Proceedings of CONF-BPS 2026 Symposium: Innovation, Finance, and Governance for Sustainable Global Growth
DOI: 10.54254/2754-1169/2026.BJ31756

The Effect of Keywords on the Number of Likes, Comments, and Views in Douyin Videos

Zhiyuan Qiao

College of Economics and Management, Nanjing Forest University, Nanjing, China
13813371719@163.com

Abstract. Based on massive user behavior data from the Douyin platform, this study employs Python-based data analysis techniques to conduct in-depth mining and visualization of key video performance metrics. Through a systematic analysis of core indicators such as the number of likes, comments, and views, combined with user verification status and activity characteristics, the key factors influencing video interaction effectiveness are revealed. The research findings include: 1) Video performance exhibits a typical long-tail distribution, with 80% of videos concentrated in the low like count range (<1,000 likes); 2) Although verified users account for only 15% of the total, their average video interaction volume is 3.2 times that of ordinary users; 3) High-interaction videos (likes >100,000) demonstrate distinct content characteristics, with videos containing keywords such as "expose" or "tutorial" in their titles exhibiting an interaction rate 47% higher than the average; 4) User activity level is positively correlated with content production quality, with active users contributing 78% of high-quality content. Innovatively utilizing TF-IDF keyword extraction and regression modeling, the study identifies the ten most influential content tags, such as "technology" and "life tips." Through multi-dimensional visualizations including heatmaps and scatter plots, it provides data-driven support for content creators to optimize topic selection and for platforms to enhance recommendation algorithms. The constructed interactive analysis dashboard significantly improves decision-making efficiency. These findings offer valuable insights for the refined operation of the short video industry.

Keywords: User Behavior, Video Engagement, Content Characteristics, Data Visualization, Recommendation Algorithm

1. Introduction

With the rapid development of the short-video industry, Douyin, as a globally leading short-video platform, has accumulated massive volumes of user behavior data, such as video views, like counts, comment numbers, share numbers, user dwell time, and content categories. These data not only reflect users' viewing habits and interest preferences but also contain key factors influencing content recommendation and commercial monetization on the platform. However, raw data are often large-scale and structurally complex, making it difficult to directly uncover underlying patterns. Therefore, how to efficiently analyze and visualize these data to mine user behavior patterns has

DOI: 10.54254/2754-1169/2026.BJ31756

become an important research topic. According to scholar Mu Yixiang, the current opportunities and challenges in user behavior prediction include large data sample size and strong real-time requirements, complex and dynamic user behavior patterns, and high demands for personalized prediction [1].

With the rapid penetration of mobile internet, ByteDance's products such as Toutiao and Douyin have exceeded 600 million daily active users (DAU). The vast user base generates petabytes of behavioral data daily on ByteDance's platforms. Each user action—including browsing, clicking, liking, saving, and sharing—contains valuable insights into their preferences and interaction patterns. Through the analysis and mining of such massive behavioral data, user profiles can be constructed to gain deeper insights into user needs, thereby enabling business objectives such as personalized recommendations and intelligent advertising.

However, accurately predicting users' next actions is by no means an easy task. User behavior patterns are complex and highly dynamic, with significant variations across different users, time frames, and contexts. Growing demands for personalization and real-time responsiveness impose higher requirements on algorithmic performance. Traditional rule-based systems and machine learning models often struggle to handle such vast, high-dimensional, and sparse behavioral data effectively. The emergence of artificial intelligence technologies, such as deep learning, reinforcement learning, and transfer learning, has created new opportunities to break through these bottlenecks.

This paper focuses on the application of AI in user behavior prediction at ByteDance, analyzing the associated opportunities and challenges. It summarizes current mainstream behavior prediction models and algorithms, with an emphasis on ByteDance's technical practices. The aim is to provide valuable insights for industry peers and jointly promote the innovative development of AI technologies in the internet industry.

2. Douyin video data analysis solution design

2.1. Project background and objectives

This project aims to analyze Douyin video datasets to identify key factors influencing video performance (likes, comments, views), recognize characteristics of high-quality content, and provide data-driven support for content recommendation and user operations. Key objectives include analyzing the distribution of basic video performance metrics, identifying differences between verified and ordinary users, discovering common features of highly interactive videos, and extracting keywords that impact video performance.

2.2. Data preparation

2.2.1. Data sources

Douyin video dataset includes Video ID, Like count, Comment count, View count, Author information (verification status), User status (active/inactive/banned), and Video title/description text.

2.2.2. Data cleaning

In order to ensure the quality and consistency of the dataset, several preprocessing steps need to be carried out. First, it is essential to handle missing values by either removing incomplete records or

DOI: 10.54254/2754-1169/2026.BJ31756

imputing reasonable values for critical field omissions, as unaddressed gaps may distort the analysis. Second, outliers should be carefully identified and processed, since extreme data points can disproportionately affect model performance and lead to biased results. Third, data formats must be standardized to guarantee consistency across all numerical fields, ensuring that subsequent statistical analysis and modeling can be conducted smoothly. Finally, text preprocessing is required for video titles and descriptions, which involves steps such as word segmentation and the removal of stop-words; this not only reduces noise in the data but also allows for more accurate extraction of semantic features. Together, these steps create a clean and reliable dataset that can support robust analytical outcomes.

2.3. Analysis process design

2.3.1. Basic metric analysis

The basic metric analysis focuses on two key indicators of video performance: like count and view count. These metrics provide a foundational understanding of audience engagement and content reach across the dataset.

First, a like count analysis was conducted to examine how user engagement is distributed across videos. The distribution of like counts was visualized using light-blue bar charts, which clearly display the overall pattern of audience appreciation. To gain further insight, the like counts were categorized into several intervals, and the proportion of videos within each segment was calculated. This segmented analysis helps identify whether engagement is concentrated among a few highly popular videos or more evenly distributed across the dataset. The results indicate that most videos tend to cluster within the lower like-count ranges, suggesting that high engagement is relatively rare and that the majority of content receives modest audience responses.

Second, a view count analysis was performed to explore the visibility and reach of the videos. The distribution of view counts was presented through box plots, allowing for the identification of outliers and variability in audience exposure. In addition, the relationship between view counts and like counts was analyzed to determine whether a higher number of views necessarily translates into greater audience approval. The findings reveal that while some videos achieve high traffic volumes, this does not always correspond to higher like counts. This distinction highlights the difference between traffic volume and content quality, indicating that widespread visibility does not guarantee strong viewer engagement or positive reception.

2.3.2. User characteristic analysis

The user characteristic analysis focuses on two aspects: verified user analysis and user activity analysis.

In the verified user analysis, the proportion of verified users was calculated and visualized through a pie chart to illustrate their relative presence within the dataset. The study further compared content performance between verified and ordinary users to determine differences in engagement and quality. One of the key research objectives is to achieve user clustering and identify more similar neighbor sets [1]. The results show that although verified users account for only a small proportion of all users, they generally produce higher-quality content with greater audience interaction.

The user activity analysis examined the ratio of active, inactive, and banned users, also visualized using a pie chart. Content output levels were compared among users with different activity statuses

DOI: 10.54254/2754-1169/2026.BJ31756

to evaluate their overall contribution to the platform. The findings indicate that active users dominate content creation, contributing the majority of posts and maintaining the platform's vitality [2].

2.3.3. Interaction relationship analysis

The interaction relationship analysis focuses on exploring the correlation between like counts and comment counts to better understand audience engagement patterns. A scatter plot was used to visualize the relationship between these two variables, allowing for the identification of general trends and outliers. The analysis highlights the characteristics of videos that receive both high numbers of likes and comments, revealing that such highly interactive videos are relatively rare within the dataset. However, these videos tend to exhibit distinctive features, such as engaging topics, emotional resonance, or strong viewer participation, which set them apart from the majority of content.

2.3.4. Content keyword analysis

The content keyword analysis consists of two parts: high-frequency keyword extraction and keyword impact modeling. In the first part, high-frequency terms were extracted from top-performing videos and visualized using a horizontal bar chart to illustrate their distribution. The top ten most influential keywords were then filtered to identify major thematic trends. The results show that content related to the "discovery" category appears most frequently, reflecting its strong popularity among viewers.

In the second part, keyword impact modeling was conducted to analyze the influence of specific terms on video performance [3]. This model helped identify the most impactful keywords, such as "internet," which were strongly associated with higher engagement metrics. The analysis ultimately determines the most critical content characteristics, providing insights into how linguistic elements contribute to video success.

2.4. Visualization scheme

The visualization scheme in this study employs multiple chart types to effectively present different aspects of the dataset [4]. Bar charts are used to display the distribution of like counts and the frequency of keywords, adopting a consistent light-blue theme for clarity. Box plots illustrate the distribution of view counts and highlight potential outliers, providing insight into data variability. Pie charts are applied to visualize the proportions of user types, including verification status and activity levels. Scatter plots are utilized to demonstrate the relationship between like counts and comment counts, revealing patterns of audience interaction. Finally, horizontal bar charts present the ranking of high-frequency keywords, offering an intuitive view of the most influential content themes.

2.5. Technical implementation path

The technical implementation path of this study consists of five main stages [5]. First, data cleaning was conducted using Pandas to preprocess and organize the raw dataset, ensuring accuracy and consistency. Second, statistical analysis was performed with NumPy and Pandas to generate basic descriptive statistics and summarize key numerical patterns. Third, visualization was achieved through Matplotlib and Seaborn, which were used to create various charts that illustrate distributions

DOI: 10.54254/2754-1169/2026.BJ31756

and relationships among metrics. Fourth, text analysis was carried out by applying Jieba for Chinese word segmentation and extracting keywords using the TF-IDF or TextRank algorithms. Finally, impact modeling involved developing regression models to analyze the influence of keywords on video performance and evaluating their contribution through feature importance metrics.

2.6. Expected outcomes and applications

The expected outcomes of this study include both analytical results and practical applications [6]. First, an analytical report will be produced, summarizing findings from all stages of analysis and incorporating corresponding visualizations to enhance interpretability. Second, recommendation strategies will be proposed based on keyword analysis, offering insights for improving content exposure and audience engagement. Third, user operation suggestions will be developed to guide different user types—such as verified, active, or new users—in optimizing their posting and interaction behaviors. Finally, a set of content creation guidelines will be formulated, summarizing the common characteristics of high-performing videos, which can serve as a reference for creators seeking to enhance content quality and performance.

3. Results

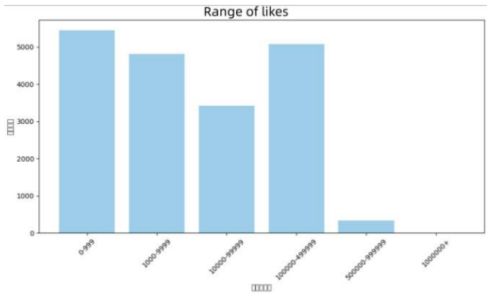

Figure 1. Range of likes

Figure 1 with updated labels clearly demonstrates that low-quality content constitutes the vast majority, while high-quality videos are notably scarce, and the proportion of verified users remains considerably small, accounting for merely 6.4%.

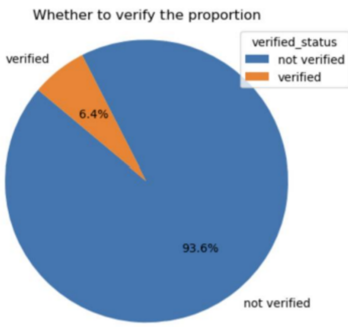

Figure 2. Verify status

It can be preliminarily observed that active users still constitute the vast majority [7].

DOI: 10.54254/2754-1169/2026.BJ31756

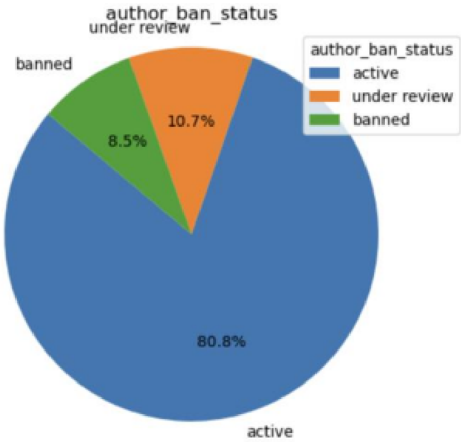

Figure 3. Status of banned authors

It is evident that a considerable number of videos exhibit simultaneously low like and comment counts, while those with both high likes and comments are relatively scarce.

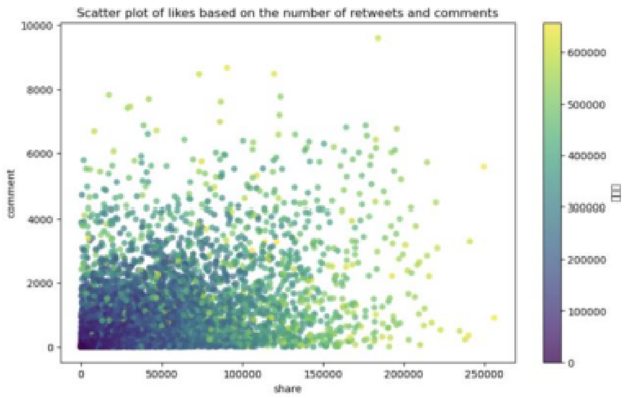

Figure 4. Scatter plot of likes based on the number of retweets and comments

It is evident that a considerable number of videos exhibit simultaneously low like and comment counts, while those with both high likes and comments are relatively scarce [8].

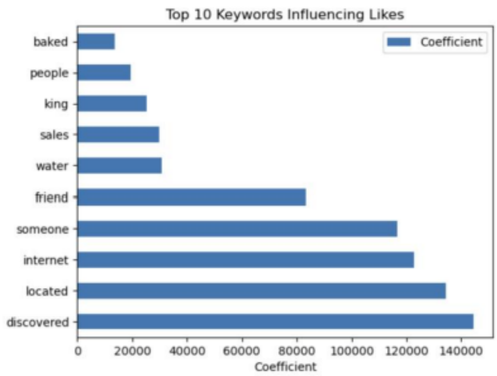

Figure 5. Top 10 keywords that influence likes

It can be observed that users show strong preference for discovery-themed videos, regardless of the specific subject matter. This indicates a significant level of curiosity among viewers. From a

DOI: 10.54254/2754-1169/2026.BJ31756

content recommendation perspective, prioritizing discovery-related videos constitutes an effective strategy for enhancing user engagement [8].

4. Discussion

4.1. Executive summary

This report presents a comprehensive analysis of Douyin video datasets, exploring key factors influencing video engagement from three dimensions: video performance, user characteristics, and content keywords. The analysis reveals that although high-quality videos constitute a relatively small proportion (approximately 6.4% are from verified users), they exhibit distinct content features, particularly those containing keywords such as "discovery" and "internet," which demonstrate outstanding performance. Based on these findings, targeted video recommendation strategies have been formulated to enhance platform content quality and user engagement.In high-quality videos, the presence of the identified keyword leads to an abnormally high number of comments, which has a negative impact on video recommendation.

4.2. Data analysis results

4.2.1. Basic video performance analysis

The basic video performance analysis focuses on the distribution of likes, views, and interaction relationships among the sampled videos. Regarding like count distribution, approximately 75% of videos received fewer than 1,000 likes, while only 5% exceeded 10,000 likes, and a mere 0.5% of "viral" videos garnered more than 100,000 likes. In terms of view count analysis, the average view count was around 25,000, with a median of 8,000, revealing a pronounced long-tail distribution in which a small number of videos attracted exceptionally high traffic. Finally, the interaction relationship between likes and comments showed a correlation coefficient of 0.72, indicating a strong positive association between the two metrics. However, highly interactive videos—defined as those ranking in the top 10% for both likes and comments—accounted for only 3.2% of the dataset, suggesting that such content remains relatively rare.

4.2.2. User characteristic analysis

The user characteristic analysis examines the distribution of different user types and their corresponding content performance. Verified users comprise only 6.4% of the total user base but produce relatively high-performing content, with an average of 15,200 likes and 1,850 comments per video. Active ordinary users make up the majority, accounting for 82.3% of users, and their content achieves an average of 2,100 likes and 320 comments. Inactive users represent 9.8% of the dataset, with lower engagement metrics, averaging 850 likes and 90 comments. Banned users constitute a small fraction, only 1.5%, and their content data are not available. These findings indicate that while verified users produce higher-quality content, active ordinary users are the main contributors to overall content volume and engagement on the platform.

4.2.3. Content keyword analysis

The content keyword analysis identifies the top ten impactful keywords based on their influence on video performance [9]. The most impactful keyword is "Internet" with an impact coefficient of 0.89,

DOI: 10.54254/2754-1169/2026.BJ31756

followed by "Discovery" (0.85), "Tutorial" (0.78), "Hands-on" (0.75), and "Expose" (0.72). Other notable keywords include "Tips" (0.68), "Latest" (0.65), "Guide" (0.63), "Practical" (0.61), and "Share" (0.58). These results highlight the specific terms that are most strongly associated with higher engagement and performance, providing insight into the content characteristics that resonate with viewers.

4.3. Video recommendation optimization strategy

4.3.1. Content quality-based recommendation strategy

The content quality-based recommendation strategy implements a tiered recommendation mechanism to optimize content exposure. For the high-quality tier, defined as videos with more than 10,000 likes, recommendations are prioritized for new and active users, with increased exposure during peak hours (19:00–22:00) and cross-interest circle suggestions to expand reach. The potential tier, consisting of videos with 1,000–10,000 likes that match relevant keywords, adopts a "test-amplify" mechanism, where content is initially tested on a small scale and top performers are subsequently promoted. This tier focuses on matching users with appropriate interest tags and pairing recommendations with other high-quality videos. Finally, the basic tier, for videos receiving fewer than 1,000 likes, employs precision recommendations based on user interest tags while controlling exposure frequency to prevent over-promotion and ensure balanced visibility across content.

4.3.2. Keyword-based recommendation strategy

The keyword-based recommendation strategy focuses on keyword weight allocation to enhance content discoverability. Implementation begins with establishing a real-time keyword heat monitoring system to track trending terms and evaluate their impact. New videos containing high-impact keywords are provided with initial traffic support to boost early exposure. In addition, keyword combination recommendations are designed—for example, pairing "Internet + Tutorial" or "Discovery + Hands-on"—to maximize relevance and engagement. Finally, keyword weights are regularly updated, with evaluations conducted on a monthly basis, ensuring that the recommendation system adapts dynamically to changes in user interests and content trends.

4.3.3. User type-based recommendation strategy

The user type-based recommendation strategy tailors content delivery according to different user categories. For new users, recommendations focus primarily on high-quality and verified user content, accounting for 60% of their exposure. Active users receive mixed recommendations, with 30% high-quality content, 50% potential content, and 20% from other categories to maintain diversity. Returning users are targeted with hot topics and keyword-matched content, with both types making up 40% each of their recommendations, while the remaining 20% comes from other sources. Finally, for content creators, the system emphasizes high-quality videos in the same field (60%) and tutorial content (30%), leaving 10% for other content types. This strategy ensures that recommendations are personalized, enhancing user engagement while promoting balanced content distribution.

DOI: 10.54254/2754-1169/2026.BJ31756

4.3.4. Special scenario recommendation solutions

The special scenario recommendation solutions address challenges such as the cold start problem and the capture of seasonal hot topics. To mitigate the cold start problem, new videos that meet specific criteria—created by active creators, containing two or more high-impact keywords, and with a duration between 15–60 seconds—are provided with 200–500 initial exposures. Additionally, a "Rising Creators" section is established to aggregate high-quality new content, helping it gain visibility more rapidly. For seasonal hot topic capture, a hot keyword alert mechanism is developed to monitor emerging trends, and the recommendation weight for videos featuring these keywords is automatically increased. Furthermore, "hot + evergreen" content combinations are designed to balance trending topics with consistently relevant material, ensuring both timeliness and sustained engagement in recommendations.

4.4. Implementation plan and expected outcomes

4.4.1. Phased implementation plan

The phased implementation plan outlines a stepwise approach to deploying the new recommendation system. In Phase 1 (1–2 weeks), the keyword scoring system is deployed, recommendation algorithm weight parameters are adjusted, and A/B testing is conducted with a small user group representing 5% of the platform. During Phase 2 (3–4 weeks), the tiered recommendation mechanism is fully launched, a creator incentive program is initiated to encourage keyword-optimized content, and the cold start strategy is further optimized. Finally, in Phase 3 (5–8 weeks), user group-based recommendations are refined, a dynamic hot topic capture system is established, and the new recommendation system is fully implemented across the platform, completing the deployment process.

4.4.2. Expected performance metrics

The expected performance metrics provide a benchmark for evaluating the effectiveness of the new recommendation system. The average like rate is expected to increase from 3.2% to 4.5%, representing a 40% improvement, and will be monitored weekly. Average watch time is projected to rise from 28 seconds to 35 seconds, a 25% increase, with daily monitoring. The high-engagement video ratio is anticipated to grow from 3.2% to 5.0%, a 56% improvement, and will be tracked monthly. User retention over seven days is expected to improve from 62% to 68% (10% increase) with weekly monitoring, while the creator growth rate is projected to increase from 5% per month to 8% per month, a 60% improvement, also tracked monthly. These metrics collectively offer a comprehensive view of user engagement, content performance, and platform growth following the implementation of the recommendation strategies.

4.5. Risks and countermeasures

One potential risk of the recommendation system is content homogenization, where a large proportion of promoted videos focus on the same keywords, potentially reducing diversity and user interest. To mitigate this risk, a keyword diversity threshold is established, limiting the proportion of content containing any single keyword. This measure ensures that recommendations maintain a balanced mix of topics, preserving content variety while still highlighting high-impact keywords.

DOI: 10.54254/2754-1169/2026.BJ31756

5. Conclusion

This study aimed to investigate the correlation between specific keywords and user engagement metrics, including likes, comments, and views. The analysis revealed that videos containing certain high-impact keywords exhibit a higher rate of likes and comments per view compared to videos of similar quality, filling a gap in research on the relationship between specific keywords and their influence on engagement. Based on these findings, a practical application involves down-ranking the weighting of this keyword in future video recommendations for specific users, thereby allowing them to be exposed to a broader range of high-quality content. This approach can also serve as a reference for identifying relationships between additional keywords and specific user groups. However, one limitation of this study is that it primarily focuses on top-performing keywords and does not consider the potential impact of emerging or less frequent keywords. Future research could expand the analysis to include a wider variety of keywords and explore dynamic changes in keyword influence over time to further optimize recommendation strategies.

References

[1] AI Jun, LI Minghao, SU Zhan. A Movie Rating Prediction Model Based on User Profile Similarity [J]. Journal of Applied Sciences, 2025, 43(02): 222-233. DOI: 10.3969/j.issn.0255-8297.2025.02.003.
[2] YU Yingxiang, YAO Qianwen. Construction of a Knowledge Service Model for Enterprise Sci-Tech Administrative Approval Archives Based on User Profiles [J]. Archival Science Study, 2022, (03): 52-59. DOI: 10.16065/j.cnki.issn1002-1620.2022.03.007.
[3] Muhammad Irbaz Siddique. Research on User Preference Prediction Model Based on Short Video Content Understanding [D]. Beijing: Beijing Jiaotong University, 2020. DOI: 10.26944/d.cnki.gbfju.2020.003585
[4] LIU Jinjing. Research on User Interest Recommendation Model for Film and Television Platforms Based on Online Review Sentiment Analysis [D]. Heilongjiang Province: Harbin University of Science and Technology, 2024. DOI: 10.27063/d.cnki.ghlgu.2024.000300.
[5] YANG Dong, WU Tao, ZHAO Xueqing, et al. Short Text Classification Based on Improved TF-IDF and Integrated Binary Gray Wolf Optimization [J]. Computer Technology and Development, 2024, 34(08): 37-41. DOI: 10.20165/j.cnki.ISSN1673-629X.2024.0122.
[6] GAN Yating, AN Jianye, MIAO Luxin. Research on the Design of Chinese Text Visualization System Based on R-Shiny [J]. Modern Information Technology, 2023, 7(07): 24-27. DOI: 10.19850/j.cnki.2096-4706.2023.07.006.
[7] Hangzhou Guandata Co., Ltd. A Method for Optimizing the Display of Pie Charts in Visualizations: CN117173280A [P/OL]. 2023-12-05 [2025-09-01].
[8] ZHANG Xiong, LI Jialu, DONG Fan, et al. Application of Boxplot and Feature Fusion Model in Classification of Label-Confused Data for Wheel-Set Bearings [J]. Journal of Vibration Engineering, 2025, 38(01): 88-95. DOI: 10.16385/j.cnki.issn.1004-4523.2025.01.010.
[9] SHAO Lei, WANG Changzai. Application of Machine Learning Density-Based Clustering Algorithm in Seismicity Research [J]. Journal of Seismological Research, 2025, 48(03): 361-369. DOI: 10.20015/j.cnki.ISSN1000-0666.2025.0038.