Academic Journal of Management and Social Sciences
ISSN: 2958-4396 | Vol. 10, No. 3, 2025

Decoding Public Sentiments: A Text Mining Analysis of National Fitness Empowering by Digital Intelligence

Zhen Ma, Haohai Gu

School of Data Science, Xi'an Eurasia University, Xi'an Shaanxi, 710065, China

Abstract: This study focuses on the public sentiments of national fitness empowering by digital intelligence (DI), and analyzes the data of 14086 comments on social media platforms Weibo. By combining the comprehensive research methods of TF-IDF, Naive Bayes sentiment classification model and LDA topic model, the results show that positive sentiments (accounting for 85.1%) focus on the technological advantages (such as AI, data function) and service upgrading of intelligent fitness, neutral sentiments (3.0%) focus on rural promotion, consumption costs and other popularization issues, and negative sentiments (12.0%) expose facility function defects and management contradictions. This study expands the field of digital intelligence empowerment of national fitness, aiming to provide support for the industry's precise optimization of services and policy design, and help the high-quality development of the digital intelligence fitness industry.

Keywords: National Fitness; Digital Intelligence; Sentiment Analysis; LDA.

1. Introduction

With the deep integration of artificial intelligence, cloud computing, big data, and blockchain technologies, digital intelligence (DI), as a comprehensive technological system, is reshaping development paradigms of the national fitness. By 2024, China's intelligent fitness app user base had reached 480 million, with 350 million active users, marking a 10.5% year-on-year growth [1].

Specifically, national fitness empowering by DI refers to the transformation of public fitness activities and related data into analyzable information resources through digital means, enabling the perception, analysis, and optimized control of fitness behaviors. Existing research on DI applications in national fitness exhibits divergent focuses: international studies prioritize technological enhancements in user experience and smart management systems. For instance, Carrington P et a. (2019) proposed optimizing sports venues and services using AR/VR technologies [2]. Masterman G (2014) advocated for intelligent information management systems to enrich fitness experiences [3]. Domestic research, however, emphasizes comprehensive DI fitness applications and policy recommendations, mainly focusing on three perspectives: technology application, people's livelihood supply and demand, and quality development. From the perspective of technology application, Shi Xiaoqiang et al. (2021) believe that based on big data and blockchain technology, it is necessary to establish a data sharing mechanism for interconnection between relevant departments [4]. As for people's livelihood supply and demand, Zhang Yuqi (2023) proposed that a digital and intelligent sports service platform can effectively promote the public's recognition of the concept of "big health" and provide key support for supply-side structural reform [5]. From the perspective of quality development, Wang et al. (2024) explored the value and dilemma of digital intelligence empowering the high-quality development of national fitness, and proposed four development paths: strengthening concepts, improving guarantees, consolidating foundations, and improving depth [6]. Comprehensive analysis shows that the existing research pays little attention to the public sentiments of digital intelligence empowering national fitness, and most of the research methods use qualitative analysis methods such as literature analysis, logical analysis method, expert interview method, rarely combing text mining technology and machine learning model to analyze the key factors of national fitness empowering by DI.

Therefore, this study will focus on the sentiment performance of national fitness empowering by DI. Combing with the Naive Bayes sentiment classification model, text mining technology (TF-IDF and LDA topic model), we conduct in-depth discussion and analysis of the public commentary of digital intelligence fitness on social media. The results can provide reference implementation suggestions for the high-quality development of the sports and fitness industry.

2. Text Mining and Sentiment Analysis Methods

2.1. Sentiment Analysis

Sentiment analysis, also known as "opinion mining", refers to the use of natural language processing (NLP), text parsing, and computational linguistics to reveal users' emotional preferences. The most common sentiment analysis method is based on statistical learning, which realizes sentiment judgment by training a classification model. In this study, the widely used Naive Bayes model, which is based on Bayes' theorem, is used to select the category with the largest posterior probability as the prediction result, where $P(class)$ is the prior probability, which represents the initial probability of the category, and $P\ (feature|class)$ is the likelihood probability, which represents the probability of the feature under the category condition, which is specifically expressed as:

$$Prediction\ class = arg \max_{class} P(class) \cdot \prod_{i=1}^{n} P\left(feature_i \mid class\right) \tag{1}$$

As for the evaluation of sentiment classification performance, the commonly used indicators include accuracy, precision, recall, F1 value, ROC curve and AUC. Except for the ROC curve, all evaluation indicators ranged between $[0, 1]$, and the closer to 1 was, the better the model performance.

2.2. Text Mining

2.2.1. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a weighting technique widely used in the field of information retrieval and text mining, and it performs well in text relevance ranking. TF-IDF calculates the importance of a word in a document by combining the word frequency and the inverse document frequency, the higher the TF-IDF value, the more important the word is in the current document d, and the less common it is in the whole corpus D, and the calculation formula is as follows:

$$TF-IDF = TF_{(t,d)} \times IDF_{(t,D)} \tag{2}$$

where:

$TF_{(t,d)}$ indicates the frequency of the word t in the document d;

$IDF_{(t,D)}$ indicates the inverse document frequency of the word t in the whole corpus D.

2.2.2. LDA Topic Model

Latent Dirichlet Allocation (LDA) is a document topic generation model, also known as a "three-layer Bayesian probability model", which includes a three-layer structure of "text-subject-word". As an unsupervised learning algorithm, LDA is able to discover hidden topic structures from large-scale text corpa. Specifically, for each document in Document Set *D*, the algorithm generation process is as follows:

(1) Select *N*, *N obeys* the $Poisson(\xi)$ distribution, and *N* represents the length of the document.

(2) Select θ, θ obeys the $Dirichlet(\alpha)$ distribution, representing the probability of the occurrence of the subject, and α represent the parameters of the Dirichlet distribution.

(3) For each word that makes up *N*, choosing the subject z_n and z_n which obey the $Multinomial(\theta)$ distribution.

In this process, β is a $K \times V$ matrix, where *K* represents the number of topics and V represents the number of words. $\beta_{ij} = P(w^j = 1 | z^i = 1)$ indicates the probability that a topic z^i generate a word w^j, θ is the Dirichlet random variable, by finding product of the joint distribution of θ, the subject z, the word w and edge distribution of the document, the union probability of the entire document set is obtained as follows:

$$P(D|\alpha,\beta) = \prod_{d=1}^{M} \int P\left(\theta_d|\alpha\right)\left(\prod_{n=1}^{N_d} \sum_{z_{d_n}} P\left(z_{d_n}|\theta_d\right) P\left(w_{d_n}|z_{d_n},\beta\right)\right) d\theta_d \tag{3}$$

In this study, LDA model is constructed to identify the topic of public commentary on social media.

3. Study Implementation and Analysis

3.1. Data Collection and Processing

The study collected data from Weibo, a leading Chinese social media platform with active public engagement, and used a web crawler. Setting the initial Uniform Resource Locator (URL) of the web page, we crawled public commentary containing keywords “digital intelligence”, “intelligence fitness” and “national fitness”. Finally, 17,347 data samples were collected, which published from January 2023 to March 2025.

First, we cleaned the raw data by addressing missing values, outliers, and duplicates to ensure quality control, ultimately retaining 14,086 valid samples. Next, we performed text preprocessing using the Jieba word segmentation tool to tokenize Weibo comments into meaningful lexical sequences. Subsequently, we removed stop words to reduce noise and improve processing efficiency, leveraging a predefined list of 2,312 terms (stopwords.txt) to filter non-informative tokens. Table 1 shows the comparison results of the data before and after processing.

Table 1. Comparison of data processing results

Original text	Post-processing text
I found a good and inexpensive fitness product. I recommend everyone to take a look: TANG music flywheel smart spinning bike, a home fitness cycling sports equipment.	['found', 'good', 'inexpensive, ' fitness', ' product', 'recommended', 'everyone', 'take', 'look', 'TANG', 'music', 'flywheel', 'smart', 'spinning', 'bike', 'home', 'fitness', 'cycling', 'sports equipment']

3.2. Basic Statistical Analysis of Text Mining

Based on the above processed text data, TF-IDF was used to preliminarily evaluate the core concerns of users about national fitness empowering by DI. As shown in Table 2, the top five keywords are "fitness", "intelligence", "sports", "health", and "gym". It can be seen that the content of the text generally shows around fitness, smart technology and healthy lifestyle, reflecting the public's high concern and extensive demand for intelligent health management and exercise efficiency.

3.3. Sentiment Analysis of Public Commentary

Firstly, this study matched the comment text with the sentiment dictionary, and labeled each comment text with a sentiment score between 0-1, dividing data into negative emotion (<0.3), neutral emotion (0.3 to 0.7), and positive emotion (>0.7). Then, the study established a sentiment classification model for 14,086 data samples, applying stratified sampling to split the dataset into training and test sets at a 7:3 ratio, and used the trained model to predict outcomes on the test set.

Table 2. The TF-IDF of top five keywords

Keyword	TF-IDF
Fitness	4.83
Intelligent	4.39
Sports	3.50
Healthy	2.23
Gym	2.21

According to the training results, the initial Naive Bayes model performs well in the positive and neutral categories, but has some errors in the negative categories. Therefore, we use the grid search method to optimize the model parameters, in order to find the best hyperparameter configuration. By adjusting the bigram feature ('ngram_range=(1,2)'), the feature number limit ('max_features'), the document frequency filter ('min_df' and 'max_df'), tuning the smoothing parameter of the Naive Bayes ('alpha'), adding cross-validation (5-fold) to prevent overfitting, and using multi-core parallel acceleration ('n_jobs=-1') to correct the probability calculation method of the ROC curve, the AUC values of all sentiment classifications are greater than the lowest value of 0.5 (As shown in Figure 1), and the model has a better effect. From Table 3, it can be seen that the accuracy of the optimized model is improved by 3.00%, and other evaluation indicators are better than those before optimization.

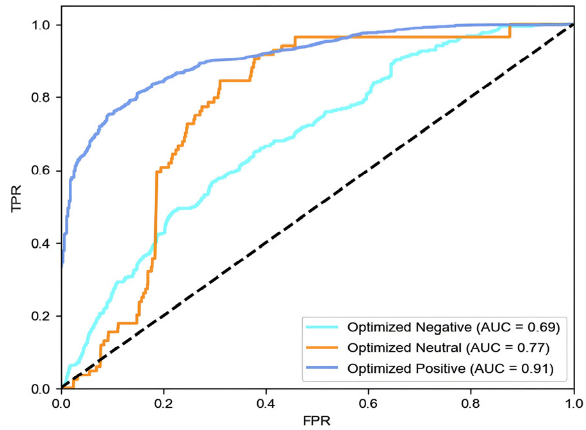

Fig 1. ROC curve of the optimized model

Table 3. Comparison of the Naive Bayes model before and after optimization

		Accuracy	Precision	Recall	F1 value	AUC
Initial model	Negative sentiments	0.86	0.01	0.01	0.01	0.53
	Neutral sentiments		0.95	0.06	0.12	0.50
	Positive sentiments		0.86	0.98	0.92	0.52
Optimized model	Negative sentiments	0.89	0.69	0.46	0.55	0.69
	Neutral sentiments		0.97	0.07	0.13	0.77
	Positive sentiments		0.90	0.98	0.94	0.91

The optimized Naive Bayes model was used to reclassify the sentiments on the dataset, and finally obtained 11983 positive sentiment texts (85.07%), 412 neutral sentiment texts (2.92%) and 1691 negative sentiment texts (12.00%).

3.4. Analysis of LDA Topic Model

Based on the text data of positive, neutral and negative sentiment texts, the LDA theme analysis was carried out to explore the similarities and differences of the public's attention to national fitness empowering by DI. Firstly, the number of topics for each sentiment classification is judged according to the perplexity, and the lower the perplexity, the better the fit of the model to the data, and the stronger the prediction ability. As seen in Figure 2, the number of best topics' number corresponding to the lowest point of perplexity for positive emotion, neutral emotion, and negative emotion is 6, 6, and 5.

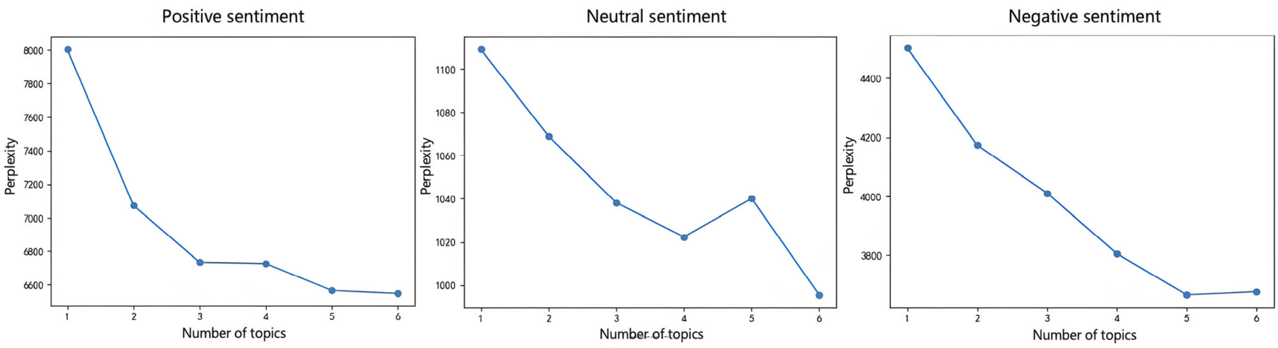

Fig 2. Perplexity of different sentiment classifications

According to the best number of topics for each sentiment category, the study uses LDA model to cluster the topics, and output the top 5 keywords with the largest probability in each topic under different sentiments, as well as the proportion of each topic in the text. Meanwhile, we summarized and analyzed the topics. For example, the keywords "Fitness", "Intelligence", "Exercise", "Health", and "AI" in Theme 1 of positive sentiments can be summarized as the topic of "Intelligent fitness", and the detailed topic clustering results are shown in Table 4.

Table 4. Clustering results of LDA topics for different sentiment classifications

Sentiment Classification	Topic Number	Topic	Keywords (TOP 5)	Percentage
Positive	1	Intelligent fitness	Fitness, Intelligence, Exercise, Health, AI	45.01%
	2	Fitness services	Service, Development, Society, Sports, Promotion	17.77%
	3	Community fitness	Fitness, intelligence, Community, Living, Apartment	15.18%
	4	Fitness technology	Intelligence, Technology, APP, Company, Fitness	7.69%
	5	Leisure and fitness	Parks, Facilities, City, Fitness, Leisure	7.42%
	6	Fitness industry development	Development, Sports, Industry, Fitness, Equipment	6.93%
Neutral	1	Rural intelligent fitness	Intelligence, fitness, rural, installation, agricultural materials	25.52%
	2	Intelligent home	Intelligence, Fitness, Sports, Equipment, Watches	24.01%
	3	Intelligent consumption	Consumption, Sports, Intelligence, Cost, Data	13.74%
	4	Intelligent devices	Fitness, Sports, Artificial Intelligence, Health, Keep	13.38%
	5	Fitness subsidy	Consumption, fitness, subsidies, intelligence, support	11.82%
	6	Intelligent fitness renovation	Community, implementation, intelligence, fitness, renovation	11.53%
Negative	1	Intelligent fitness facilities	Intelligence, Fitness, Facilities, Function, Comfort	30.11%
	2	Community fitness renovation	Community, Renovation, Facilities, Fitness, Environment	19.85%
	3	Intelligent fitness management	Community, management, fitness, exercise, intelligence	17.87%
	4	Diagnosis and treatment of fitness	Intelligent diagnosis and treatment, fitness, personal account, medical insurance, payment	16.55%
	5	Home fitness	Intelligence, treadmill, watch, convenience, home	15.62%

According to the results of the LDA topic analysis, positive, neutral, and negative sentiments show different concerns in the DI fitness. Positive sentiments focus on the technological advantages and service enhancements of intelligent fitness, such as the theme of "Fitness services", which focus on industrial development and consumption upgrading, showing positive expectations. Neutral sentiments discuss the popularization and application of intelligent fitness, for example, "rural intelligent fitness" discusses rural promotion, and "Intelligent consumption" involves fitness costs and consumption issues. Negative sentiments reveal problems in intelligent fitness facilities and management, reflecting dissatisfaction with the status. For example, "Intelligent fitness facilities" criticize equipment for being inadequate, and "Community fitness renovation" discuss charging and management issues. Overall, all of the three sentiments share a common focus on the long-term development of DI fitness.

Based on this, we put forward industry and policy optimization suggestions. For users with positive feedback, companies can deepen the application of intelligent fitness, increasing research and development to improve the experience of intelligent fitness tools, in order to consolidate the market growth. For the users with neutral feedback, it is necessary to refine the scenario strategy, promote DI fitness in rural areas accurately, strengthen supervision of household products, and guide benign consumption. For the users with negative feedback, it is necessary to improve maintenance mechanism of fitness facilities, and strengthen the supervision of technology application.

4. Summary

This study focuses on the public sentiment of national fitness empowered by digital intelligence, combining the Naive Bayes sentiment analysis model, TF-IDF and LDA topic model to analysis social media comment data. The results reveal the public's diverse attitudes towards DI fitness. It showed that positive sentiments focused on the technological advantages and service improvement of DI fitness, neutral sentiments discussed its popularity and economy, while negative sentiments revealed problems in facilities and management. The results of this study can provide a decision-making basis for policy design, technology iteration and operation management, helping the DI fitness industry achieve a balance between technological upgrading and user needs, and finally promoting the high-quality development of national fitness.

Acknowledgments

Supported by: Shaanxi Provincial Sports Bureau 2024 Regular Research Project "Research on Digital Intelligence Technology Empowering Public Participation in National Fitness" (Project No. 20240201).

References

[1] China Report Hall. Analysis of the current situation of the fitness software industry in 2024: the scale of fitness software users in China will reach 480 million. Information on: https://www.chinabgao. com/freereport/ 95685. html.2024-08-26.

[2] Carrington P, Bernart E. Computer science and sports: The digital evolution of physical competition. XRDS: Crossroads, The ACM Magazine for Students, 2019, Vol. 25(No. 4): p. 8-9.

[3] [Masterman G. Strategic Sports Event Management. 2014, p. 120-124.

[4] SHI Xiaoqiang, DAI Jian. Research on the Situation Requirement, Realistic Foundation and Objective Measures of the Development of National Fitness in China during the Period of the 14th Five-Year Plan. China Sport Science,2021, Vol. 41(No. 04): p. 3-13+59.

[5] Zhang Yuqi. The "digital intelligence" of national fitness and the upgrading of public sports consumption. Chinese Sports Science Society. Abstract Collection of Papers of the 13th National Sports Science Conference Poster Exchange (Sports Industry Branch). Shanghai University of Sport; ,2023, p. 14055.

[6] WANG Ying, ZHANG Feng-biao. High-Quality Development of Digital Intelligence Empowering National Fitness. Sports Culture Guide,2024(No. 01): p. 35-41.