Media and Communication Research (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/mediacr.2024.050407
ISSN 2523-2584 Vol. 5 Num. 4

The Topic Mining for Sentiment Analysis of Tokyo Olympic Games by Using LDA and Sequence Association Rules

Kuanwei Huang[1,*], Weicong Mo[1]

[1]*Business School, Lingnan Normal University, Zhanjiang, China*
[*]*Corresponding author*

Keywords: Tokyo Olympics; topic mining; sentiment analysis; sequence association rules

Abstract: This paper studies the topic of " Tokyo 2020(summer) Olympics " in Sina Weibo as the research target for topic mining by using latent dirichlet allocation (LDA) model and applying the sequence association rules algorithm to obtain the keywords among the different topic categories. Then we conduct the text content with sentiment analysis. Among the topics of the Tokyo Olympics, more than 74% of netizens have positive comments on the association rules under the keyword “gold medal”, which parses out the competition sports that attract public attention, such as swimming, table tennis, shooting and other competitions that are well received and that will make the netizens pay more attention within. Finally, this study expects that it would continue to discover the sports hotspots for relevant departments with reference and assist the direction for the development of sports in China.

1. Introduction

In the past few decades, the means by which people access information about the Olympic has continuously evolved, progressing from newspapers and radio to television and the internet. The 2020 Tokyo Olympics, in particular, highlighted the prowess of social media in this evolving landscape. To mitigate the regrettable absence of on-site spectators, the International Olympic Committee, in the opening of the “Tokyo Olympics Social and Digital Media Guidelines,” explicitly encourages athletes and other individuals certified for the 2020 Tokyo Olympics to share their experiences through social and digital media, along with their friends, family, and supporters[1]. This marks a discernible broadening of access to social media compared to preceding Olympics, with social media emerging as the primary conduit for acquiring real-time information about Olympic events. Besides, the rapid expansion of the internet, coupled with the proliferation of various network terminal devices, has led to an increasing number of individuals engaging with social networking platforms. This trend underscores the dynamic landscape of internet usage in China, emphasizing the significant role played by platforms like Sina Weibo in facilitating social interactions and information dissemination.

Based on the aforementioned data, it is evident that the number of internet users continues to grow, enabling netizens to express their opinions and viewpoints on specific events through online social platforms. Valuable public sentiment information can be extracted from online discourse. Weibo, as a mainstream social networking platform, serves as a repository for a vast amount of online discourse.

Social events are disseminated through Weibo media, and under the impetus of netizens, these events generate trending topics within the Weibo community. Topics on the Weibo hot search list typically represent the concentrated search and attention behaviors of Weibo users during a specific period. Under a single topic, there are reposted posts and comments related to that topic. Weibo public opinion is regarded as the collective opinion of netizens on a widely discussed topic within a certain period, and it constitutes an aggregation of public psychology with a certain degree of publicity[2] .

This study, therefore, aims to conduct a big data analysis based on Weibo hot search related to the Tokyo Olympics, and examines the sports' themes and online sentiments that users are interested in. Furthermore, it aims to delineate the associated content and features of these themes, providing an in-depth analysis of the topics that attract significant public attention.

2. Literature review

2.1 Latent Dirichlet Allocation (LDA) model

The LDA model, also known as a three-layer Bayesian probability model. Its fundamental concept involves representing document-topic and topic-word distributions as multinomial distributions with a Dirichlet prior probability[3]. The LDA model utilizes unsupervised learning algorithms to unearth latent topic information embedded in the corpus, providing a probabilistic representation of the topics for each text in the corpus. Considering diverse requirements, numerous studies have extensively explored and built upon the LDA model[4].Williams and Betak utilized Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) to identify topics within a text database related to railway equipment incidents. Yong and colleagues conducted comparative studies between LDA and other topic models, experimenting with representative corpora. Bastani et al employed LDA as an intelligent method to analyze consumer complaints, aiming to extract potential topics from the complaint texts[5-7]. Due to the widespread attention and in-depth research on LDA across various industries, this paper selects the LDA model for topic mining in large-scale Weibo comment datasets.

2.2 Text sentiment analysis

Text sentiment analysis refers to the process of utilizing natural language processing and text classification techniques to identify users' subjective emotions, opinions, and attitudes from textual data [8]. This methodology finds widespread applications in sentiment monitoring and information prediction. The inception of social media sentiment analysis can be traced back to the analysis of Twitter's social media data [9]. Bollen, based on Twitter data, categorized emotions into six emotional dimensions, identifying the most representative emotions each day[10]. Dodds attempted to elucidate patterns of happiness from the perspective of sentiment analysis. Sentiment analysis methods can be broadly categorized into two types: those based on sentiment dictionaries and those leveraging machine learning[11]. Sentiment analysis based on dictionaries involves extracting feature words from the test text, looking up the sentiment value of these words in a sentiment dictionary, and then classifying the sentiment based on the accumulated sentiment values [12]. Two common approaches for selecting sentiment dictionaries are to either reference existing ones, such as HowNet, Senti-WordNet, and Inquirers, or to construct dictionaries based on research data, as done by scholars like Feldman, who extracted sentiment words through manual annotation and Bootstrapping[13]. Machine learning-based sentiment analysis methods involve training a classifier on a text corpus and subsequently using the classifier to classify new texts [14]. With the continuous advancement of artificial intelligence and deep learning research, many scholars have applied deep learning to sentiment analysis. Pang pioneered the use of machine learning for sentiment analysis of movie review texts. In sum, sentiment analysis relies on extensive sentiment corpora, making it challenging

for individuals to label sufficient training data[15]. Individual annotations often carry some subjectivity, leading to lower accuracy in trained models. This represents a bottleneck in current sentiment analysis research.

2.3 Association rules

Association rules, as a common data mining method, initially found application in association analysis between products and have now been widely adopted in various fields [16]. In the field of library and information science, association rules are commonly used to mine the correlation between keywords. Peng selected conference papers from iConference as their research subjects, applying association rules to mine associations between keyword pairs in five conference documents[17]. They derived main research topics from these association rules. Zhang explored keyword association rules in the field of health informatics articles, summarizing the key research content in this domain[18].

The evaluation of association rules mainly involves three indicators: Support, Confidence, and Lift [19]. Taking text mining as an example, let I be the set of all words, and any subset of I is referred to as an itemset, denoted to X. The number of words contained in each itemset is called the length of the itemset, and an itemset with length k is called a k-itemset. The entire corpus is denoted as D, where |D| represents the number of documents in the corpus. The frequency with which an itemset appears in the corpus D is called its support, denoted as count(X). The relevant definitions are as follows:

(1) Support: support(X)=count(X)/|D|. Given a minimum support threshold denoted to support (min), if support(X)≥support(min),then X is a frequent itemset.

(2) Confidence: confidence(X⇒Y)=support(XY)/support .(XY) represents the proportion of texts that contain both X and Y among those containing X. When the confidence is greater than the threshold, the corresponding rule is a strong association rule.

(3) Lift: lift(X⇒Y)=support(XY)/(support(X)support(Y)). This indicator serves as the criterion for evaluating association rules. If the confidence is less than 1, it is generally considered an ineffective rule. A confidence equal to 1 indicates that X and Y are independent of each other. If the confidence exceeds 1, especially surpassing 3, it suggests that the mined association rules have significant value.

This study will utilize topic extraction as a foundation, applying association rule algorithms to conduct association mining on keywords related to the identified topics. Further analysis will be performed on the relationships between topic keywords, aiming to enhance the interpretation of topics.

2.4 Word2Vec

Word2Vec is a word embedding model developed by Google in 2013 based on the principles of deep learning. Its primary purpose is to transform textual information from an unstructured form to a vectorized form [20]. Word2Vec allows the transformation of words into vector forms by studying textual content, representing semantic information through word vectors [21]. Additionally, as a natural language processing tool, one of Word2Vec's major features is addressing the dimensionality challenge by representing word features based on contextual information. Based on different methods of training word vectors, Word2Vec can be further categorized into CBOW and Skip-Gram. CBOW uses contextual words as input to predict information about the target word, while Skip-Gram uses the target word as input to predict contextual information. Through a comparative analysis of these two training methods, it is observed that CBOW exhibits better performance in handling small-scale corpora, whereas Skip-Gram demonstrates superior predictive effectiveness when dealing with large-scale corpora. Word2Vec operates on the idea that two words with similar contextual environments have approximate meanings. After extensive training on a large corpus, Word2Vec can effectively represent word vectors, and the quantification of the numerical relationship between word pairs is

achieved through calculating the cosine similarity of their vectors. This study utilizes a richly collected corpus from social media to train the Skip-Gram model of Word2Vec and calculate the similarity of word pairs in Chinese.

3. Data pre-processing

Sina Weibo is a microblogging service characterized by its low entry barrier, instant and widespread dissemination, and features such as search and sharing. It has become a crucial platform for the propagation of online public sentiment. In terms of information source selection, this study opts for the widely followed trending topic on Sina Weibo, "Tokyo Olympics," as the subject of research for collecting data on sentiment evolution and establishing a thematic space. The search index related to the Tokyo Olympics reached its lowest point at 600,000 times, while it peaked between July 29-30, reaching as high as 12 million times. Moreover, an examination of the information index reveals multiple peaks since July 23, with the highest peak reaching 70 million times. From this, it can be observed that the Baidu Index for the Tokyo Olympics exhibits characteristics such as rapid outbreak of public sentiment and a prolonged duration. It possesses distinct emotional propagation features in the new media environment, manifesting in both mobile and non-mobile network sentiments.

This study takes the topic “Tokyo Olympics” on Sina Weibo as an example and utilizes web scraping to collect user data. The acquired data fields include usernames, microblog content, timestamp, mobile tool types, and the number of reposts, comments, and likes. The study selects the opening day of the Tokyo Olympics, “July 23,” as the starting point for analyzing the evolution of sentiment in online public discourse. The endpoint is set at the closing ceremony on August 8, using “Tokyo Olympics” as the overarching topic, and collecting all data from microblogs under this topic. The gathered raw data is semi-structured and contains noise. After preprocessing, a dataset containing 6,822 Chinese comments is obtained.

4. Results

4.1 Results of LDA

In Figure 1, the optimal result was achieved when the number of topics was determined to be 16. Each topic was filtered to obtain the most representative vocabulary from 30 significant keywords. Table 1 shows that in the “Tokyo Olympics” topic, high-concept topics can be extracted, namely, two major categories of topics distinguished by competitive sports and non-competitive sports. Non-competition topics are challenging to summarize topic names from keywords, so they are differentiated using alphabetical letters.

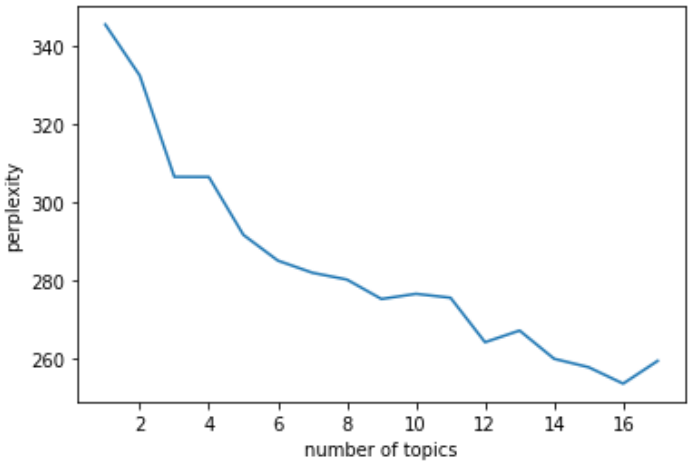

Figure 1: Number of topics by LDA

Table 1: The results of topic mining for keywords (translate in English)

Topic category	Topic	The high-frequency keywords for the topics
Competitive sports	Volleyball	Women's volleyball, Sports, Athletes, Spirit, Arena, Strength, Competition, Sound
	Table Tennis	Video, Ma Long, Fan Zhendong, Men's single table tennis, Watching, Wonderful, Competition, Coach
	Swimming	Women, Gold medal, Finals, Men, Silver medal, Bronze medal, Swimming, Theme
	Table Tennis	Table Tennis, Finals, Silver medal, Competition, Champion, Mixed Doubles, Win, Gold medal
	Judo	Gold medal, Country, National flag, Face mask, Judo, Motherland, Theme, National anthem
	Weightlifting	Weightlifting, Men, Gold medal, Champion, Legion, Video, Athlete, Shi Zhiyong
	Basketball	Competition, Group matches, Women's basketball, Team, Missed, Player, leading, Versus
	Shooting	Opening ceremony, Netizen, First gold medal, Yang Qian, Shooting, Sport, video, Culture
	Diving	Diving, Closing ceremony, Quan Hongchan, Winning the gold medal, video, Medal table, Five rings, Chen Yuxi
	Athletics	Su Bingtian, Finals, 100 meter athletics, Grades, Men, Semifinals, History, Athletics
	Skateboarding	Epidemic, COVID-19, Competition, Participation, Media, Theme, Athletes, Skateboarding
Non-Competitive sports	A	Referee, Motion, Strength, Rule, Medal, Team member, eyes, Foul
	B	Opening ceremony, Video, Live broadcast, Life, Competition, Art, audience
	C	Olympic Athletes, Official, Lottery draw, Assistance, Dream, Schedule, Set, Value
	D	Medal, Competition, Kids, Representative, Influence, Friend, Training, Problem
	E	Athlete, Sports, World, Coach, Competition, Arena, Country, Video

4.2 Results of Association Rules

This study employs the Apriori algorithm to mine association rules among key topics. Initially, each of the 99 distinct keywords undergoes dimensionality enhancement using word2vec, resulting in a vector space of 100 dimensions. Subsequently, Python is utilized for association rule analysis on the dimensionally enhanced textual topic keywords. The parameters are set with a minimum support of 0.35 and a minimum confidence of 0.01. The objective is to explore potential associations among keywords across different topics (i.e., inter-topic). Table 2 presents some of the association rules identified between specific keywords.

Table 2: Results of Inter-Topic Keyword Association Mining (partial illustration)

Association rules	Confidence	Support	Lift
{ Judo,First Gold medal }⇒{ Legion,Yang Qian,Skateboarding}	1	0.4	2.381
{ Medal table,Swimming,Chen Yuxi }⇒{ Judo,Skateboarding,Diving }	1	0.4	2.381
{ Medal table,Swimming,Chen Yuxi }⇒{ Judo,Diving }	1	0.4	2.326
{ Medal table,Swimming,Chen Yuxi }⇒{ Skateboarding,Diving }	1	0.4	2.326
{ Swimming,Skateboarding,Chen Yuxi }⇒{ Judo,Diving }	1	0.41	2.326
{ Judo,Swimming,Chen Yuxi }⇒{ Skateboarding,Diving }	1	0.41	2.326
{ Win gold medal,Swimming,Skateboarding,Chen Yuxi }⇒{ Judo,Diving }	1	0.4	2.326
{ Win gold medal,Judo,Swimming,Chen Yuxi }⇒{ Skateboarding,Diving }	1	0.4	2.326
{ Medal table,Swimming,Skateboarding,Chen Yuxi }⇒{ Judo,Diving }	1	0.4	2.326

To further analyze the implicit keyword knowledge contained in these association rules, the keyword “Win gold medal”, “Gold medal” or “First gold medal” from the competition sports are used as the antecedent set to categorize and filter the obtained association rules. It is discovered that the consequent set shares keywords with six competition themes: table tennis, swimming, weightlifting, shooting, diving, and skateboarding. These 10 keywords include: “watching”, “swimming”, “women's”, “weightlifting”, “legion”, “Shi Zhiyong”, “Yang Qian”, “Quan Hongchan”, “Chen Yuxi”, and “skateboarding”. Using these 10 keywords as the antecedent set, association algorithms are applied to the consequent set of non-competition sports, resulting in keywords such as “athletes”, “medals”, “videos”, and “training”. The relationships obtained from associating the topic of the competition sports and the non-competition sports keywords in sequence are presented in Table 3.

The results have achieved the discovery of latent knowledge within a specific domain's textual collection. The results of sequence association rule have successfully represented the semantic dimensions of textual content. Extracting information aids in uncovering valuable implicit knowledge and facilitates a deeper understanding of knowledge within a specific domain. Therefore, through association rules, knowledge extraction can not only be realized in massive texts but also effectively describes the semantics between different pieces of knowledge.

Table 2: Results of Sequence Association Rules (Partial illustration)

Sequence association rules of competition sports and non-Competition sports topic	Confidence	Support	Lift
{ Judo,Skateboarding,Diving }⇒{ Medal table,Swimming,Chen Yuxi }	0.952	0.4	2.381
{ Swimming,Skateboarding,Diving }⇒{ Medal table,Judo,Chen Yuxi }	0.976	0.4	2.38
{ Swimming,Skateboarding,Chen Yuxi }⇒{ Medal table,Judo,Diving }	0.976	0.4	2.38
{ Judo,Swimming,Diving }⇒{ Medal table,Skateboarding,Chen Yuxi }	0.976	0.4	2.38
{ Judo,Swimming,Chen Yuxi }⇒{ Medal table,Skateboarding,Diving }	0.976	0.4	2.38
{ Skateboarding,Diving }⇒{ Medal table,Swimming,Chen Yuxi }	0.93	0.4	2.326
{ Judo,Skateboarding,Chen Yuxi }⇒{ Medal table,Swimming,Diving }	0.93	0.4	2.326
{ Skateboarding,Diving }⇒{ Medal table,Judo,Swimming,Chen Yuxi }	0.93	0.4	2.326
{ Judo,Skateboarding,Diving }⇒{ Medal table,Swimming }	0.952	0.4	2.323
{ Judo,Skateboarding,Chen Yuxi,Diving }⇒{ Medal table,Swimming }	0.952	0.4	2.323

4.3 Results of sentiment analysis

Using the sentiment orientation analysis interface from Baidu AI Open Platform, we determined the trinary sentiment attitudes of netizens in the comments, categorizing them as positive, neutral, or negative which denoted to the symbol "+ ", " | ", and "-" respectively. Subsequently, the analysis focused on the keyword set "Gold medal" extracted as the antecedent from competition sports topic related to competitive events and the keyword set from non-competition sports topic as the consequent.

In terms of proportions, both competitive and non-competitive sports topic exhibit positive sentiment percentages of over 74%. Among the competitive events, shooting shows the highest positive sentiment (99%), followed by Swimming (90%). Table Tennis and Skateboarding a relatively lower positive sentiment compared to other competitive events (80%). Overall, the positive attitudes and emotions towards sporting competitions in China receive favorable evaluations. Particularly, in the thematic analysis focused on the "Gold Medal" during the Tokyo Olympics, the sentiments expressed by netizens significantly tend to be positive.

5. Conclusions

The topic mining for the Tokyo Olympics themes on the Weibo platform reveal that netizens with the key word "Gold Medal" in both competition sports and non-competition sports show active (i.e., positive) sentiments, accounting for over 74% of the mainstream. By considering the competitive theme "Gold Medal" as the antecedent set, filtering out the associated rules containing keywords of competitive theme as the antecedent set, and secondarily selecting rules with the consequent set containing keywords of non-competitive themes, we obtained the sequence association rules between

different competitive and non-competitive topic. This study utilized LDA and sequential association rules to uncover sports activities from the Tokyo Olympics' themed comments that have the potential to compete for gold medals. It would urge relevant authorities to play a more pivotal role, making full use of available resources to propel the flourishing of sports culture. Furthermore, the results of this study, highlighting outstanding athlete in sports such as diving and athletics, that could be the ambassadors for these sports so that to promote the development of sports, encourage to host official and grassroots sports. We expect that it would sustain the trends in sports development and the sports culture among the populace, encouraging active participation of the entire population in sports activities in China.

References

[1] Yang, B. Social media brings multiple impacts to the Tokyo Olympics. Youth Journalist,2021,15, 97-98.

[2] Jiao, D. W. Emotions and Their Influence in Weibo Public Opinion. Jiang-huai Tribune, 2013,(3):129-132.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. Journal of machine Learning research,2003,3, 993-1022.

[4] Tan Chunhui, Xiong Mengyuan. Comparative analysis on the evolution of hot topics in data mining research at home and abroad based on LDA model [J]. Information scienc,2021,39(4):174-185.

[5] Williams, T., &Betak, J.A comparison of LSA and LDA for the analysis of railroad accident text. Journal of Ubiquitous Systems and Pervasive Networks, 2019,11(1), 11-15.

[6] YONG C, HUI Z, RUI L, et al. Experimental explorations on short text topic mining between LDA and NMF based schemes[J]. Knowledge-Based Systems,2019,163(1):1-13.

[7] BASTANI K, NAMAVARI H, SHAFFER J. Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications, 2019,127:256-271.

[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems,2013,26, 3111-3119.

[9] GO A,BHAYANI R,HUANG L. Twitter sentiment classification using distant supervision[R]. Cs224n project report. Palo Alto: Stanford University,2009.

[10] BOLLEN J,PEPE A,MAO H. Modeling public mood and emo-tion: twitter sentiment and socio-economic phenomena[J]. Computer science,2009,44(12) : 2365 - 2370.

[11] Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter.PloS one,2011,6(12), e26752.

[12] HU M,LIU B. Mining and summarizing customer reviews[C]/ /Tenth ACM SIGKDD international conference on knowledge discovery and data mining. Seattle: ACM,2004:168-177.

[13] FELDMAN R. Techniques and applications for sentiment analysis[M]. New York: ACM,2013.

[14] VOLKOVA S,WILSON T,YAROWSKY D. Exploring demo-graphic language variations to improve multilingual sentiment analysis in social media[C]/ /Proceedings of conference on empirical methods in natural language processing.Seattle: ACL,2013: 1815-1827.

[15] Pang,B. ,Lee, L.,&Vaithyanathan, S.Thumbs up? sentiment classification using machine learning techniques. arXiv.2002.

[16] Cui Yan, BAO Zhiqiang. Overview of Association rule Mining [J]. Computer Application Research,2016,33(02):330-334.

[17] Peng Xixian, Zhao Yuxiang, Zhu Qinghua.Research topic of iConferences based on Association rule Mining [J]. Information Journal,2013,32(12):1303-1314.

[18] Zhang Yue, Ni Junmin, Wang Jian, Song Xiaokang, Zhao Yuxiang.Topic analysis of health informatics based on association rule mining: A case study of d Health conference [J]. Journal of Information Resource Management, 2020, 10(06): 90-100.

[19] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [J].Advances in neural information processing systems, 2013(26):3111-3119.

[20] Zhang Yue, Sun Xiaoling, Zhu Qinghua. Research on characteristics and rules of Public opinion Communication in Public Emergencies: A case study of Sina Weibo and Sina News Platform [J]. Information journal, 2014,33(4):90-95.