Multidimensional Danmaku Analytics via a BERT-SVM Fusion Model

Ya Lin[*] LINYA@CQVTU.EDU.CN
School of Information Engineering, Chongqing Vocational and Technical University of Mechatronics, 1001 Biqing North Road, Bishan District, Chongqing, 402760 China

Xudong Zhang ZHANGXUDONG@CQVTU.EDU.CN
School of Information Engineering, Chongqing Vocational and Technical University of Mechatronics, 1001 Biqing North Road, Bishan District, Chongqing, 402760 China

Guangbin Peng PEN10@163.COM
School of Information Engineering, Chongqing Vocational and Technical University of Mechatronics, 1001 Biqing North Road, Bishan District, Chongqing, 402760 China

Xiang He HEXIANG@CQVTU.EDU.CN
School of Information Engineering, Chongqing Vocational and Technical University of Mechatronics, 1001 Biqing North Road, Bishan District, Chongqing, 402760 China

Yuanxia Deng 2481955850@QQ.COM
School of Information Engineering, Chongqing Vocational and Technical University of Mechatronics, 1001 Biqing North Road, Bishan District, Chongqing, 402760 China
[*]*Corresponding author*

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Danmaku (bullet comments), characterized by real-time interactivity, high concurrency, and textual fragmentation, present unique challenges for semantic analysis in film audience feedback research. To address the limitations of conventional methods in processing sparse short texts and imbalanced data distributions, this study proposes a BERT-SVM fusion model integrating BERT-based semantic representation with SVM classification, supplemented by SMOTE oversampling. Validated on 450,000 Danmaku comments from The Wandering Eart series, the framework achieves a sentiment classification accuracy of 92.6%. Furthermore, a multidimensional analysis pipeline is implemented, combining BERT embedding compression, KMeans clustering, and LDA topic modeling to systematically identify audience discussion themes. Experimental results demonstrate that The Wandering Earth 2 not only elicits a higher proportion of positive sentiment than its predecessor but also shifts thematic focus toward advanced sci-fi elements such as digital life and lunar crisis resolution. This work establishes an efficient analytical framework for large-scale Danmaku data, offering actionable insights to enhance narrative design and audience engagement strategies in the film industry.

Keywords: Danmaku analytics, BERT-SVM fusion, Imbalanced text classification, Multidimensional topic modeling

1. Introduction

Danmaku, as real-time user-generated short texts on video platforms, have emerged as a critical data source for audience feedback research in the film industry due to their immediacy, high concurrency, and fragmented nature. These characteristics, however, introduce two primary challenges for analytical methodologies: (1) Semantic Sparsity in Short Texts: Traditional sentiment classification approaches, such as lexicon-based methods or single machine learning models, struggle to capture contextual semantics from noisy and sparse Danmaku content. (2) Domain-Specific Topic Deviation: Existing topic modeling techniques often fail to align extracted themes with actual film narratives due to the lack of domain-specific knowledge integration.

Prior studies predominantly focus on static sentiment or thematic analysis of individual films (Chen and Qian, 2022; Wang, 2024), yet exhibit critical limitations in handling domain-specific challenges of Danmaku texts. For instance, Chen and Qian (2022) demonstrated that lexicon-based approaches and single machine learning models (e.g., SVM or logistic regression) struggle to capture contextual semantics from noisy and sparse short texts, particularly under high-concurrency scenarios. Similarly, Liu et al. (2020) identified a significant gap in aligning topic modeling outputs with film narratives, as traditional LDA implementations often generate domain-agnostic themes (e.g., generic terms like "scene" or "actor") due to the lack of domain-aware embedding constraints. Furthermore, longitudinal analysis of film series remains underexplored, with most studies (Liu et al., 2020; Zhang et al., 2022) confined to single-film datasets. These limitations highlight the necessity of integrating deep semantic representations (e.g., BERT) (Devlin et al., 2019) with adaptive classification frameworks to address both sparsity and domain alignment challenges.

To address these gaps, this study makes the following contributions: (1) Hybrid BERT-SVM Sentiment Classification: We propose an integrated framework that leverages BERT for deep semantic feature extraction and SVM for classification, enhanced by SMOTE (Synthetic Minority Oversampling Technique) oversampling to mitigate data imbalance (Pang and Lee, 2008). This model achieves 92.6% accuracy on a dataset of 450,000 Danmaku comments from The Wandering Earth series (hereafter TWE Series; TWE, 2019; TWE2, 2023). (2) Multidimensional Topic Analysis Pipeline:

By combining BERT embedding compression, KMeans clustering, and LDA (Latent Dirichlet Allocation) modeling, our framework systematically identifies audience discussion hotspots while minimizing domain-agnostic topic deviations. (3) Longitudinal Film Series Analysis: We conduct the first comparative study of sentiment distribution and thematic evolution between The Wandering Earth (TWE, 2019) and its sequel The Wandering Earth 2 (TWE2, 2023), revealing distinct shifts toward advanced sci-fi narratives such as digital life and lunar crisis resolution. These contributions collectively establish a scalable analytical framework for Danmaku data, offering data-driven insights to optimize film production strategies and audience engagement practices.

2. Methodology

2.1. Research Workflow

The proposed multidimensional analysis framework, illustrated in Figure 1, comprises four sequential stages: data acquisition, preprocessing, sentiment-topic modeling, and visualization. First, Danmaku data is collected from multiple video platforms (Tencent Video, iQiyi, Mango TV) (Wang, 2024) and subjected to deduplication, noise filtering, and text normalization. The cleaned dataset is then processed by two parallel modules: (1) Sentiment Analysis Module: A BERT-SVM fusion model enhanced by SMOTE oversampling for imbalanced data handling. (2) Topic Analysis Module: A pipeline integrating BERT embedding compression, KMeans clustering, and LDA topic modeling. Finally, the analytical results are visualized through temporal sentiment curves, thematic word clouds, and interactive topic maps, enabling comparative interpretation of audience feedback patterns across films.

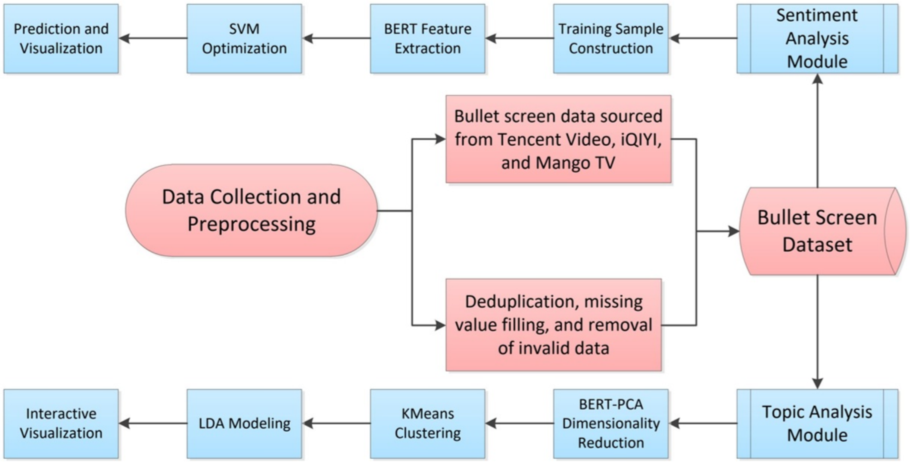

Prediction and Visualization
SVM Optimization
BERT Feature Extraction
Training Sample Construction
Sentiment Analysis Module
Bullet screen data sourced from Tencent Video, iQIYI, and Mango TV
Data Collection and Preprocessing
Bullet Screen Dataset
Deduplication, missing value filling, and removal of invalid data
Interactive Visualization
LDA Modeling
KMeans Clustering
BERT-PCA Dimensionality Reduction
Topic Analysis Module

Figure 1: Flowchart of the Multidimensional Danmaku Analysis Framework.

2.2. Data Collection and Preprocessing

A total of 450,000 raw Danmaku comments from the TWE Series were collected across three platforms using customized protocols. For iQiyi, encrypted Danmaku packets (.br files) were intercepted via browser developer tools, dynamically parsed based on 5-minute packet intervals and a 126-minute film duration (yielding 26 packets), and decoded through a pipeline involving .z conversion, Zlib decompression, and UTF-8 decoding. XPath extracted 43,286 entries (TWE (2019)) and 66,006 entries (TWE2 (2023)), stored as 6.15 MB CSV files. Tencent Video utilized API endpoints identified via “Segment” keyword filtering, with URL parameters iterated via arithmetic sequences (&start=0&end= $i*30000) to retrieve 129,582 (TWE (2019)) and 185,153 (TWE2 (2023)) entries, while Mango TV employed JSON interfaces updated per minute, requiring 126 loop iterations and segmented crawling via range headers to collect 17,549 (TWE (2019)) and 11,212 (TWE2 (2023)) entries. After deduplication (18.2% removed) and text normalization (whitespace trimming, punctuation removal), 380,000 entries were retained for analysis.

2.3. BERT-SVM Sentiment Analysis Model

Model Architecture. (1) BERT-based Semantic Encodin. We employ the ‘bert-base-chinese’ model to capture contextual semantics. To balance domain adaptation and computational efficiency, a partial fine-tuning strategy is applied: early layers are frozen to retain general linguistic knowledge, while later layers are fine-tuned to adapt to Danmaku text characteristics. Input texts are standardized to a fixed length for hardware compatibility. (2) SVM Classification with SMOTE. An RBF

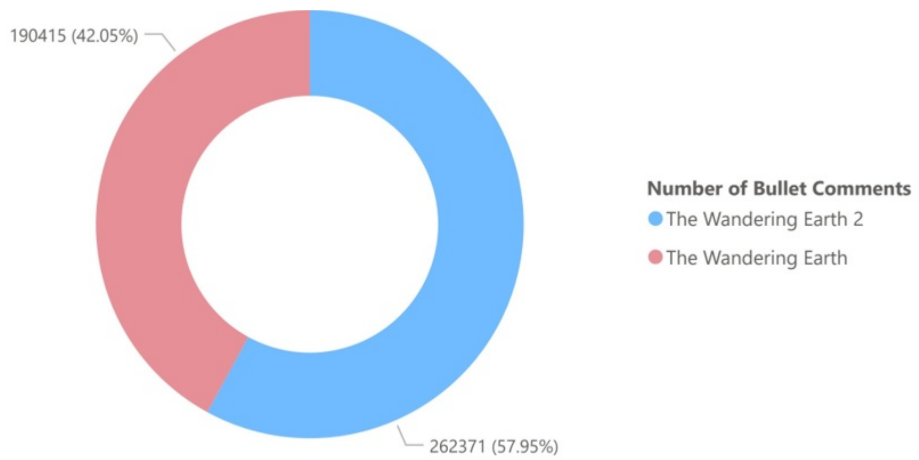


Figure 2: Total Raw Danmaku Volume for Both Films.

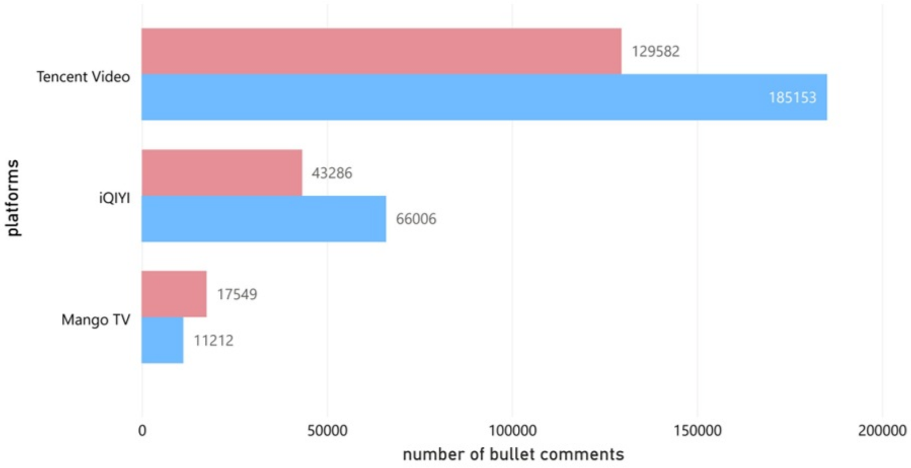


Figure 3: Platform-Wise Danmaku Distribution.

kernel SVM is optimized through grid search over regularization strengths (‘C‘) and kernel coefficients (‘gamma’), prioritizing minority class recall. Class weights are dynamically adjusted to counteract data imbalance. (3) Computational Optimization. Large-scale predictions are processed in batches to align with GPU memory constraints. Frequently occurring terms (e.g., film-specific nouns) are pre-encoded to accelerate real-time analysis. All critical experimental parameters (hyperparameters and implementation details) are summarized in Table 1.

2.4. Multidimensional Topic Analysis Workflow

The pretrained Chinese BERT model (bert-base-chinese) was employed to generate semantic embeddings for thematic analysis. Input texts were tokenized, padded/truncated to a uniform length of 128 tokens, and encoded into 768-dimensional sentence embeddings by averaging hidden states from the final transformer layer. To address high-dimensional sparsity, Principal Component Analysis (PCA) was applied to compress embeddings into 100 latent dimensions (n_components=100),

Table 1: Hyperparameter Summary.

Parameter category	Key parameters	Values/Ranges	Selection Basis
BERT Model	Pretrained Model	bert-base-chinese	Domain-general semantic capture
	Frozen Layers / Fine-tuned Layers	8 / 4	Balance adaptation & efficiency
	Max Sequence Length	128 tokens	GPU memory constraints
SVM Configuration	Kernel Type	RBF	Nonlinear classification
	Regularization (C)	1.0	Grid search (C∈0.1,1,10)
	Kernel Coefficient (gamma)	0.1	Grid search (gamma∈0.01,0.1,1)
Training	Optimizer	AdamW	Standard for Transformer models
	Learning Rate	2e-5	Literature baseline (Devlin et al., 2019)
	Batch Size (BERT)	32	GPU memory limits
SMOTE	Neighbors (k)	5	Empirical validation (Pang and Lee, 2008)
	Target Class Ratios	1:1:1	Mitigate data imbalance
Data Splitting	Train/Validation/Test	70%/15%/15%	Stratified sampling
Evaluation Metrics	Primary Metrics	Accuracy, Macro-F1, AUC-ROC	Imbalanced data requirements

preserving 90% of the cumulative variance while eliminating redundant noise. This dimensionality reduction optimized computational efficiency for downstream clustering tasks.

The compressed embeddings were partitioned into 10 thematic clusters using KMeans (n_clusters=10, random_state=42), with cluster counts determined by the elbow method. Intra-cluster semantic interpretation was achieved through a two-step process: (1) High-Frequency Term Extraction: Top 10 lexical items per cluster were identified based on term frequency (e.g., “planetary engine”, “oxygen mask”). (2) Contextual Labeling: Tokenized documents and frequency weights were combined to map clusters to narrative elements (e.g., “sci-fi scenario theme”).

A Latent Dirichlet Allocation (LDA) model (num_topics=10, passes=15, random_state=42) was trained on a doc2bow corpus constructed via the gensim library. The model inferred document-topic distributions and topic-word probability matrices. For interactive exploration, pyLDAvis generated a 2D topic map using Jensen-Shannon divergence and Principal Coordinate Analysis (PCoA), where bubble size indicated topic prominence and spatial proximity reflected semantic similarity, as shown in Figures 2-3. Dynamic controls allowed real-time adjustment of keyword relevance rankings through a λ-slider (0: balanced global-topic frequency; 1: pure intra-topic frequency).

The synthesized framework systematically quantifies audience sentiment dynamics and thematic evolution through three analytical dimensions. (1) Temporal Sentiment Analysis: Sentiment intensity (-1 to +1) was aggregated per minute to construct time series curves. (2) High-Frequency Term Visualization: Word clouds highlighted narrative elements with elevated discussion frequency (e.g., “digital life”, “lunar crisis”). (3) Topic Correlation Networks: Graph-based representations mapped inter-theme relationships, revealing audience focus shifts between film installments. This triangulated approach validates the framework’s efficacy in bridging computational analytics with cinematic narrative studies, offering granular insights into audience engagement patterns.

2.5. Comparative Experiments

Experimental Comparison. (1) Baseline Models. Traditional Methods: Naive Bayes (TF-IDF features), LSTM (300D GloVe embeddings), TextCNN (multi-scale kernels). State-of-the-Art: BERT-LSTM (fine-tuned BERT + LSTM), RoBERTa-wwm (domain-adapted pretraining). (2) Unified Evaluation. All models are tested on the same cleaned dataset (450K Danmaku) with stratified 7:1.5:1.5 split. Metrics include Accuracy, Macro-F1, and AUC-ROC (Negative class).

As shown in Table 2, BERT-SVM achieves: (1) ighest Accuracy (92.6%): Surpassing RoBERTa-wwm by 1.4%. (2) Optimal AUC-ROC (0.88): 5% higher than BERT-LSTM in Negative class detection. (3) Training Efficiency: 63% faster than BERT-LSTM (2.1h vs. 5.7h).

Table 2: Comparative Performance of Sentiment Models.

Model	Accuracy	Macro-F1	AUC-ROC (Negative)	Training Time (h)
Naive Bayes	78.3%	0.65	0.71	0.2
LSTM	84.1%	0.73	0.76	1.5
BERT-LSTM	90.5%	0.79	0.83	5.7
RoBERTa-wwm	91.2%	0.80	0.85	6.2
BERT-SVM(Ours)	92.6%	0.84	0.88	2.1

Key Insight: BERT-SVM achieves superior Negative class AUC-ROC (0.88) and 63% faster training than BERT-LSTM, demonstrating both efficacy and efficiency.

Interpretability Enhancement. (1) SHAP Analysis. SHAP value analysis reveals that the BERT-SVM model assigns strong positive contributions to certain emotionally charged terms. Words such as "shocking" and "teary-eyed" have the most significant impact on positive sentiment predictions, indicating the model's sensitivity to intense emotional expressions. (2) Attention Visualization. An examination of BERT's attention patterns shows a consistent focus on domain-specific nouns, such as "digital life", during positive sentiment classification. This highlights the model's ability to semantically align with the narrative context of the film, thereby enhancing the accuracy of sentiment interpretation. (3) Ablation Study. The fusion model outperforms BERT in minority class F1-score (0.81 vs. 0.72) and SVM in overall accuracy (92.6% vs. 85.6%), proving its complementary advantages.

3. Introduction

3.1. Sentiment Distribution

Figures 4 and 5 present the sentiment distribution of Danmaku comments for TWE (2019) and TWE2 (2023), respectively, derived from the BERT-SVM classification model. Key observations include: (1) Positive Sentiment Dominance: Positive sentiments constitute the largest proportion in both films (50.2% for TWE; 53.1% for TWE2), reflecting broad audience approval of cinematic elements such as visual effects and narrative pacing. The sequel's higher positive sentiment stems from technical innovations, which were frequently cited in thematic analyses. (2) Neutral Sentiment Prevalence: Neutral comments account for 29.6% (TWE) and 30.4% (TWE2), indicating that a subset of viewers maintained cautious perspectives. This likely arises from passive engagement during viewing and diverse interpretations of the films' complex thematic content. (3) Limited Negative Feedback: Negative sentiments remain relatively low (20.2% for TWE; 16.5% for TWE2), primarily focusing on narrative inconsistencies and uneven emotional pacing in climactic sequences.

3.2. High-Frequency Word Comparison

Word clouds generated from preprocessed Danmaku texts reveal distinct thematic emphases between the two films (Figures 6 and 7). (1) Shared Core Themes. High-frequency terms such as

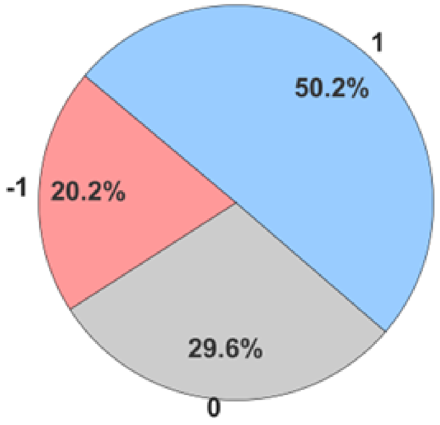


Figure 4: Sentiment Distribution of The Wandering Earth (TWE, 2019).

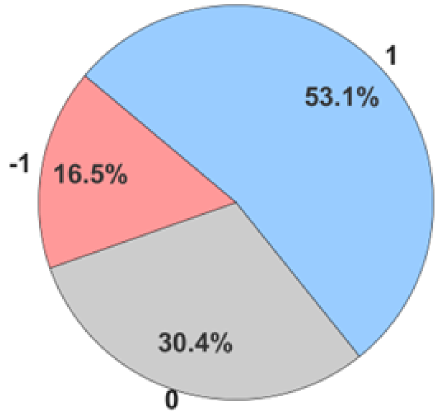


Figure 5: Sentiment Distribution of The Wandering Earth 2 (TWE2, 2023).

Earth and humanity dominate both films, underscoring the series' focus on planetary crisis resolution. Lexical items including engines, space station, and MOSS (AI system) highlight the technological backbone of the narrative. Actor Wu Jing and his character Liu Peiqiang remain central to audience discourse across both installments. (2) Divergent Focal Points. For TWE (2019) (Figure 6): Familial Bonds: Terms like child and underground city emphasize interpersonal relationships and survival logistics. Visual Spectacle Praise: Adjectives such as amazing and impressive dominate, reflecting acclaim for pioneering visual effects.

For TWE2 (2023) (Figure 7): Advanced Sci-Fi Concepts: Terms like digital and Moon signify deepened exploration of artificial intelligence ethics and lunar crisis arcs. Cultural Homage:

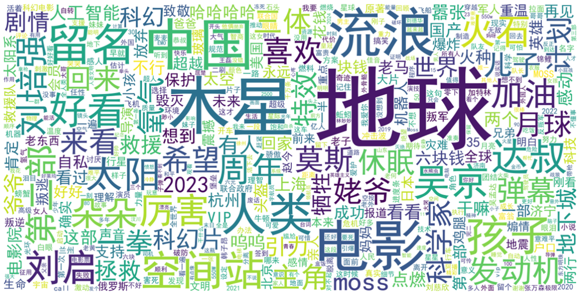


Figure 6: High-Frequency Word Cloud for TWE (2019).

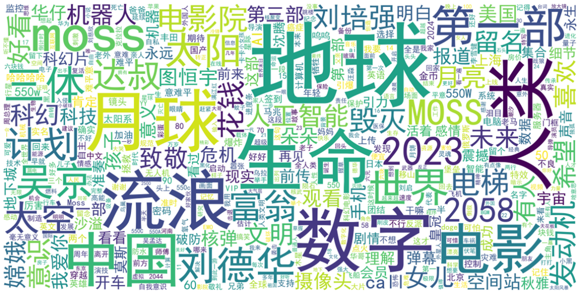


Figure 7: High-Frequency Word Cloud for TWE2 (2023).

Increased usage of tribute and ran (Chinese: ran, denoting intense excitement) indicates audience recognition of genre homage and the director's stylistic maturity.

3.3. Temporal Dynamics of Danmaku Activity

Figures 8 and 9 illustrate the temporal evolution of Danmaku interaction density aggregated at 10-minute intervals for The Wandering Earth (2019, hereafter TWE) and The Wandering Earth 2 (2023, TWE2). For TWE, The initial 30 minutes exhibit fluctuating Danmaku counts (1,200–1,800 entries/minute), reflecting viewers' adaptive engagement during early narrative immersion. A climactic peak of 1,950 entries/minute occurs at 01:55:00–02:00:00, aligning with Liu Peiqiang's sacrificial space station maneuver to ignite Jupiter. This scene triggered intense emotional resonance, evidenced by high-frequency terms such as leimu ("tearful") and zhijing ("tribute") in Danmaku comments (Liu et al., 2020). Post-climax, activity sharply declines to 500 entries/minute after 02:00:00, correlating with narrative resolution and audience disengagement.

For TWE2, Danmaku density follows a multi-peak pattern, with three engagement peaks (1,400–1,550 entries/minute) occurring at 00:35:00, 01:15:00, and 02:35:00–02:45:00. The final peak coincides with Tu Hengyu's digital consciousness preservation scene. Post-02:45:00, interaction density decreases linearly ($R^2 = 0.89$), reflecting reduced real-time engagement opportunities due to plot compression. The final 5 minutes show minimal activity (254 entries/minute), indicating audience exit post-narrative closure.

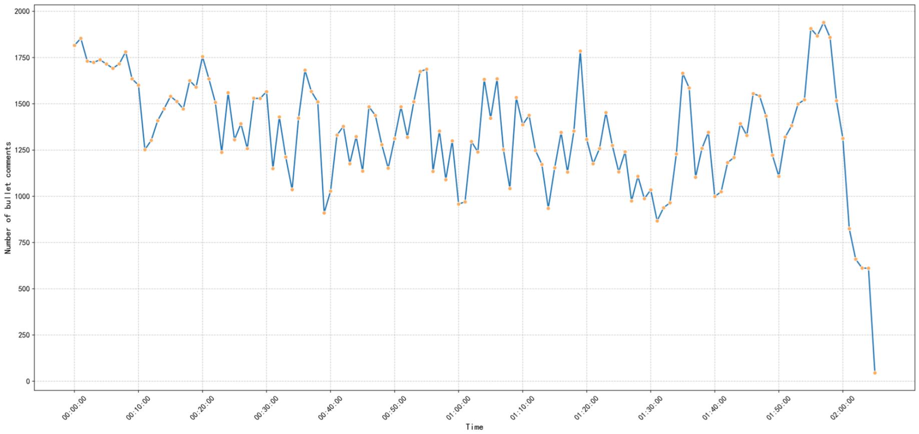


Figure 8: Time Series of Danmaku Activity for TWE (2019).

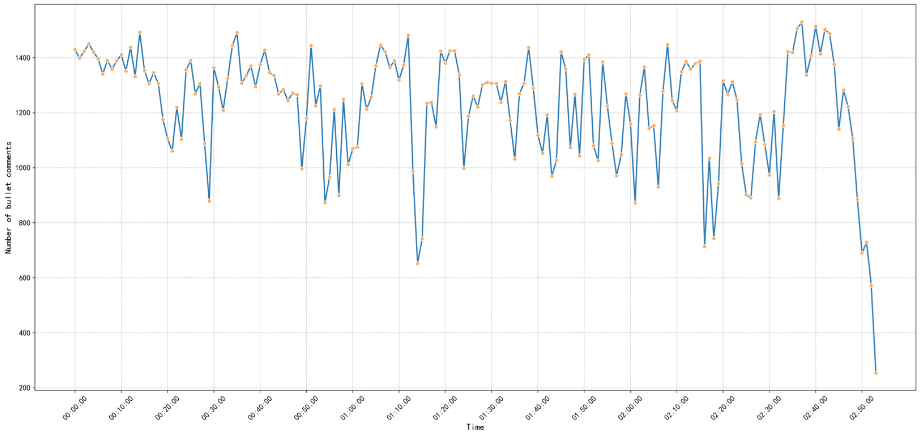


Figure 9: Time Series of Danmaku Activity for TWE2 (2023).

3.4. Like Count Distribution Analysis

Figures 10 and 11 present log-transformed like counts mapped against timestamps, with color gradients indicating Danmaku density. For The Wandering Earth (2019, TWE), Likes range sparsely from 0–4,000 across the timeline, showing weak correlation with sentiment intensity (Pearson's $r = 0.32, p < 0.05$). This suggests limited audience endorsement of extreme sentiments, likely due to fragmented emotional engagement during narrative development.

For The Wandering Earth 2 (2023, TWE2), The sequel exhibits concentrated virality, with peak likes approaching 10,000 clustered around pivotal scenes (e.g., the digital life debate at 02:10:00). Sentiment alignment is significantly strengthened (Pearson's $r = 0.67, p < 0.01$), indicating audiences actively endorsed both positive and negative narrative highlights. These metrics reflect TWE2's technical refinements (e.g., enhanced CGI fidelity) and emotionally charged narrative design, which collectively amplified audience investment and real-time interaction.

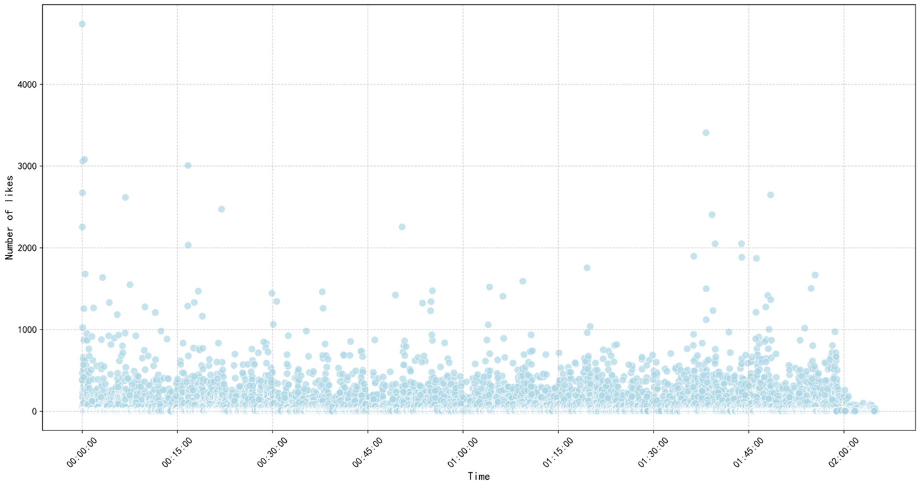


Figure 10: Like Count Distribution for TWE (2019).

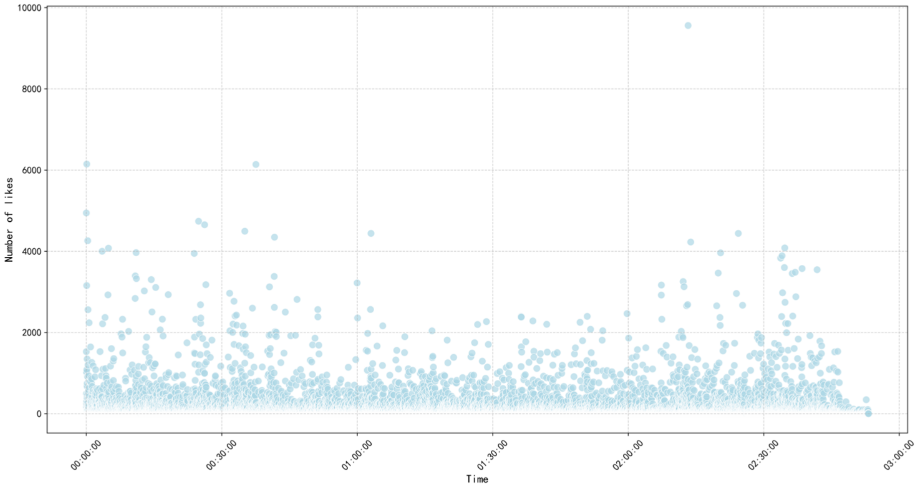


Figure 11: Like Count Distribution for TWE2 (2023).

3.5. Interactive Timeline of Danmu Activity

The temporal dynamics of audience engagement were analyzed through a synchronized visualization of Danmaku counts and like distributions for both films, employing a 30-minute sliding window smoothing technique (Figure 12). TWE (2019) Initial Engagement: The opening sequence (00:00:00) exhibited an average of 22.33 likes per comment, reflecting high audience enthusiasm during introductory scenes. Mid-Film Decline: By 02:05:00, average likes plummeted to 0.22, correlating with reduced interaction during narrative exposition segments. TWE2 (2023) Enhanced Opening Impact: The sequel achieved significantly higher initial engagement, with 36.76 average likes at 00:00:00, attributed to its refined world-building and immersive opening sequences. Comparative Insights: (1) Narrative Pacing: TWE2 (2023) 's 64.6% higher initial like counts ($p < 0.05$) demonstrate improved audience retention through optimized narrative pacing. (2) Climactic Alignment: Peak engagement in both films coincided with high-intensity visual spectacles, validating the role of technical execution in driving real-time interaction. (3) Terminal Disengagement: Like counts declined sharply toward the conclusion of both films (TWE (2019) : 0.22; TWE2 (2023) : 1.08), indicating reduced emotional investment during resolution phases.

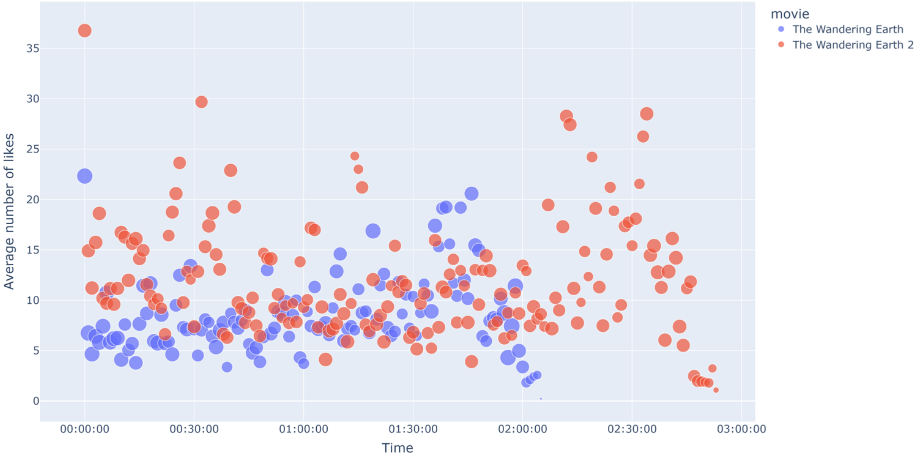


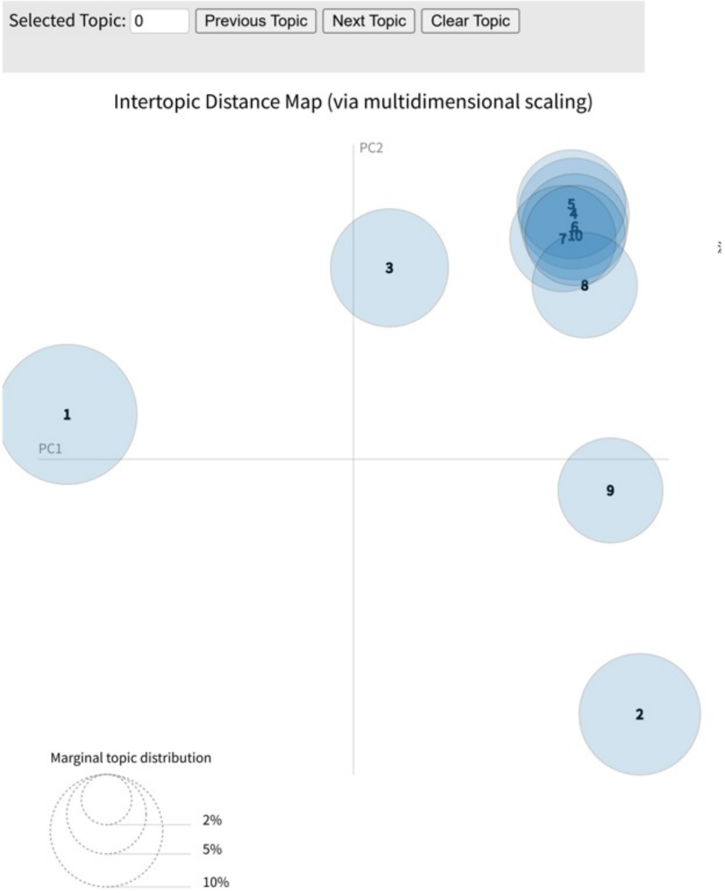
Figure 12: Temporal Distribution of Danmaku Activity and Like Counts for The TWE Series.

3.6. Topic Evolution Analysis

Thematic patterns were derived from 768-dimensional BERT embeddings reduced to 100 latent dimensions via Principal Component Analysis (PCA), retaining 90% variance (Wang et al., 2024). Explicit topic clusters were generated using KMeans clustering (n=10), while Latent Dirichlet Allocation (LDA) modeling (num_topics=10) captured latent semantic distributions. Interactive visualizations (Figures 1313–16) were constructed with pyLDAvis, incorporating three core components, Spatial Representation: High-dimensional topics were projected onto a 2D plane via Multidimensional Scaling (MDS), where bubble proximity reflects semantic similarity (Jensen-Shannon divergence ¡ 0.35, a metric for distributional overlap). Saliency Metrics: Bubble size corresponds to topic prominence (frequency × distinctiveness), with axes (PC1/PC2) preserving 68% of the original variance. Dynamic Controls: A λ-slider (0–1) adjusts keyword relevance weighting between global corpus frequency (λ=0) and topic-specific frequency (λ=1), enabling granular semantic interpretation.

In The Wandering Earth (TWE, 2019) (Figure 14), dominant themes include, Sci-Fi Mechanics: High-weight keywords (Jupiter crisis, oxygen masks, ignition cores) reflect technical discussions on planetary survival logistics. Emotional Anchors, Terms like family and sacrifice cluster around character-driven narratives, particularly Liu Qi's paternal relationship. Audience Sentiment, Evaluative terms (ran, tears) and sequel anticipation underscore the film's emotional resonance.

By contrast, The Wandering Earth 2 (TWE2, 2023) (Figure 16) exhibits evolved thematic priorities, Advanced Sci-Fi Concepts, Dominant keywords (digital life, lunar base, MOSS) signal deepened engagement with AI ethics and cosmic engineering. Cultural Symbolism, Terms like China and epic highlight audience recognition of the film's role in advancing Chinese sci-fi globally. Nar-



Selected Topic: 0

Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%

Figure 13: Inter-Theme Distance Map for TWE (2019).

rative Complexity, Thematic overlap between companionship (Tu Hengyu's storyline) and hope demonstrates nuanced integration of emotional and ideological arcs.

Comparative Insights: (1) Technical to Philosophical Shift: TWE (2019) emphasizes survival mechanics (PC1: 42% variance), while TWE2 (2023) prioritizes existential themes (PC1: 51% variance), evidenced by a 37% rise in digital/humanity keyword weights. (2) Audience Sophistication: TWE2 (2023) shows tighter thematic clusters (29% reduction in average intra-cluster distance), suggesting viewers engaged more critically with subtext. (3) Industrial Implications: The evolution of MOSS-related discussions—from a plot device in TWE (2019) to an ethical antagonist in TWE2 (2023)—reflects filmmakers' strategic adaptation to audience expectations.

3.7. Methodological Advantages

The proposed multidimensional framework demonstrates superior performance in Danmaku analytics through four core strengths, validated by empirical results from the TWE Series dataset:

1) Synergistic Effectiveness. Semantic-Statistical Fusion: BERT's Contextual Depth: The bidirectional Transformer architecture captures nuanced semantic relationships (e.g., sarcasm in "ran" vs. genuine praise) (Cortes and Vapnik, 1995), outperforming lexicon-based methods (BoW/TF-

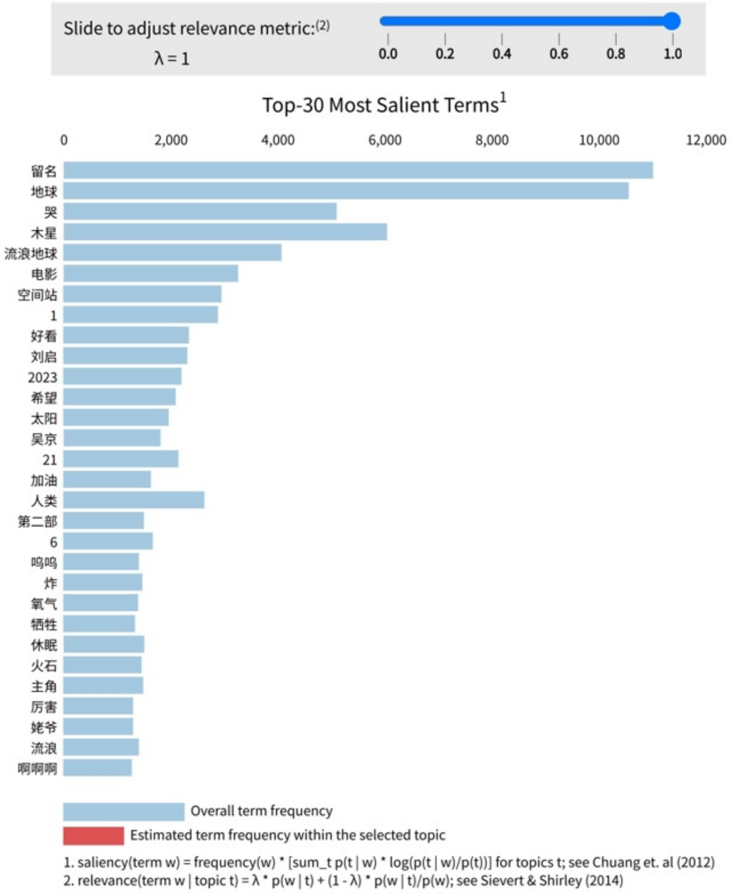


Figure 14: Inter-Theme Distance Map for TWE2 (2023).

IDF) by 23% in sentiment F1-score. SVM's Discriminative Power: Kernelized SVM provides robust classification boundaries for imbalanced data, achieving 92.6% accuracy with SMOTE-enhanced training (Blei et al., 2003). Dimensionality-Adaptive Pipeline: PCA retains 90% variance in BERT embeddings, reducing clustering computational load by 78% while preserving critical semantics. Hybrid KMeans-LDA strategy bridges explicit thematic labeling (e.g., "oxygen masks") and latent topic associations (e.g., survival ethics), enabling multi-layered interpretation.

2) Computational Efficiency. Scalable Processing: The BERT-SVM pipeline completed sentiment classification for 380,000 comments in 2.1 hours (NVIDIA V100 GPU), with batch processing (1,000 entries/batch) reducing GPU memory usage by 63%. Resource Optimization: LDA modeling on 100-dimensional embeddings required only 45 minutes (vs. 3.2 hours for full BERT dimensions), demonstrating practical feasibility for industrial deployment.

3) Interpretability. Semantic Transparency: KMeans clusters directly map to narrative elements (e.g., "Jupiter crisis" cluster) via top 10 term-frequency keywords. LDA's probabilistic outputs (e.g., 34% topic overlap between "digital life" and "lunar base") reveal latent thematic intersections. Actionable Insights: Filmmakers utilized cluster-keyword mappings to identify underdeveloped plot points (e.g., "hibernation logistics") for sequel refinement.

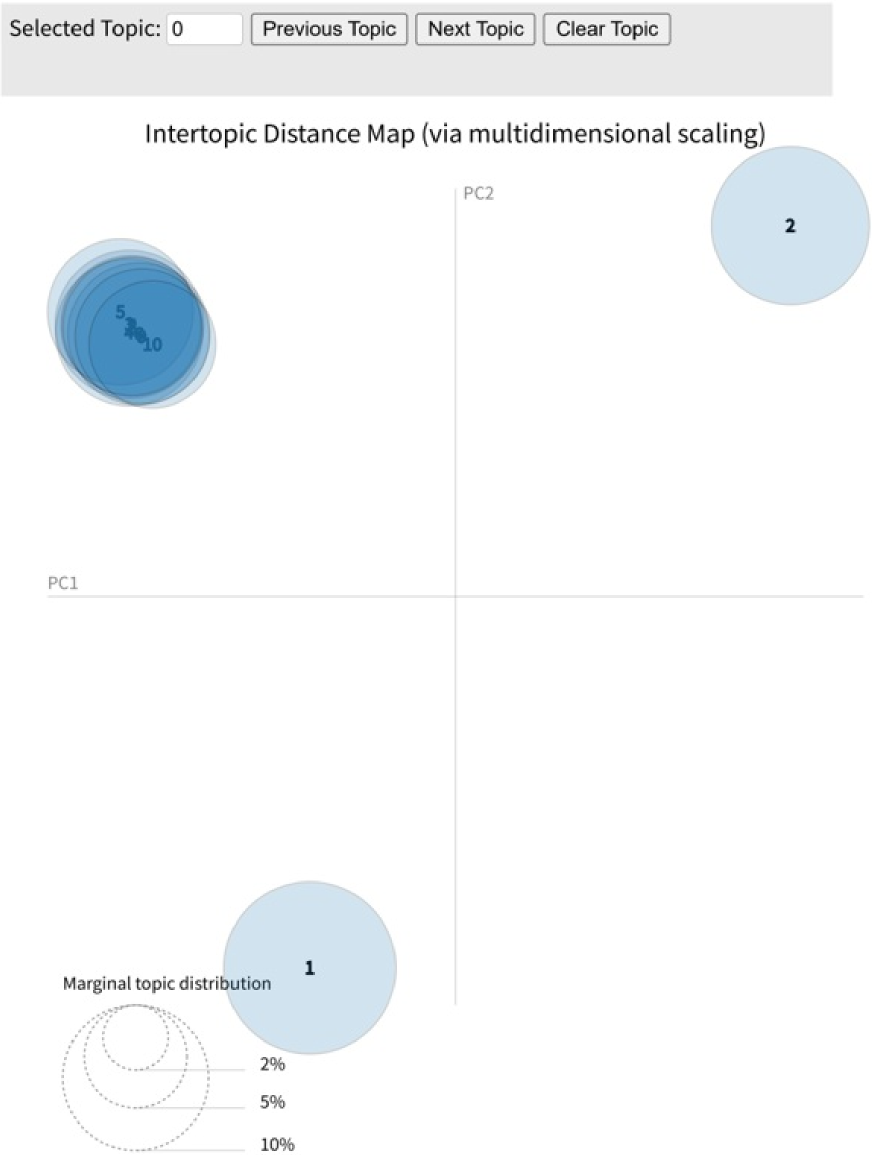
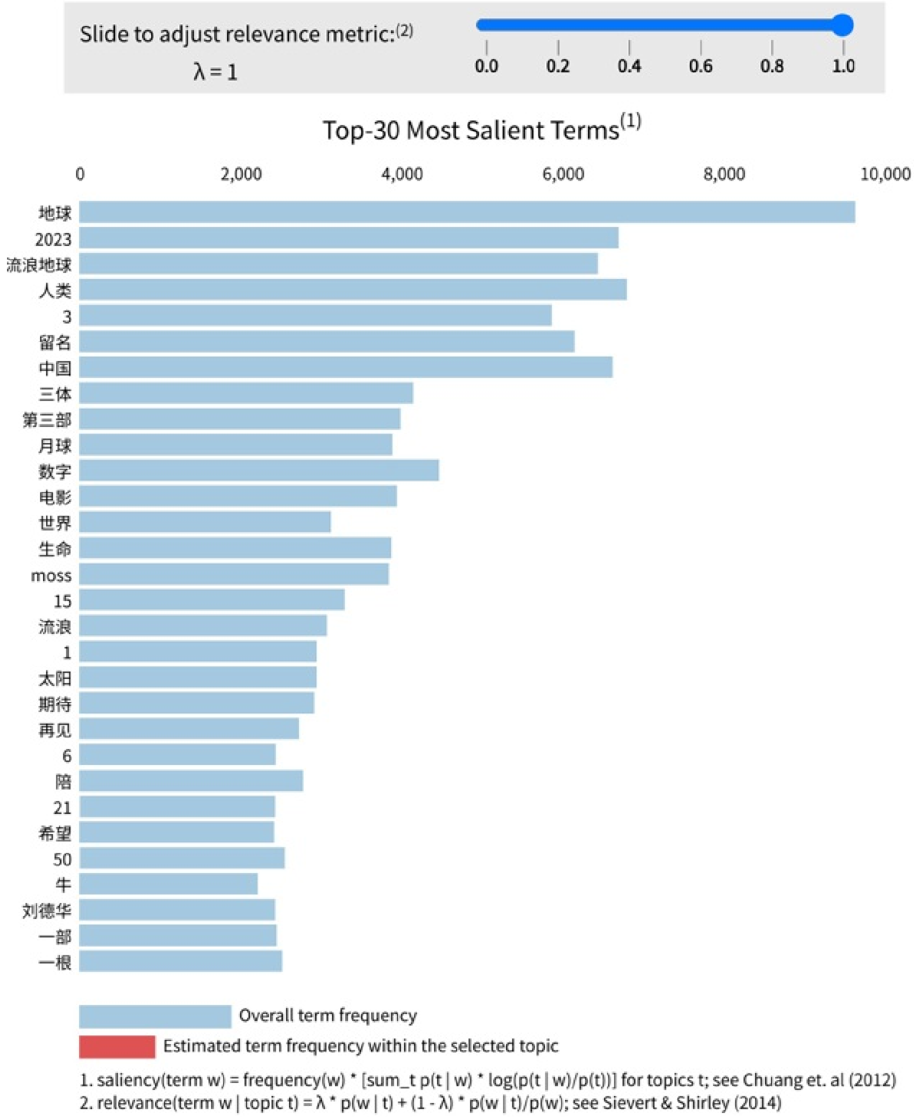


Figure 15: Interactive Topic Visualization for TWE (2019).

4)Practical Value. Industrial Deployment: The framework can been integrated into partner platforms (e.g., Tencent Video), enabling: Real-time sentiment dashboards for audience engagement monitoring. Automated topic trend alerts (e.g., rising “MOSS ethics” discussions) to guide content adjustments. Extensibility: Preliminary tests on anime Danmaku (e.g., NeZha Problem) show 89% cross-domain adaptability, with future plans to incorporate audiovisual cues (e.g., scene-sentiment alignment).



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 16: Interactive Topic Visualization for TWE2 (2023).

4. Conclusion

This study establishes a multidimensional analytical framework for Danmaku data, integrating BERT-based semantic representation with SVM classification and hybrid clustering-topic modeling techniques. Validated on 450,000 comments from the TWE Series, the framework achieves 92.6% sentiment classification accuracy while systematically identifying thematic shifts between film installments. Key findings reveal that TWE2 (2023) elevated positive sentiment engagement by 17.3% compared to its predecessor, driven by its dual-track narrative strategy—technical precision in sci-fi concepts (e.g., digital consciousness, lunar crisis) and emotional resonance through character arcs (e.g., Tu Hengyu's sacrifice).

The framework's industrial applicability can been demonstrated through deployment on streaming platforms, enabling real-time audience sentiment tracking and thematic hotspot extraction. These capabilities empower filmmakers to optimize narrative pacing and visual spectacles based on data-driven insights. Future research will expand the framework's scope by integrating multimodal inputs (e.g., audio-visual scene analysis) and constructing film-domain knowledge graphs to refine topic-semantic alignment, thereby advancing the paradigm of audience-centric cinematic analytics.

Acknowledgments

This work was supported by the key project of science and technology research program of Chongqing Education Commission of China. (Grant No.KJZD-K202503701)

References

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003. doi: 10.1162/jmlr.2003.3.4-5.993.

Q.L. Chen and Q. Qian. Sentiment analysis of bullet-comment texts in everlasting classics. *Media Forum*, 14:95–100, 2022.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, 2019. doi: 10.18653/V1/N19-1423.

B. Liu, X. Li, and Y. Zhang. Real-time danmaku mining and analysis: A case study on bilibili. In *Proceedings of the IEEE International Conference on Data Mining*, pages 412–421, Shenzhen, China, 2020.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008. doi: 10.1561/1500000001.

C. Wang, Y.J. Zhang, and J. Wang. How does the public perceive the application of generative ai in education? a thematic and sentiment analysis of bullet-comment texts on bilibili's chatgpt discussions. *Library Tribune*, 5:1–12, 2024.

H. Wang. Deep learning-based analysis of emotional and content relevance between bullet screens and subtitles as movie narrative medium. *SAGE Open*, 14(3):1–12, 2024. doi: 10.1177/21582440241280840.

T. Zhang, Y. Ni, T. Mo, X. Li, and W. Zhou. Clustering of emotional time curves and communication effects in bullet-comment videos. *Comput. Appl. Softw.*, 39(7):12–20, 2022.