Research on Douyin Data Analysis Based on Deep Learning

Yuxia Du[a,*] and Joey S. Aviles[b]

Graduate School, Angeles University Foundation, Angeles City 2009, Philippines

[a]du.yuxia@auf.edu.ph, [b]aviles.joey@auf.edu.ph

Abstract. There are many types of data on the Douyin platform, including user behavior data, video content data, interaction data, etc., and the amount of data is huge. Deep learning technology, with its powerful data processing capabilities, is able to efficiently process these complex data and extract valuable information and features. By analyzing Douyin data, you can gain insight into users' interests, behaviors, preferences, and other characteristics, so as to help brands or creators target audiences more accurately. This precise targeting helps to improve the pertinence and attractiveness of the content, increasing user stickiness and conversion rates. And analytics can reveal what types of content are more popular, when content is more effective, how users interact with it, and more. This information is essential for optimizing content strategies, which can help creators adjust the direction, format, and release time of content to better meet user needs and improve the dissemination and influence of content. So as to improve the marketing effect and enhance competitiveness.

Keywords: Deep learning algorithms, Douyin, data mining, User profile

1 Background

As of December 2023, the number of short video users in China reached 1.053 billion, an increase of 41.45 million from December 2022, accounting for 96.4% of the total number of netizens. As a leader in the field of short videos, the number of users of Douyin has exceeded the 1 billion mark, of which the number of users in the Chinese market accounts for a large proportion. This shows that Douyin has a large user base and provides a rich data source for data analysis. Douyin has strong user stickiness, with an average daily usage time of more than 7 hours per person, and a high user retention rate. This reflects the high dependence and frequent use of Douyin by users, further increasing the complexity and challenge of data analysis. As a platform with hundreds of millions of users, Douyin's user behavior, preferences, trends and other data are an important basis for business decisions. Through the analysis of Douyin data, enterprises can understand user needs, market hotspots, competitive situations, etc., so as to formulate more targeted marketing strategies, product planning and operation strategies to enhance business competitiveness. Douyin data analysis helps the platform better understand user needs and feedback, find users' pain points and dissatisfaction in

A. K. Draman Mud et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Social Sciences (ICBDSS 2024)*, Advances in Computer Science Research 116,
https://doi.org/10.2991/978-94-6463-562-1_22

the process of use, and carry out product optimization and upgrades. For example, by analyzing the user's viewing history, interactive behavior and other data, the content recommendation algorithm can be optimized to provide users with more personalized and accurate recommended content, and improve user experience and satisfaction. As a leader in the field of social media and short videos, Douyin's data reflects the consumption concepts, cultural trends and social hotspots of young people. Through the analysis of Douyin data, we can gain insight into market dynamics and trends, and provide reference for enterprises to formulate long-term development strategies. For example, analyzing data such as hot topics and popular elements can predict cultural phenomena and consumer trends that may be popular in the future for a period of time.

Douyin data analysis involves cutting-edge technologies such as big data processing, machine learning, and deep learning, and the application and research and development of these technologies require continuous data support and verification. Through the analysis of Douyin data, the shortcomings and potentials of technology application can be discovered, and technological innovation and progress can be promoted. For example, optimizing recommendation algorithms and improving data processing capabilities require analysis and experiments based on large amounts of data. In a rapidly changing market environment, Douyin faces many risks and challenges, such as user churn, intensified competition, and policy adjustments. Through the analysis of Douyin data, potential risks and problems can be discovered in a timely manner, and corresponding response strategies and measures can be formulated to reduce the impact of risks on the platform. For example, by analyzing data such as user retention and activity, you can predict the risk of user churn and take steps to intervene.

2 Review of Relevant Literature

There are many studies on Douyin data analysis, and in terms of user analysis, it is mainly through the study of user gender, region, age characteristics and other information to portray user portraits, so as to determine users' preferences for purchasing product categories and purchase time preferences [1]. In terms of commodity analysis, current research mainly focuses on the classification and clustering of commodities, and determines the popularity of commodities by studying the behavioral information of the commodities [2]. E-commerce big data mainly solves supply chain problems, analyzes the products that are currently sensational on the market, and makes predictions about the future demand for goods, and merchants can reasonably allocate inventory according to the forecast [3]. In terms of online payment, e-commerce big data analysis can detect abnormal behaviors to ensure the security of user payments[4].

For example, Adomavicius and others summarized the characteristics and consumption rules of users by analyzing their reading time and click-through rate and other behaviors [5], so as to construct a user portrait model. Bhtacharyya et al. analyzed the similarity between users by analyzing the keyword text used by users in FaceBook[6], so as to discover the interaction between friends in social relationships. Pazani et al. classified and analyzed the user-generated interest tags and summarized the user interest profiling method[7]. Li et al. used the analysis of user social tags to discover users' interests and

hobbies, and used the clustering method to divide users' interest characteristics [8]. Hawalah et al. mapped the model to a reference ontology based on user preferences and interests[9], and then used it as an analysis and mining task for short-term and long-term interests, so as to construct a widely accepted data ontology system that is easy to understand users. Cui et al. used a combination of SOM and K-means algorithms to cluster and analyze all user documents and each user document respectively to construct user portraits of users' interests [10].

3 Research Content

User Persona is a virtual user model created based on user data and behavioral characteristics to better understand and describe the target user group. The concept was first coined by Alan Cooper, the father of interaction design, who described it as "a virtual representation of a real user, a target user model built on a set of attribute data". With the development of the Internet, the connotation of user portraits has been gradually enriched, and now it usually includes information such as the demographic characteristics of users, online browsing content, online social activities and consumption behaviors, and abstracts a labeled user model.

The main content of user portraits includes but is not limited to the following aspects:

(1) Basic information: such as age, gender, occupation, income level, etc., which are the basic data for building user portraits.

Behavior patterns: Behavior data such as browsing, clicking, purchasing, and sharing on the Internet can reflect users' interests, preferences, and consumption habits.

(2) Needs and motivations: By analyzing user behaviors and feedback, user needs and motivations can be inferred, and then users can be provided with more personalized services or products.

(3) Pain points and problems: Identifying and solving the pain points and problems encountered by users in the process of using products or services is the key to improving user experience.

The core work of building user portraits is to use the massive logs stored on the server and a large amount of data in the database to analyze and mine, and label users. These "tags" are identifiers that can represent a certain dimension of the user's characteristics, such as "like sports", "shopaholic", etc. By tagging users, users can be divided into different groups and different marketing strategies or product plans can be developed for different groups.

Personas play an important role in business operations and product development. It can help enterprises more accurately understand user needs and market trends, optimize product design and service processes, and improve user experience and satisfaction. At the same time, user portraits are also the basis of precision marketing and personalized recommendations, through the analysis and application of user portraits, enterprises can achieve more accurate marketing and personalized recommendation services, and improve marketing effectiveness and conversion rate. In general, user portraits are an important tool for enterprises to understand and serve users in the digital era. By building

user portraits, enterprises can better grasp market dynamics and changes in user needs, and provide strong support for the long-term development of enterprises.

The user portrait of the Douyin platform is a young, active, strong spending power and highly social attributes. They pursue fashion, trend, and personalized experiences, and have a high demand for high-quality goods and services. At the same time, they are also an important force in the creation, dissemination and consumption of content on the Douyin platform. It is a complex and multi-dimensional concept, which is constructed based on the user's social attributes, living habits, consumption behavior, and interactive behavior on the Douyin platform. We collect data by analyzing Douyin user portraits, and analyze and predict them

Douyin data analysis can predict trends and results in multiple aspects, including the following:

(1) User behavior prediction

By analyzing users' viewing history, likes, comments, and sharing behaviors, it predicts users' interests and preferences for specific types of content, so as to provide users with more personalized content recommendations. It can also analyze the level of activity of users in different time periods, predict the active time of users, and help content creators and advertisers publish content or run ads at the best time.

(2) Content trend forecasting

By monitoring video views, likes, comments, and retweets, you can predict which topics or content are likely to go viral, so you can guide content creators to follow trends and create more engaging content. It can also use algorithms such as machine learning to predict the type, style, or element of content that may be popular in the future based on historical data and current trends, providing a reference for content creation and marketing.

(3) Business effect prediction

According to user portraits and interest preferences, predict the effect of advertising in different user groups, help advertisers accurately place ads, and improve advertising conversion rates. For e-commerce content, by analyzing users' purchase behavior and product browsing data, it predicts the sales trend and potential demand of products, and provides a basis for merchants to adjust inventory and marketing strategies.

(4) Platform development forecast

By analyzing data such as user registration and active users, the user growth trend of the platform is predicted, and the long-term development plan of the platform is provided. Combined with the prediction results of user behavior and content trends, the health and diversity of the platform's content ecology are evaluated, and guidance is provided for optimizing the content ecology and improving user experience.

(5) Risk assessment and response

By analyzing data such as user retention rate and activity, we can predict the risk of user churn and formulate corresponding countermeasures to reduce user churn. Use natural language processing and other technologies to monitor the quality of content, predict possible violating content, take measures to intervene and deal with it in advance, and maintain the healthy ecology of the platform.

4 Research Process

The Douyin data analysis process is a systematic and complex effort that aims to provide strong support for the operation, content creation, and marketing strategies of the Douyin platform by digging deep into and interpreting the data. Here are the main steps of Douyin data analysis:

(1) Data collection

Data collection is the first and most basic step in Douyin data analysis. There are various types of data collected, including basic user information, such as age, gender, region, etc., which can help build user portraits. There is user behavior data, such as watch time, likes, comments, shares, retweets, etc., which reflect the user's activity and interaction. There is video content data: such as video type, topic, tags, etc., which helps to understand the popularity and distribution of the content. There is market data: such as competitors, industry trends, etc., which is essential for developing a marketing strategy.

(2) Data cleaning and sorting

Data cleansing is required because the raw data collected may contain errors, duplicates, incompleteness, or inconsistent formatting. The purpose of data cleaning is to eliminate noise and outliers in the data and ensure the quality and accuracy of the data. Data wrangling is the classification, grouping, and sorting of cleaned data for subsequent analysis.

(3) Data analysis

Data analysis is the core step of Douyin data analysis. At this stage, it is necessary to use various analysis methods, such as comparative analysis, trend analysis, user portrait analysis, etc., to dig deep into the information and patterns behind the data. Specific analysis methods include:

The results of playback analysis, likes, comments, shares and retweets, user portrait analysis, and data analysis are usually presented in a visual way such as reports and charts for easy understanding and interpretation.

(4) Data interpretation and application

In the process of interpreting data, it is necessary to focus on the outliers, trend changes and reasons behind the data, so as to provide strong support for decision-making. Based on the results of data analysis, you can adjust your content creation strategy, optimize your promotion plan, and improve your user profiles. For example, based on the results of user portrait analysis, formulate a more accurate targeted promotion strategy. Optimize the quality and type of video content based on the results of content analysis. Adjust product development and marketing strategies based on market trend forecasts.

5 Research Results

The data for this experiment is mainly from the Kaggle dataset, which refers to a series of datasets available on the Kaggle platform for data science, machine learning, and data analytics competitions. Kaggle is an online platform that hosts various data science

competitions and offers the ability to host and share databases. These datasets cover a wide range of topics from everyday life to professional fields, such as video game sales, air quality, biomedical data, image recognition, and more. The Kaggle dataset is an important resource in the field of data science, which provides users with rich data and hands-on opportunities to improve their data science skills and ability to solve real-world problems (please refer to Figure 1).

(1) Through data analysis, we can see that the top four types of videos that Douyin users are more interested in are: life, food, health, sports, etc. The daily life category includes food making, home decoration, travel sharing, parenting experience, etc., showing users' life clips and practical skills. This kind of video is close to the user's daily life and is easy to resonate and pay attention to. Entertainment and funny videos attract viewers with their light-hearted and humorous content, which is easy to elicit resonance and laughter. The content includes jokes, funny performances, parody shows, etc., bringing joy to the audience through exaggerated expressions and humorous lines.

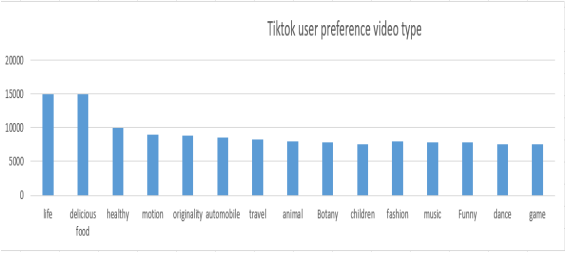

Fig. 1. Distribution of Douyin video types

(2) The use of Douyin by all ages (please refer to Figure 2):

The age group of users who use Douyin is relatively broad, but it is mainly concentrated in the younger group. Here's a specific analysis of the age distribution of Douyin users:

Young people: The main user group of Douyin is young people, especially users under the age of 30. Specifically, users aged 18-24 have the highest proportion of users, accounting for about 35%, followed by users aged 25-30, accounting for about 27%. In addition, users aged 31-35 also account for a certain percentage, about 16%.

Other age groups: While Douyin's main users are younger people, it is also expanding its reach to other age groups. For example, the post-85 generation has a relatively high proportion of female users, and the proportion of users in lower-tier cities exceeds 6 percent. Among the post-80s, male users account for a relatively high proportion and have a high preference for automobiles, mother and baby, and food videos.

(3) According to data analysis, the time periods with high user activity on Douyin are mainly concentrated in the following periods (please refer to Figure 3):

Before work in the morning (about 6:00-9:00): This is the time when many people wake up and get ready for work or school, and users may use this time to browse their phones, including TikTok for entertainment or information. In addition, the morning is also the prime time for some inspirational, workplace content to be released.

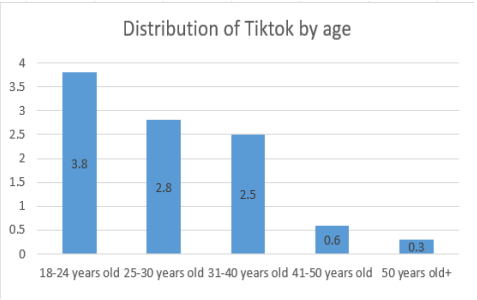

Fig. 2. Distribution of Douyin users' age groups

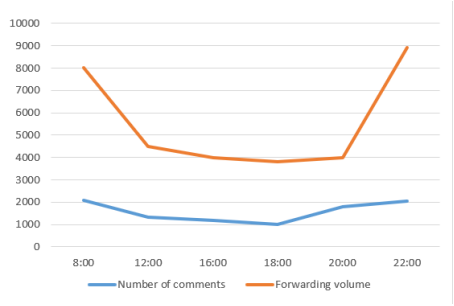

Fig. 3. Douyin user activity

Lunch break (about 12:00-14:00): This time period is the time when most people have lunch and take a lunch break, and many people will choose to brush Douyin during this time period to relax themselves. At the same time, this is also an intensive time for dinner and leisure communication between colleagues, and the probability of talking and sharing recommended Douyin videos is also higher.

From work to bedtime (about 18:00-24:00): This time period is one of the most active times for Douyin users. Many people use this time to browse their phones after work, including scrolling through TikTok to kill time or get entertainment. Especially in the evening from 21 o'clock to 23 o'clock, many people have already finished dinner and have sufficient leisure time, and many people have the habit of browsing their mobile phones and swiping Tik Tok to help them sleep before going to bed. This period is also the prime time for the release of chicken soup, emotional and other content.

It is important to note that these time periods are only generalized based on general experience and user habits, and do not fully represent the active time of all Douyin users. In addition, as user habits and the content ecology of the Douyin platform continue to change, these time periods may also be adjusted.

(4) From the perspective of the time of e-commerce live broadcast, the traffic at different times is different (please refer to Figure 4).

The time-sharing traffic of the Douyin e-commerce live broadcast room refers to the audience traffic obtained in the live broadcast room in different time periods. This traffic situation is affected by a variety of factors, including the weight of the live broadcast

room, real-time data performance, user behavior, etc. Most of the first wave of traffic in the live broadcast room depends on the weight of the account itself. The platform will give the live broadcast room an initial weight positioning based on the account's previous live broadcast data, so as to determine the initial traffic at the start of the broadcast. The activity of the account is also an important factor that affects the weight. Active accounts are more likely to get recommendations and traffic tilt from the platform. Evenings and weekends are usually the peak hours for Douyin users, so the live stream traffic during these hours will also be relatively large. However, specific peak hours may vary depending on region, user habits, and other factors. At the time nodes such as 7 o'clock and 8 o'clock on the hour, the traffic on the hour will be large because the recommendation of the traffic pool is also calculated according to the time. In addition, an even multiple of 5 (e.g., 7:10, 7:20) may get more referral traffic. The traffic in the live broadcast room fluctuates in different time periods. Streamers need to adjust their live streaming strategies based on real-time data to cope with changes in traffic.

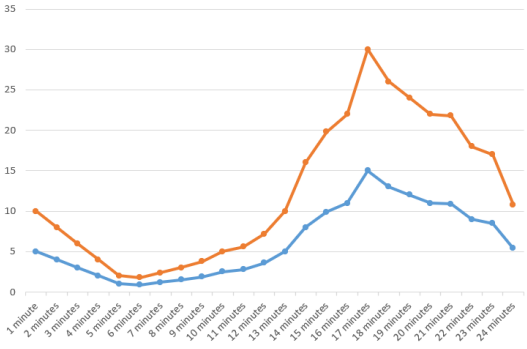

Fig. 4. The activity of the live broadcast room

6 Summary

While Douyin data analysis provides many conveniences for users and creators, it also has some shortcomings:

(1) Data privacy and security issues

To enable the Douyin data analysis function, users need to provide personal information, including viewing history, likes, comments and other data. This data, if mismanaged or misused, can lead to the disclosure of personal privacy that could be used for commercial espionage, harassment, or other illegal purposes. Although Douyin has strict rules on the use of data, users still need to be vigilant and cautious about their data privacy. As the volume of data increases, so do the security risks during data storage and transmission. Hacker attacks, data leaks, and other incidents may pose a threat to the security of users' personal information.

(2) Information overload and interference

Data analytics provides a wealth of data about user behavior and preferences, but too much information can lead to information overload for creators, making it difficult to

sift through valuable content. Too much data analysis can interfere with the creative process of creators, making them rely too much on data and ignore their own creative inspiration and intuition, which can affect the quality and uniqueness of the work.

References

1. Tian M, Sun C, Wu S. An EMD and ARMA-based network traffic prediction approach in SDN-based internet of vehicles. Wireless Networks. 2021: 1-13.
2. Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networksfJ. IEEE transactions on neural networks and learning systems. 2020，32(1):4-24.
3. Liang Rongling, Nian Qifeng. Visualization Analysis of Hurun Rich List Data Based on Python Crawler [J]. Computer and Information Technology, 2022, 30 (6): 46-50
4. Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications[J]. Al open, 2020,1: 57-81.
5. Jiang W, Luo I. Graph neural network for traffic forecasting: A survey [J]. Expert Systems with Applications, 2022: 117921.
6. Zhou Hong. Data Visualization Exploration of Correlation Factors Analysis of Digital Technology Application in Higher Vocational Education Based on SPSS [J]. Digital Technology and Application, 2023, 41 (3): 114-116
7. Fan Luqiao, Gao Jie, Duan Banxiang A domestic popular tourist attraction data visualization system based on Python+Flask+ECharts [J1. Modern Electronic Technology, 2023.46 (9): 126-130
8. Li Yuanxue, Wang Guangming. Research on the Application of Data Visualization in Disaster News Reporting [J]. News World, 2023, (3): 19-23
9. Li Tai, Cheng Xinyue, Li Aidi, et al. Design and Implementation of a Multi Source Transportation Big Data Visualization System Based on Distributed Architecture [J1. Urban Construction Theory Research (Electronic Version), 2023, (10): 119-121
10. Cui B, Zhu H. Abnormal Detection of Electric Management System based on Spatial-temp oral User Profile[Jl. Procedia Computer Science, 2018, 139:269-274