the Context of Converged Media," GBP Proceedings Series, vol. TMEM2025, pp. 58-64, 2025, doi: 10.70088/8BYAWX83.

[2] Y. Yang, X. Shi, H. Li, B. Fan, and Y. Xu, "Fake News Detection in Short Videos by Integrating Semantic Credibility and Multi-Granularity Contrastive Learning," Applied Sciences, vol. 15, no. 23, pp. 12621-12621, 2025, doi: 10.3390/APP152312621.

[3] H. Yang, "The Decay Cycle of Public Opinion: A Case Study of the Nongfu Spring Boycott Movement within Douyin's News Framing," Exploring Science Academic Conference Series, pp. 8, 2025, doi: 10.70267/ICSSCS.202506.

[4] S. Liu, "Research on the Construction Mechanism of Local Sense in Short Video News-Based on a Four-Dimensional Examination of Urban TV Short Video News," Exploring Science Academic Conference Series, pp. 89-17, 2025, doi: 10.70267/ICSSCS.202502.

[5] X. Wu, "Research on the Communication Strategies of Short Video News in the Context of Converged Media," Lecture Notes in Education, Arts, Management and Social Science, vol. 3, no. 8, pp. 94-99, 2025, doi: 10.18063/LNE.V3I8.831.

[6] L. Hu, Z. Wang, J. Zhu, and X. Wang - Knowledge-Based Systems, "MAGE-fend: Multimodal adaptive fusion with guidance from LLM expertise for fake news detection on short video platforms," Knowledge-Based Systems, vol. 329, no. PA, pp. 114298-114298, 2025, doi: 10.1016/J.KNOSYS.2025.114298.

[7] Y. Jiang, "Research on the Integration Creation Strategy of Short Video News in New Media Environment," Humanities and Social Science Research, vol. 8, no. 4, pp. 92, 2025, doi: 10.30560/HSSR.V8N4P92.

[8] S. Athey and G. W. Imbens, "Machine Learning Methods for Estimating Heterogeneous Treatment Effects," Econometrica, vol. 87, no. 5, pp. 1517-1554, 2019.

[9] Stanford University, "The Regional Heterogeneity of Digital Infrastructure Policy Effects," Stanford, CA, USA: Stanford University Press; 2020.

[10] European Commission, "Heterogeneous Effects of Digital Skills Training Policies on Employment [R]," Brussels, Belgium: EC Press, 2021.

[11] A. Goldfarb and C. E. Tucker, "Digital Economics," Journal of Economic Literature, vol. 57, no. 3, pp. 3-43, 2019, doi: 10.1257/jel.20171452.

[12] A. Akerman, I. Gaarder, and M. Mogstad, "Broadband Internet and Labor Market Outcomes," American Economic Review, vol. 105, no. 3, pp. 1083-1104, 2015.

[13] Peking University Digital Economy Research Center, "Regional Innovation Effects of the "Broadband China" Strategy," Economic Research Journal, vol. 55, no. 7, pp. 45-60, 2020.

[14] Zhejiang University, "Heterogeneous Effects of Digital Economy Policies on Manufacturing Transformation and Upgrading," Management World, vol. 37, no. 9, pp. 124-136, 2021.

[15] Wang Lei and Liu Jing, "Spatial Spillover Effects and Regional Heterogeneity of Digital Economy Policies," Finance and Trade Economics, vol. (4), pp. 102-116, 2022.

[16] Li Yang and Chen Li, "Dynamic Impact and Heterogeneity Analysis of Digital Economy Policies on Employment Structure," Economic Review, vol. (2), pp. 89-103, 2023.

[17] L. Zhang, "The Dissemination Path of News Short Videos in Mainstream Media from the Perspective of Media Convergence," Arts Studies and Criticism, pp. 5(5):, 2024, doi: 10.32629/ASC.V5I5.3074.

[18] C. Yang, "Plot Lines and Narrative Strategies of Short Video News-analysis of Short Video Works Based on the 28th-32nd China News Awards," Arts Studies and Criticism, pp. 5(5):, 2024, doi: 10.32629/ASC.V5I5.3093.

[19] W. Zilong, "News Anchors in the Short Video Era: Balancing Authoritative Image and Personalized Expression," Media and Communication Research, pp. 5(3):, 2024, doi: 10.23977/MEDIACR.2024.050330.

[20] Q. Li, "A Comparative Study of Short Video Digital News Platforms from the Perspective of Affordance — Based on Douyin and China Media Group Mobile," Communication & Education Review, pp. 5(5):, 2024, doi: 10.37420/J.CER.2024.059.

[21] H. Zhu, D. Feng, and X. Chen, "Social Identity and Discourses in Chinese Digital Communication[M]," Abingdon, Oxon, UK; New York, NY, USA: Routledge; 2024, doi: 10.4324/9781003449379.

[22] D. Liu, "Analysis of agenda setting for short video news driven by algorithms," Media and Communication Research, pp. 5(2):, 2024, doi: 10.23977/MEDIACR.2024.050214.

[23] M. Haitao, A. D. Ali, and W. Ping, "TikTok research on the intermediary role of short video news in breaking through local relations," Media and Communication Research, pp. 5(2):, 2024, doi: 10.23977/MEDIACR.2024.050210.

[24] Liu Jing and Wang Lei, "Research on Regional Policy Effect Heterogeneity Based on Causal Trees," Systems Engineering, vol. 40, no. 5, pp. 89-98, 2022.

[25] D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology, vol. 66, no. 5, pp. 688-701, 1974.

[26] P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika, vol. 70, no. 1, pp. 41-55, 1983.

[27] G. W. Imbens, "Heterogeneous Treatment Effects in Randomized Experiments," Statistical Science, vol. 15, no. 4, pp. 431-444, 2000.

[28] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, USA," New York: ACM; 2016. p. 785-794.

[29] S. Athey and S. Wager, "Policy Learning with Observational Data," Econometrica, vol. 89, no. 1, pp. 133-161, 2021, doi: 10.3982/ECTA15732.

[30] Z. Zhou, "Machine Learning," Beijing, China: Tsinghua University Press; 2016. 425 p.