Spatio-Temporal Patterns of Fitness Behavior In Beijing Based on Social Media Data

Tian Bin
College of Applied Arts and Science of Beijing Union University

Meng Bin (✉ mengbin@buu.edu.cn)
College of Applied Arts and Science of Beijing Union University

Zhi Guoqing
College of Applied Arts and Science of Beijing Union University

Qi Zhenyu
College of Applied Arts and Science of Beijing Union University

Chen Siyu
College of Applied Arts and Science of Beijing Union University

Liu Jian
College of Resource Environment and Tourism of Capital Normal University

Research Article

Keywords: Social media, fitness behavior, LDA model, geodetector, Beijing Spatio-temporal patterns, Weibo

Posted Date: November 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-967228/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Using social media data, this paper employs FastAI, Latent Dirichlet Allocation (LDA) and other text mining techniques coupled with GIS spatial analysis methods to study temporal and spatial patterns of fitness behavior of residents in Beijing, China, from the perspective of residents' daily behavior. Using LDA theme model technology, it is found that fitness activities can be divided into four types: running-based fitness; riding-based fitness; fitness in sports venue; and fitness under professional guidance. Emotional analysis revealed that, residents can get a better fitness experience in sports venues. There are also obvious differences in the spatio-temporal distribution of the different fitness behaviors. Fitness behavior of Beijing residents has a multi-center spatial distribution pattern, with a wide coverage in northern city areas but obvious aggregation areas in southern city areas. In terms of temporal patterns, the residents' fitness frequency shows an obvious periodic distribution (weekly and 24 hours). And there are obvious differences in the time distribution of fitness behaviors for each theme. Additionally, based on the attribution analysis of a geodetector, it is found that the spatial distribution of fitness behavior of residents is mainly affected by factors such as catering services, education and culture, companies and public facilities.

1. Introduction

Health is a theme that is an important part of many people's lives. It is largely influenced by individual biological factors (e.g., disease, heredity, gender, age) \(^1\,^2\), but fitness behavior is considered to be one of the most changeable influencing factors \(^3\). Fitness behavior is susceptible to the subjective feelings of residents, the availability of activity venues and activity time and other factors \(^4\). The Chinese government has always attached importance to the implementation of a national fitness plan. The National Fitness Plan (2016-2020) proposed that national fitness is an important way and means to reflect the comprehensive strength of the country. However, the current per capita fitness activity participation in China is not high. For example, only 33.9% of people regularly participate in physical fitness activities, and the fitness rate of residents aged 20-69 is only 14.7\(^5\). From the perspective of the availability of fitness venues, China's per capita area of sports facilities in 2017 \(^6\) was 1.66 m\(^2\), compared to 16 m\(^2\) in the United States and 19 m\(^2\) in Japan \(^7\). From the perspective of industrial development, the contribution rate of fitness and leisure activities in China's sports industry in 2017 was 3.26%. This value is much lower than the 15% in the United States \(^8\), indicating that the Chinese people's awareness of fitness is relatively weaker. The number and distribution of fitness venues is also insufficient, and the development of related industries is relatively low. Some study found that the geographical environment affects the form, structure and method of sports REF. Conversely, the demand for sports can prompt changes in the existing geographical environment \(^9\,^12\). Therefore, it is particularly important to study the spatio-temporal distribution of fitness behaviors in urban areas.

Many studies within China and internationally have considered patterns of urban fitness behavior. From the perspective of residents, many authors have examined the characteristics of fitness motivation, form
and frequency in order to optimize the layout of urban fitness venues\textsuperscript{13, 14}. From the perspective of fitness venues, relevant research pays more attention to the transport range, accessibility, distribution characteristics, utilization rate, carrying capacity and other influencing factors of these locations\textsuperscript{15-20}. However, most of these studies are based on questionnaire or statistical data and are restricted by factors such as questionnaire design and statistical limitations, and affected by subjective factors such as the recall and skills of the surveyors involved. As the temporal and spatial scale of the sample coverage is often small, there is a certain risk to the reliability of the resulting data and conclusions\textsuperscript{21-24}.

With increasing public participation, social media has become an important means for residents to share their daily lives with one another. Weibo is one of the most widely used social media platforms by Chinese residents. By the end of 2016, Weibo had more than 300 million monthly active users, of which mobile users accounted for more than 90\%. Application of big data such as social media data in the study of fitness behavior of residents can effectively make up for the number of samples and objectiveness that could not be achieved by previous data collection and survey methods. For instance, using social media data, some studies have analyzed the spread of social media-based fitness guidance in virtual space\textsuperscript{25}. They combined Weibo check-in data with urban point of interest (POI) data to study the spatial distribution and influencing factors of urban fitness space\textsuperscript{26}.

Currently, there are few studies of the combined use of text, spatial information and urban POI social media data. This study uses Weibo text data with geographic location information. Such integrated data can not only identify the user's spatial location, but also the user's feelings through text analysis technology. Specifically, this study uses Weibo data to analyze fitness behavior of urban residents from the perspective of big data research. The study examines the temporal distribution of fitness behavior, the spatial distribution of fitness behavior and the main factors affecting fitness behavior.

\section*{2. Results}

\subsection*{2.1 Overall distribution}

In order to better reveal the spatial distribution of fitness behaviors of residents in Beijing(Figure 1), this study analyzes the overall fitness Weibo data in terms of space and time. From the perspective of spatial distribution, we constructed a grid of 1km*1km within the 5th Ring Road in Beijing, and counted the number of fitness behaviors in each grid in order to more intuitively discover the overall distribution characteristics of residents' fitness behaviors. The fitness behaviors of residents in Beijing are generally distributed in multiple centers, with a wide coverage in the northern city and obvious clustering areas in the southern city. Specifically, areas with dense fitness activities are mainly concentrated in densely populated residential areas, such as Fengtai Town, Majiapu, Xiluoyuan, Songjiazhuang, Chaoyang District, Beitaipingzhuang, and Yuanda Road in Chaoyang District. Regions and so on. Secondly, there are more distributions in large parks, such as the Olympic Forest Park. On the contrary, in some working areas where Beijing urban residents are relatively concentrated, such as CBD and Zhongguancun, the concentration is not high.
From the perspective of time distribution (Figure 2), spring and summer are the most frequent seasons for residents to exercise. Among them, March is the month where residents have the largest number of exercises throughout the year. From the perspective of working days, Wednesday is the most frequent day for residents to exercise. The number of fitness behaviors peaks twice in a day. The specific performance is: Beijing residents generally start fitness activities at 5 am, and reach a small climax at 9 am. After that, the fitness behavior continued to increase, and reached the peak of the day around nine o’clock in the evening, and then dropped rapidly.

2.2 Thematic analysis

This study uses an empirical setting method to determine the number of fitness behavior topics, including viewing the subject terms of the classification results, comparing whether the differences between different results are obvious, etc., after multiple tests and calculating the optimal classification results. The microblog content related to the fitness behavior of Beijing residents is processed using LDA theme model technology. After repeated tests, it is finally determined that the best results will be obtained by dividing it into four theme categories. Figure 3 shows the effect of topic classification results on a two-dimensional plane after multiple dimensionality reductions. Among them, the size of each circle represents the number of samples contained in different topics, and the differentiation between different topics is represented by the distance between different circles. The classification results show that when the number of topics is four, fitness behaviors can be well divided into four categories, and there are obvious differences between the categories.

Table 1. Keywords and types of LDA themes.

<table>
<thead>
<tr>
<th>Subject Category</th>
<th>Percentage of Total Sample Size</th>
<th>Keywords</th>
<th>Fitness Activity Type Characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme One</td>
<td>32.5%</td>
<td>Exercise, clock in, run, start, hold on</td>
<td>Running based fitness behavior</td>
</tr>
<tr>
<td>Theme Two</td>
<td>26.8%</td>
<td>Fitness, cycling, clock in, kilometer, hold on, start</td>
<td>Cycling based fitness behavior</td>
</tr>
<tr>
<td>Theme Three</td>
<td>21.9%</td>
<td>Today, clock in, now, play ball, minutes</td>
<td>Sports fitness behavior in venues</td>
</tr>
<tr>
<td>Theme Four</td>
<td>18.8%</td>
<td>Cheer, hour, gym, oneself, feeling</td>
<td>Gym and other professional fitness behaviors</td>
</tr>
</tbody>
</table>

In-depth analysis of the semantic characteristics of the high-frequency keywords of the four themes, based on which condenses the behavioral characteristics of each theme. The verb "run" appeared in the high-frequency keywords of theme 1, combined with encouraging words such as "clock in" and "start", the characteristics are more obvious, so the behavioral characteristics are summarized as "Running based fitness behavior". The verb "cycling" appeared in the high-frequency keywords of the second theme, as well as fitness-related words such as "kilometer" and "hold on". It can be clearly seen that this
type is mainly based on long-distance cycling fitness, so the summary of its behavior is characterized as "Cycling based fitness behavior". The verb "play ball" appears in the high-frequency keywords of topic three, plus the qualifying words "today" and "minute", it can be inferred that such behaviors are performed in professional sports venues such as basketball courts, table tennis halls, and badminton halls. Therefore, the behavioral characteristics are summarized as "Sports fitness behavior in venues". The noun "gym" appeared in the high-frequency keywords of theme four, as well as words expressing emotions such as "Cheer", "oneself", and "feeling". It can be inferred that this type of behavior is performed in areas such as the gym using professional fitness equipment. The behavior of fitness, therefore, summarizes its behavioral characteristics as "Gym and other professional fitness behaviors". The number of the four themes in the total sample and the keywords are shown in Table 1.

2.3 Emotional evaluation of fitness behavior

The level of emotional value can directly reflect the feelings of residents when they perform fitness behaviors, and it is also an important index to enrich the research on the characteristics of the spatio-temporal distribution of residents' fitness behaviors. In response to this analysis, this research uses manual screening and keyword extraction to eliminate check-in Weibo automatically generated by APP from all Weibo data. The purpose is to reduce the impact of such Weibo data on the emotional analysis of residents' fitness behaviors.

In this study, sentiment analysis tools were used to calculate the sentiment value of each Weibo, and the sentiment value was used to indicate the mood of residents during the fitness behavior. According to the results of the subject classification, we calculate the average value of the emotional value of each category of fitness behavior (Figure 4). Using this method can intuitively reflect the emotional characteristics of residents in the four fitness behaviors. The results show that residents have better overall fitness experience. Among them, performing physical fitness behaviors in venues generally results in a better fitness experience and a more comfortable mood. In contrast, residents sometimes do not get a good fitness experience using professional fitness equipment in the gym or at home.

2.4 Spatio-temporal patterns of fitness behavior

Based on the above research results, this research will also separately interpret the microblog text information of the four fitness behaviors and produce the corresponding nuclear density distribution map (Figure 5), in order to interpret their intrinsic attributes and spatial distribution characteristics.

2.4.1 Theme 1: Running based fitness behavior

The main representative of theme one is running-oriented fitness behaviors. Most of these behaviors are residents expressing their feelings after the long-distance running, or recording their every exercise by clocking in. In this part of the check-in data, about half of the check-in records are automatically generated using sports apps. It can be seen that sports apps have a greater impact on residents in this
type of fitness behavior. A small number of fitness behaviors involve other sports related to running at the same time, such as skipping rope and hiking.

From the perspective of spatial distribution, the overall fitness behavior of this theme presents a distribution pattern of "more gathering areas, more east and less west". This type of fitness behavior is mainly concentrated near residential areas, such as near Songjiazhuang and Wangjing. Secondly, parks near residential areas are mostly distributed, such as Ritan Park. In addition, it is also distributed near some colleges and famous attractions.

2.4.2 Theme 2: Cycling based fitness behavior

Theme 2 mainly represents fitness behaviors based on cycling. The main types include urban cycling, cycling to scenic spots, night cycling fitness, and the popular shared bicycle cycling check-in. Similarly, there are many check-in records automatically generated by residents using sports APP after exercising. In addition, some fitness-related topics initiated in Weibo are also one of the important factors affecting residents' fitness behaviors.

From the perspective of spatial distribution, the overall situation is "one center, two sub-centers, more in the north and less in the south". This theme fitness behavior has the distribution characteristics of clusters around universities and parks, such as Capital Normal University, China Agricultural University, Beijing Language and Culture University, Beijing University of Aeronautics and Astronautics, Yuandadu Ruins Park, Olympic Forest Park, etc. And mainly distributed between the North Fourth Ring to the North Fifth Ring.

2.4.3 Theme 3: Sports fitness behavior in venues

Theme three mainly includes recreational fitness behaviors such as basketball, swimming, and badminton. This type of fitness behavior has more stringent requirements on the venue than running and cycling. Sports such as table tennis and tennis need to be performed in professional fitness venues. In the relevant texts, many residents mentioned that they had to commute long distances in order to play ball. Therefore, some residents would also perform other sports such as cycling and running while performing fitness behaviors on this subject.

From the perspective of spatial distribution, the overall appearance of the spatial clustering features spreading from the center to the surroundings. The central area is located between Minzu University of China, Peking University, Liudaokou, and Beijing Normal University. The distribution feature of this theme is similar to that of theme 2, but it is more concentrated in the vicinity of universities. At the same time, there are also small gathering areas near the International Trade Center.

2.4.4 Theme 4: Gym and other professional fitness behaviors

Theme 4 mainly includes fitness behaviors that use professional equipment to train in the gym. The fitness content is richer than the first three themes. In addition to the emotional expression of residents
after fitness, it also includes numerical records and fitness opinions using ellipsometers, dumbbells and other equipment. Residents who do this kind of exercise tend to have more stringent requirements for their own health and body shape. Most of them have clear fitness programs, fitness intensity, and summarize the details of changes in their health and weight.

From the perspective of spatial distribution, this type of fitness behavior mainly has three concentrated areas. This type of agglomeration area has a more significant feature, that is, it is mainly located near large residential areas and commercial centers, such as Liudaokou in the north, Guomao in the east, and Jiaomen West in the south.

2.4.5 Comprehensive analysis

By observing the “hot-spot” map of fitness behaviors of residents in the research area (Figure 6) we can clearly see that there were significant hot-spots in fitness behaviors of the different themes and the distributions of hot-spots were quite different. (1) Overall there was more activity in the north of the research area than in the south, with some significant hot-spots in the area of Xueyuan Road-Lincui Road-Osen Park, near the International Trade Center and the Beijing Central Business District, as well as near large residential areas in the south and east. (2) Different fitness behaviors were unevenly distributed in space. Theme 1 had discrete hot spots in the east and south of the area, and fitness behaviors of Theme 4 had discrete hot-spots in the centre.

In summary, the aggregation of centers of fitness behaviors in the research area was relatively high, and there was an obvious trend of gradually decreasing activities from the center of the aggregation to the surroundings. Additionally, a small number of high-density areas were also formed in some peripheral areas, and the overall distribution pattern was of large aggregation areas as the main body with scattered small aggregation areas.

We observed temporal distribution differences between the four fitness behavior themes using two time frames: 24 h and one week (see Figure 7).

In relation to the 24-hour daily cycle, residents’ exercise time was mainly concentrated in the evening, reflected by the gradual increase in the number of exercisers from 17:00 in the afternoon, and reaching a peak around 20:00 in the evening. Additionally, some residents also performed fitness behavior from 7 am to 10 am. The types of fitness activities at this time were mainly Theme 1 (running) and Theme 2 (cycling).

Taking a week as the cycle, residents’ fitness activity was mainly concentrated on Sunday and Monday. For the fitness behaviors of Theme 1 and Theme 2, which are less affected by the need for specific venues, their time distribution characteristics were similar, and were focused on Monday and Wednesday although they had a high degree of participation almost every day. For fitness activities that have specific needs for venues or facilities, there was a large difference in daily participation. The fitness behavior of
Theme 3 was mainly took place on rest days, and with partial distributions on Mondays and Thursdays. The fitness behavior of Theme 4 mainly occurred on Sunday, Monday and Friday.

2.5 Related factors to fitness behavior

Based on the findings described above, this study used a geodetector tool to further explore relevant factors that affect the spatial differentiation of fitness behavior of residents. First, an evaluation index system for influencing factors was constructed based on a total of 14 explanatory variables in 6 categories including various elements of the city, environmental conditions, land prices, traffic convenience, population distribution, and location conditions (see Table 2). Based on that, factor detection and interaction detection functions of the geographic detector were used to reveal influences on the spatial selection of fitness behavior of residents in Beijing.

Table 2. Index system of influencing factors on spatiotemporal differentiation of residents' fitness behavior.

<table>
<thead>
<tr>
<th>Influencing Factors</th>
<th>Explanatory factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Various urban elements</td>
<td>Number of food service facilities (X1), number of sports and leisure facilities (X2),</td>
</tr>
<tr>
<td></td>
<td>number of public facilities (X3), number of business service facilities (X4), number</td>
</tr>
<tr>
<td></td>
<td>of accommodation facilities (X5), number of educational and cultural facilities (X6),</td>
</tr>
<tr>
<td></td>
<td>number of companies and enterprises (X7), number of wholesale and retail facilities</td>
</tr>
<tr>
<td></td>
<td>(X8), number of residential service facilities (X9)</td>
</tr>
<tr>
<td>Environmental conditions</td>
<td>Distance to the nearest park (X10)</td>
</tr>
<tr>
<td>Land Price</td>
<td>Land value (X11)</td>
</tr>
<tr>
<td>Convenience of transportation</td>
<td>Distance to bus and subway stations (X12)</td>
</tr>
<tr>
<td>Population Distribution</td>
<td>Population density (X13)</td>
</tr>
<tr>
<td>Location conditions</td>
<td>Distance to CBD (X14)</td>
</tr>
</tbody>
</table>

Using the factor detection function of the geodetector can effectively explore the degree of influence of each influencing factor on the spatial choice of residents' fitness behavior (Table 3). The research results show that number of food service facilities, number of residential service facilities, and number of educational and cultural facilities are the main influencing factors of fitness behavior. At the same time, for further research on fitness behaviors of different themes, we found that the influencing factors of fitness behaviors of various categories are significantly different. The running based fitness behavior is greatly affected by the distribution of residents’ services, catering and accommodation facilities, indicating that it is more dependent on supporting service facilities in nearby communities. The cycling
based fitness behavior is greatly affected by land price, education and catering facilities. The sports fitness behavior in venues and the gym and other professional fitness behaviors are mainly affected by the distribution of educational and cultural facilities, but the former is also affected by the distribution of public facilities, while the latter has greater requirements for the distribution of residential service facilities.

The interaction detection function of the geodetector can indicate whether the combined effect of two different factors will enhance or weaken the factor explanatory power for the dependent variable, and it can effectively reveal the impact of the two factors on the spatial choice of fitness behavior of residents. The results suggest that: (1) In this study, the explanatory power of the 14 influencing factors after pairwise interaction is greater than the explanatory power individually, which indicates that the fitness behavior of residents is jointly restricted by the influencing factors of various dimensions, and any two influencing factors will enhance the factor explanatory power for the dependent variable; (2) Overall, residents are influenced by whether there are catering service facilities available to them, such as restaurants and street snacks, when they exercise. Such areas are often accompanied by extremely high traffic flow or by densely populated areas; (3) 12 indicators of four fitness behaviors were performed using pairwise interactive detection (see Table 3). The results confirm that different theme of fitness behaviors are influenced by different factors. For instance, the fitness behavior of Theme 1 was more likely to be affected by two factors: the distribution of catering facilities and population density; the fitness behavior of Theme 2 was more susceptible to the interaction of land prices and the distribution of educational and cultural facilities.

Table 3. Results of the attribution analysis.
<table>
<thead>
<tr>
<th>Factor action detection</th>
<th>Type characteristics</th>
<th>Main impact factors</th>
<th>Secondary impact factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>X1, X9, X6</td>
<td>X4, X5, X11</td>
<td></td>
</tr>
<tr>
<td>Running based fitness behavior</td>
<td>X9, X1, X5</td>
<td>X8, X6, X3</td>
<td></td>
</tr>
<tr>
<td>Cycling based fitness behavior</td>
<td>X11, X6, X1</td>
<td>X12, X4, X2</td>
<td></td>
</tr>
<tr>
<td>Sports fitness behavior in venues</td>
<td>X6, X11, X1</td>
<td>X4, X3, X12</td>
<td></td>
</tr>
<tr>
<td>Gym and other professional fitness behaviors</td>
<td>X6, X1, X11</td>
<td>X4, X9, X12</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interaction detection</th>
<th>Type characteristics</th>
<th>Interaction factors</th>
<th>Factor explanatory power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>X1∩X3</td>
<td></td>
<td>0.446</td>
</tr>
<tr>
<td>Running based fitness behavior</td>
<td>X1∩X13</td>
<td></td>
<td>0.423</td>
</tr>
<tr>
<td>Cycling based fitness behavior</td>
<td>X6∩X11</td>
<td></td>
<td>0.438</td>
</tr>
<tr>
<td>Sports fitness behavior in venues</td>
<td>X5∩X6</td>
<td></td>
<td>0.354</td>
</tr>
<tr>
<td>Gym and other professional fitness behaviors</td>
<td>X3∩X6</td>
<td></td>
<td>0.432</td>
</tr>
</tbody>
</table>

### 3. Discussion

Firstly, it is necessary to study residents' fitness behavior from the perspective of time, space and text. We have shown that text information mining methods using big data can efficiently process text information in social media data, and geographical spatio-temporal analysis methods can effectively analyze the spatio-temporal information contained within social media data. The combination of the two can be used to fully explore residents' fitness behavior patterns and characteristics hidden within massive data. This research has constructed a complete and efficient framework for the extraction and analysis of fitness-related information from the Weibo social media platform.

Second, we have used LDA theme model technology to analyze fitness behaviors of different themes from the perspective of users. The results of sentiment analysis show that Beijing residents generally have better fitness experience, but residents have poor fitness experience in the gym. Therefore, improving the service facilities in such areas can effectively improve residents' fitness experience. The results of spatial analysis showed that there were significant differences in the aggregation areas of fitness behaviors in the different themes. On the whole, the main hotspots were concentrated near the colleges and residential areas of the North Fourth Ring Road, and show the differentiation characteristics of more in the north and less in the south. Indeed, there are differences in the spatio-temporal distribution of the fitness behaviors of the different themes as the characteristics of each type of fitness are different, and
are affected by factors such as time cost, economic conditions, place environment, and personal subjective wishes.

We found significant differences in the weekly activity intensity and 24-hour activity intensity of fitness behavior in the four different fitness activity themes, and they show obvious temporal characteristics and periodic characteristics. Among them, the fitness behavior of Theme 1 (running) and Theme 2 (cycling) had relatively even distribution of weekly activity intensity and generally occurred during the day. The fitness behavior of Theme 3 and Theme 4 had relatively discrete distributions of weekly activity intensity, and were mainly concentrated at night. These characteristics were mainly related to the type of fitness and its convenience. Fitness behaviors such as running and cycling are less affected by the venue and the environment, and the cost is low. Hence, they are very popular among residents. Additionally, the four fitness behaviors have a higher intensity of daytime activities on Monday (the first day of the working day), and the underlying influence mechanism is worth exploring.

Thirdly, the spatio-temporal distribution of fitness behavior of residents in Beijing is affected by multiple factors. The results of single-factor detection and interaction detection show that the spatial distribution of catering service facilities and public facilities generally had a high explanatory power for the choice of fitness behavior of residents. Additionally, it was also affected by factors such as land price, population density, and the number of companies. From this we can see that when residents choose fitness locations, they will often choose fitness venues near restaurants. Additionally, the neighborhood of bustling business districts is also a favorite fitness place for residents. However, from the perspective of the four types of fitness behaviors, the interaction among factors such as accommodation facilities, educational and cultural facilities, land prices, and population density also has a strong explanatory power. Therefore, rationally strengthening the construction of Beijing's urban fitness service facilities, especially in densely populated areas, and increasing the diversity of fitness service facilities near some residential areas of the Fifth Ring Road, can effectively implement a reasonable distribution of Beijing's urban fitness service facilities, improve residents’ fitness environment and improve residents’ quality of life.

This research also has some shortcomings. First, this research classifies fitness behavior of residents, the number of themes is determined by empirical methods, and the representativeness of the classification results may be insufficient. Secondly, this research only performed investigation on Weibo users. Due to the biased nature of Weibo users (e.g. by age), it is necessary to be cautious to extend the results to the general population. Future research needs to combine big data with small data to achieve a comprehensive understanding of fitness behavior of the whole population, and provide more effective empirical methods and research frameworks for research on residents' daily behaviors and activities.

4. Data And Methods

This research used Sina Weibo, which is a kind of social media data, and it does not involve any private information of individuals. The research was performed in accordance with relevant guidelines and
regulations, and also performed in accordance with the Declaration of Helsinki. The methods used in our research are all obtained through public channels and improved according to research needs.

4.1 Data and pre-processing

This study takes the city within the Fifth Ring Road of Beijing as the research area. This area represents less than 5% of the total area of Beijing, but has nearly half of the population of Beijing. As of the end of 2017, the permanent population density in the Fifth Ring Road in Beijing was about 11,000 people per square kilometer, which was more than 8 times the average population density of the city. It is the main place where residents' fitness behavior occurs.

The social media data came from Sina Weibo. According to research needs, this research has made targeted improvements in data acquisition methods. Sina Weibo official API and web crawler tools were used to crawl Weibo data in Beijing. A total of more than 13 million individual Weibo data items were obtained in the research area in 2017 (i.e., time, location, user UID, text information).

Table 4. Preliminary screening rule chart.

<table>
<thead>
<tr>
<th>Preliminary screening rules</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>The target users sent less than 800 posts in 2017</td>
<td>Excluding advertising account, marketing account and other non-personal account microblog</td>
</tr>
<tr>
<td>Users have posted in Beijing for three months or more</td>
<td>Ensure that social media users are residents of Beijing</td>
</tr>
<tr>
<td>Text content includes fitness related keywords (such as ball, cycling, running, fitness, etc.)</td>
<td>Exclude a large number of posts unrelated to fitness behavior</td>
</tr>
</tbody>
</table>

In order to study fitness behavior of residents in Beijing, we first needed to screen all of the data. Table 4 shows the preliminary screening rules and their processing purposes. According to the preliminary screening rules, more than 250,000 Weibo data points related to fitness behavior of residents in the research area were obtained. Next, the text content of the Weibo data was cleaned. A custom dictionary for cleaning the data for fitness behavior was developed using a stop words database and a keyword thesaurus, and useless information such as emojis, advertisements, and lotteries in the Weibo text were removed.

The administrative division data used in this study comes from National Geomatics Center of China (http://www.ngcc.cn/ngcc/).

4.2 Research methods

4.2.1 Text content recognition

BERT (Bidirectional Encoder Representation from Transformers) is a language model released by Google in November 2018. It can greatly improve the accuracy of text recognition and is a major breakthrough in
natural language processing (NLP) \(^{26}\). We use the BERT model, compiles with the python language, and combines FastAI technology \(^{27}\) to realize the semantic processing of massive microblog text data and extract microblog data related to residents' fitness behavior. By importing a large number of manually labeled Weibo data and multiple iterations, the final recognition accuracy can reach more than 93.8%.

4.2.2 Text theme detection

LDA (latent dirichlet allocation) is a well-known theme model, proposed by David M. Blei et al.\(^{28}\) that has been widely applied in the processing of big data\(^{29-32}\). It is a typical unsupervised learning algorithm that does not require a large number of manual annotations of the training set during training. The core formula is as equation (1):

\[
P(w|d) = P(w|t) \times P(t|d)
\]

In the formula, \(P(w|d)\) refers to the probability that the word \(w\) appears in the document \(d\), \(P(w|t)\) refers to the probability that the word \(w\) may appear in the topic \(t\), and \(P(t|d)\) refers to the probability that the topic \(t\) appears in the document Probability in \(d\).

This study used the python language to call the LDA model to apply subject classification to the Weibo text, and hence provide structured thematic data for the analysis of the characteristics of fitness behavior of residents in Beijing.

4.2.3 Text sentiment extraction

Text sentiment analysis, also known as sentiment orientation analysis, refers to the process of analyzing, processing, summarizing and reasoning texts with subjective sentiment orientation\(^{33}\). There are generally two methods of sentiment analysis, one is based on machine learning, which uses classification analysis to solve sentiment analysis, which requires a large amount of high-precision training data set; the other is based on sentiment dictionary, through calculation the sentiment score of each word and phrase is comprehensively calculated for the sentiment value of the data, which requires a large number of labeled sentiment dictionaries. In this study, considering the characteristics of the data and the scalability of the research framework, a sentiment dictionary-based method is used to perform sentiment analysis on Weibo data.

SnowNLP is a Chinese processing tool based on python language. This research uses the sentiment analysis function of this tool to extract the sentiment value of fitness microblog text information. The result of the operation will indicate the sentiment value of each Weibo, and use a number from 0 to 1 to indicate the probability of whether it is biased towards positive or negative sentiment (0 means negative, 1 means positive).

4.2.4 Geodetector
In order to further analyze the causes of the spatial characteristics of fitness behavior of the residents in the research area, this study used geodetector tools for attribution analysis. Geodetector was developed by Wang Jinfeng and others. It consists of risk detectors, factor detectors, ecological detectors and interaction detectors. It is a set of statistical methods used to explore geographic spatial differentiation and find its explanatory variables. One of the advantages of geodetector over traditional statistics is that it can study the influence of the coupling between two factors on the dependent variable. Fitness behavior of residents is affected by factors such as transportation and public facilities, and there are also coupling effects among influencing factors. The use of geographic detectors effectively revealed the influencing mechanisms on the distributions of fitness behavior of residents.

Declarations

Authorship contribution statement

Bin Tian was responsible for writing the main body of the paper; Bin Meng provided technical and methodological guidance; Guoqing Zhi, Zhenyu Qi, Siyu Chen and Jian Liu helped to process the experimental data.

Author(s)’ disclosure statement(s)

No competing financial interests exist.

References


Figures

Figure 1

Spatial distribution of Bodybuilding behavior theme Weibo data
Figure 2

Time distribution of Bodybuilding behavior theme Weibo data
Figure 3

LDA Intertopic Distance Map (via multidimensional scaling)
Figure 4

Emotional mean value of four thematic behaviors
Figure 5

Nuclear density map of fitness behaviors for each theme
Figure 6

Map of “hot-spots” of fitness behavior of each of four themes identified from social media data.
### Figure 7

Time distribution of fitness behavior