Available Online: 27 April 2026 DOI: 10.54254/2977-3903/2026.32983

Scrolling, clicking, collecting: predicting Bilibili hits with machine learning

Xiang Zuo

Alibaba business college, Hangzhou normal university, Hangzhou, China

2757524120@qq.com

Abstract. As a leading video social platform in China, Bilibili has a huge user base and diverse content ecology, so the correlation between user behavior and content characteristics is crucial for creators to optimize content strategies. This paper takes Bilibili as the research object to explore the relationship between creators' content characteristics and user behavior. Based on the "Must-watch every issue" video dataset of Bilibili, it conducts data preprocessing, feature engineering, and empirical analysis through multiple machine learning models including random forest regression, classification and KMeans clustering, combined with Term Frequency-Inverse Document Frequency (TF-IDF) text analysis and exploratory data analysis. User interactive behaviors are the core factors affecting video playback volume, with collection and likes having the most significant impact. Release time has a stable effect on playback volume, and high IP-focused content is more likely to achieve stable high playback. The constructed random forest model can effectively identify blockbuster videos with a recall rate of over 91%.

Keywords: Bilibili, playback volume, user behavior, content characteristics, machine learning

1. Introduction

Prior studies have explored social media content and user engagement, finding visual quality, emotional appeal and interactivity drive engagement while the effect is platform-specific [1]. Research on Bilibili have divided video popularity into explicit and implicit types, proving bullet screen sentiment, title length and video duration have heterogeneous impacts on them [2]. However, few studies have systematically verified the independent explanatory power of interactive indicators for Bilibili's playback volume, nor have they examined the actual effect of interaction rate given potential target leakage. This paper explores the relationships between user behaviors (like, collect, comment, etc.) and content characteristics (title semantics, release time, content theme, etc.), and verifies the effect of interactive rate on playback volume prediction. Based on Bilibili's "Must-watch every issue" dataset, it adopts data preprocessing, feature engineering and multiple machine learning models for empirical analysis. This study clarifies the core factors affecting Bilibili's video playback volume, providing data support for platform algorithm optimization and practical strategies for creators to improve content communication [3].

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

2. Research methodology

2.1. Data source

This study adopt the Bilibili "Must-watch Every Issue" dataset as the research sample, which covers high-quality videos labeled "Must-watch Every Issue" across all content partitions of Bilibili. The dataset includes video attribute information (title, release time, duration, partition) and user behavior indicators (playback volume, like, coin, collect, comment, share), fully reflecting the basic characteristics of Bilibili's high-quality content and user feedback, and serving as an ideal sample for analyzing playback volume factors and identifying blockbuster videos [2].

2.2. Data preprocessing

Systematic preprocessing was conducted to eliminate noise and lay a foundation for modeling: first, redundant fields were filtered out to retain variables related to the research questions; second, Unix timestamp of release time was converted to datetime type, and derivative features such as release hour and weekday were constructed; third, the top 0.5% of extreme playback volume samples were removed to mitigate the impact of right-skewed distribution; finally, missing value inspection was carried out, and no obvious missing values were found in core fields, so no imputation was required.

2.3. Feature engineering

Targeted feature engineering was implemented for different variable types: interactive features including total interaction (sum of like, collect, coin, share, comment) and interaction rate (total interaction/(playback volume+1)) were constructed to measure user interaction intensity [4]; video titles were preprocessed (de-noising, word segmentation, stop word filtering) and converted into 500-dimensional vector features via TF-IDF to capture semantic differences [5]; log1p transformation was applied to playback volume to weaken extreme value influence, and auxiliary features such as title length were built; video partitions were onehot encoded to be compatible with machine learning algorithms. The processed data was saved as a unified dataset for subsequent analysis.

2.4. Data analysis methods

Considering the research objectives and data characteristics, multiple methods were selected for empirical analysis based on actual research needs:

1. Exploratory Data Analysis (EDA): Descriptive statistics and visualization were used to explore the distribution of core variables, the correlation between interactive indicators and playback volume, and temporal differences in playback volume, which intuitively revealed data laws and provided a basis for subsequent modeling [6].

2. Random Forest Regression: Two contrast models (with/without interaction rate) were built for playback volume prediction. This algorithm was selected for its strong ability to handle non-linear relationships, resistance to overfitting, and feature importance output function, which can verify the target leakage risk of interaction rate and identify core influencing factors [7].

3. Random Forest Classification: The continuous playback volume problem was converted into a binary classification task (top 20% playback volume as blockbuster videos). With the class_weight = balanced parameter set, the algorithm effectively addresses the class imbalance problem and meets the platform's demand for blockbuster content screening.

4. TF-IDF + KMeans Clustering: TF-IDF extracted title semantic features, which were reduced to 50 dimensions via TruncatedSVD to reduce computational complexity. KMeans clustering was then performed, and the optimal cluster number was determined by the silhouette coefficient, realizing the effective division of video content themes [5].

5. Dimensionality Reduction Visualization & Correlation Analysis: Principal Component Analysis (PCA) was used for clustering result visualization, and correlation heat maps analyzed feature correlations, making high-dimensional feature analysis results more intuitive and interpretable.

3. Results and analysis

3.1. Exploratory Data Analysis (EDA) results

The original playback volume exhibited an obvious rightskewed longtailed distribution, and the log1p transformation rendered its distribution approximately normal, verifying the necessity of logarithmic processing. Most video titles were 10–30 words long, with no obvious linear correlation between title length and playback volume, indicating its limited explanatory power for playback volume.

Interactive indicators (like, collect, etc.) were significantly positively correlated with playback volume, and collect behavior had a higher growth slope with playback volume, reflecting users' long-term recognition of content. Strong correlations were also observed among interactive indicators, which warns that subsequent modeling should account for multicollinearity. Temporal analysis showed videos released at 19:00–22:00 and from Friday to Sunday had significantly higher average playback volume; vacation months also saw higher playback volume, consistent with users' leisure time distribution and platform activity characteristics [3].

High-frequency words in titles were mainly content-oriented words such as "game", "animation" and "challenge", as well as IP-related words such as "Genshin Impact" and "Honkai". This revealed that Bilibili's high-quality videos are dominated by practical and entertaining content, and hot IPs and well-known creators play an important role in high-playback videos.

3.2. Playback volume prediction: random forest regression results

Two contrast random forest regression models (Model A with interaction rate, Model B without) were constructed with consistent parameters. The results showed Model A had slightly better test set performance (R^2 = 0.7068, RMSE = 2,677,172.82) than Model B (R^2 = 0.6183, RMSE = 2,849,704.47), indicating that the interaction rate improved predictive performance but only to a limited extent, with no abnormal performance surge induced by target leakage. Both models had no obvious gap between training and test set errors, showing good generalization ability without serious overfitting [7].

Feature importance analysis of Model A showed total interaction, like and collect were the top three core features for playback volume prediction, with significantly higher importance than other variables; interaction rate had a medium-to-upper weight and did not become the dominant variable, further verifying no serious target leakage. Model residuals were concentrated near 0 with a relatively symmetrical distribution, with a slight systematic deviation: overestimation in the low playback volume range and underestimation in the high range, which is a characteristic of random forest models that smooths extreme values and helps reduce the overall prediction error, and title semantic features only had auxiliary explanatory effects.

3.3. Blockbuster video identification: random forest classification results

Blockbuster videos were defined as the top 20% of playback volume, and a random forest classification model was built with the same feature set as Model A. The model performed stably on training and test sets: the test set achieved an accuracy of 0.8728, precision of 0.6250, recall of 0.9102, F1 value of 0.7411 and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) of 0.9617. The training and test set indicators were highly consistent, with an AUC difference of only about 0.01, indicating no obvious overfitting and strong generalization ability.

ROC curves of both sets were close to the upper left corner, with AUC values far higher than the random classification level (0.5), showing strong discrimination ability for blockbuster and non-blockbuster videos. Confusion matrix analysis found the model achieved a recall rate of 91.02% for blockbuster videos, and few blockbusters were misclassified as nonblockbusters; there were a certain number of false positives, which is reasonable in business scenarios because the cost of missing real blockbusters is higher than screening a small number of ordinary videos as candidates [6, 7].

Feature importance results were highly consistent with the regression model: total interaction, like and collect were still the core variables for distinguishing blockbuster videos, verifying the rationality of feature engineering and model construction. With the characteristics of high recall and relatively conservative precision, the model is suitable as an initial screening tool for blockbuster videos, efficiently locating potential high-performance content for subsequent secondary screening.

3.4. Content theme clustering results (TF-IDF + KMeans)

500-dimensional TF-IDF features were reduced to 50 dimensions via TruncatedSVD, and KMeans clustering was performed with k values tested from 2 to 6. The silhouette coefficient showed $k = 4$ (0.476) was the optimal cluster number. PCA visualization found an abnormal "double clustering" phenomenon in Cluster 0 (obvious internal separation in 2D space), but the $k = 5$ scheme had a lower silhouette coefficient (0.411), indicating that the visual separation was a distortion of the highdimensional semantic structure during lowdimensional projection due to information loss in PCA, rather than real independent themes.

Word cloud analysis of high-frequency words showed four clusters with clear semantic distinctions: Cluster 0 (pan-entertainment & game experience, broad themes) with core words like "China", "game" and "Luo Xiang"; Cluster 1 (high-cost original content) dominated by "homemade", "animation" and "hardcore"; Cluster 2 (single IP-driven vertical content) with core words such as "Genshin Impact" and "character"; Cluster 3 (high IP focus, fan attributes) focused on "Honkai: Star Rail" and "PV" [2, 5].

Figure 1 presents the radar chart of the four clusters across mean interaction, mean view, viral rate, mean duration, and mean title length. Cluster 3 shows outstanding performance in all indicators, especially in viral rate and mean view, confirming the strong communication advantage of IP-focused content. Cluster 0 has moderate overall values but a relatively high viral rate, reflecting the polarized performance of broad-themed content. Cluster 1 and Cluster 2 have stable but lower performance, consistent with their positioning as niche or vertically themed content. These results visually verify that content theme significantly affects communication effects, with IP-concentrated vertical content being more likely to achieve stable high interaction and playback.

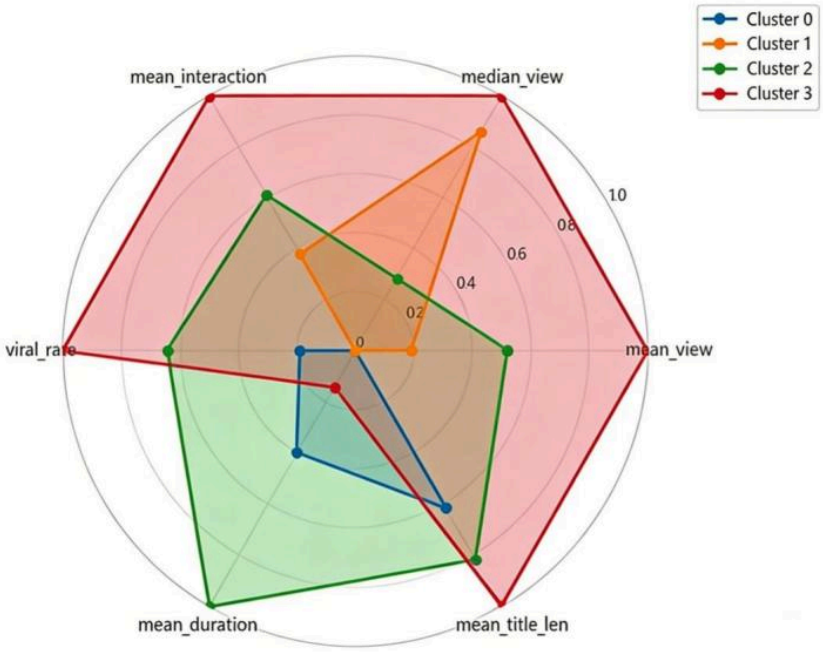

Figure 1. Radar chart of key indicators across four video content clusters (mean interaction, average playback volume, viral rate, average duration, and average title length)

4. Conclusion

This study explores the relationship between content characteristics and user behavior on Bilibili, and identifies the core factors influencing video playback volume and blockbuster attributes through empirical analysis of high-quality video samples. The results confirm that user interactive behaviors (total interaction, likes, and collections) are the decisive factors for playback volume and blockbuster identification, with interaction rate exerting a mild positive effect without significant target leakage due to the logarithmic processing of playback volume. Temporal features also show stable impacts: videos released in evening peak hours and weekends achieve higher playback volume, matching users' leisure time rules. Text clustering further reveals that content themes shape playback performance—IP-focused vertical content with clear audiences has stable high playback and interaction, while broad pan-entertainment content presents polarized performance. Additionally, the constructed random forest models perform well in playback prediction and blockbuster screening, with the classification model reaching a 91% recall rate for blockbusters, demonstrating practical application value for platform operation and creator strategy optimization .This study has limitations: the sample is restricted to Bilibili's "Mustwatch Every Issue" videos, and thus lacks representativeness for ordinary videos. Moreover, it only analyzes explicit user behaviors and ignores implicit factors like watch duration and barrage sentiment. Future research can expand the sample to include ordinary videos for comparative analysis, integrate multi-dimensional user behavior data, and combine qualitative research such as creator interviews to more deeply explore the causal mechanisms underlying content creation and user feedback, and further optimize the prediction model's accuracy and generalization ability.

References

[1] Chen, C., Zhang, C., & Zhou, Y. (2014). A study of user engagement on social media: The role of content characteristics and interactivity in online communities. *Journal of Information Science, 40*(6), 820–833.

[2] Li, H., & Fu, J. (2021). User interaction patterns and viewing behavior on Bilibili: An empirical analysis. *Journal of Information Technology & Politics, 18*(3), 250–268.

[3] Xu, Y., & Zhang, T. (2019). Exploring the impact of release timing and content features on video consumption behavior. *Journal of Media Economics, 32*(2-3), 142–159.

[4] Kaplan, A. M., & Haenlein, M. (2014). Social media: Back to the roots and back to the future. *Journal of Systems and Information Technology, 16*(3), 104–116.

[5] Zhang, X., & Zhou, Q. (2022). Text mining and clustering of video titles to identify thematic influence on user engagement. *Information Processing & Management, 59*(5), 102865.

[6] Wu, B., & Wang, Z. (2018). Understanding the effect of emotional cues and interactivity on engagement in short-form online videos. *Computers in Human Behavior, 86*, 263–271.

[7] Cheng, X., Shen, H., & Fan, J. (2020). Social media video popularity prediction using machine learning techniques. *Journal of Big Data, 7*(1), 78.