[1]Dengyun Zhu
[2]Hailong Gai
[3]Hongzhi Yu
[4]Rong Jing
[5,*]Fucheng Wan

Media Buzzword Analysis Integrated with Phrase Vectors and Topic Model

Abstract: - Buzzword analysis is one of the important research contents of natural language processing, and the research results can provide technical support for public opinion analysis. The purpose of extracting media buzzwords is to analyze the rules and changes of language change within a range. The traditional word feature-based buzzword extraction had some problems, such as low accuracy and low coverage, and this paper proposes a media buzzword analysis based on the combination of phrase vector and topic model, the core idea is to integrate the semantic similarity features, and use visualization technology to more intuitively show the overall language change rules. Visualization analyses uses a large number of corpus statistics, calculate the distance between words, and then convert into similarity, through word similarity calculation to show the distribution relationship between different words, and finally quantitative perspective to analyze. Our model is better than the traditional system, and the research results can provide corpus and model support for subsequent research directions.

Keywords: Buzzword Analysis, Phrase Vector, Topic Model, Semantic Similarity, Visual Analysis.

I. INTRODUCTION

In essence, Internet buzzwords are a special new word, automatic recognition of these words is the basis for further processing and analysis [1], the extraction of buzzwords is based on the use of buzzwords in the short term rapid increase and decline of this characteristic, through the analysis of a large amount of data of real Internet forums to characterize the use of words in the cross-year time period, so as to quantify and measure the popularity of words [2]. According to the law that the usage distribution of Internet buzzwords increases rapidly in a short period of time, the frequency of use of candidate words at different time nodes is counted, and a probability model is established for candidate words by year, and the popularity of candidate network terms is measured and the word popularity score is measured by calculating the KL distance between models in adjacent time periods, and the Internet buzzwords are automatically obtained through sorting [3].

With the development and popularization of computer networks, the rise of rich and colorful network culture, popular words and network languages came into being, and they spread and popularized rapidly in the network world. Buzzwords and internet slang are an important part of new words [4]. In September 2006, the National Language and Writing Committee issued and published the “Report on the Life of Chinese and Chinese Dialects (2006)”, which has been published annually since then, providing strong corpus support and important reference materials for linguistic research. In short, during this period, the research of new words was combined with information technology means, quantitative and qualitative, and innovation and breakthroughs were achieved in research methods and technical means. Di Zihuan et al. [5] analyzed the dissemination of language information from the perspective of information ecology. Xie Xiaoming et al. [6] started from the economic attributes of language to discuss the performance of network language as a factor of production and the development of language industry in cyberspace. Abbe [7] expounded the definition of online language, introduced the characteristics of online language, and analyzed the positive and negative impacts of online language on teaching Chinese as a foreign language, hoping to provide reference significance for teaching Chinese as a foreign language. Liu Yi [8] accurately separates abbreviations from Internet buzzwords through the concept definition of abbreviations, discusses the abbreviation methods used by various Internet popular abbreviations, presents the specific process of

[1] Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730030, China; Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu, China
[2] Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730030, China
[3] Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu, China
[4] Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730030, China
[5] Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, Northwest Minzu University, Lanzhou, Gansu, China
*Corresponding author: Fucheng Wan

abbreviation formation, and finally briefly explains the significance of the clarity of acronym structure cognition for regulating the network environment. Yang Yixi [9] starts from the classification of network language, analyzes its diachronic characteristics according to its functional level, and discusses the causes of popular communication of network language from the dimensions of culture, technology, politics, user subject, and dominant force. Cheng Runfeng [10] believes that the socialization of network language is a complex situation formed by the comprehensive action of multiple mechanisms. Examining this situation from different perspectives, the formation mechanism of different dimensions can be found. Du Zhenshuo et al. [11] analyzed the propagation mechanism of online language, focused on a series of problems arising in the process of network language transmission, and explored ways to promote the benign development of network language, in order to improve the ability to use network language. Sun Liling [12] adopts the attribute correlation analysis method to consider the correlation between online buzzwords and social hot issues and mainstream value orientation, and analyzes the relationship between the emergence and establishment of online languages and the transformation of economic system, cultural transformation and social structural adjustment, and concludes that the production of online buzzwords meets the needs of the current development of the times, and value selection plays an inward-looking function in the development of online languages. Li Yanhong [13] From the perspective of collective behavior, it can be seen that the formation of the network public opinion field is promoted by a variety of factors, showing the group power of cyberspace and the collective behavior in discourse production, and discourse power is the focus of the online public opinion field. The above is all from a qualitative point of view on the network.

Compared with the previous research work, the contribution of this paper is to use the method of combining quantitative and qualitative data to analyze Internet buzzwords, including data display of different vocabulary and visual analysis.

II. BUZZWORD CHARACTERISTICS

Media buzzwords reflect hot events and public concerns in social life over a period of time [14]. These words are not only reported in the news, newspapers and radio for a long time and with great frequency, but they are also closely followed. The top ten media phrases in 2020 include new crown pneumonia, anti-epidemic, civil code, double circulation and other words, reflecting major social hot events and new policies, and also summarize the changes in social psychology and social development in a certain period of time in highly condensed terms [15]. At the same time, it is inevitable that some abbreviations, abbreviations and unregistered words will appear, such as double carbon, double reduction, etc. The words “big white” and “ceiling” in the top ten popular words in 2022 have given new meanings to the old ones. Vigorous efforts, courageous march forward, and Chinese-style modernization are widely displayed in the public eye, and are important information and indicative slogans extracted from the theme and report of the congresss [16].

In addition to the characteristics of widespread dissemination and popularity in a certain period of time, media buzzwords also have the characteristics of clear introduction, prominent highlights of events, and often can accurately summarize the current social, cultural or political situation, as well as people’s emotional experience, so as to attract widespread attention and discussion, and are greatly influenced by current affairs and politics [17]. Considering the time and frequency of use of the buzzword itself, the time, month and phrase cumulative frequency of use are regarded as one of the characteristics of the catchphrase. The news corpus contains a large number of official political terms, which appear frequently and should not be recognized as buzzwords. Once custom stop words are partially removed, the impact of such words is reduced through unsupervised clustering and topic extraction. Some colloquial label words are not easy to identify, and noun phrases are used to dig out, such as the second middle school and the third middle school, the study of Chongde and the practice of history, and the core values of socialism.

Internet buzzwords present the characteristics of simplification and regularization of grammar, popularization of rhetorical devices [18], diversified semantics, and complex causes. Based on the characteristics of Internet buzzwords, this paper uses Weibo corpus as the basic data to extract Internet buzzwords [19].

Mutual information refers to the correlation between two collections of events and is used to measure the information shared by two random variables.

Information extraction is a text processing technology that extracts specified types of entity, relationship, event and other factual information from natural language text and forms structured data output. As the information extraction technology matures, it gradually moves from closed to open, and tries its best to break through the domain restrictions of the corpus. By making use of a large number of open-source corpus provided by the Internet, it makes continuous improvement, so that it can also show good results on large-scale corpus and achieve higher

robustness. However, the richness of language will also lead to great errors in the superficial understanding of many contents. At the same time, information extraction depends on the knowledge base, which can indeed improve the accuracy of information extraction and requires high reliability of knowledge base.

III. BUZZWORD EXTRACTION SCHEME

In 2003, Zhang Pu introduced the scientific definition and characteristics of buzzwords in the DCC-based research on dynamic tracking and assisted discovery of buzzwords, gave the formal characteristics of the buzzword curve, and proposed the possibility of computer-aided discovery of buzzwords. In view of the problems of diverse forms of Internet buzzwords and difficulty in obtaining unlogged words from conventional word segmentation, Wu Baozhen proposed a method to obtain Internet buzzwords based on total segmentation, using the total segmentation algorithm to obtain all possible word sets, and then using vector space models and language filtering rules to obtain candidate word sets. However, the segmentation results of this method show geometric growth, and the longer the sentence, the more results, and the efficiency of the word segmentation system decreases sharply [20]. Finally, according to the frequency of use, duration and degree of change, the buzzword scoring formula is proposed to obtain the buzzwords, and the experimental results have a 90% overlap rate with the annual buzzwords released by authoritative institutions. In 2015, Tang Yongli based on large-scale network corpus, using conditional random fields to segment network corpus, extracting entry information on the Internet encyclopedia platform and constructing a popular candidate vocabulary set through Chinese input method cell thesaurus export and other steps, according to the law of rapid improvement of popular words in a short time to establish a candidate word model. Calculate the KL distance between models in adjacent time periods to calculate the popularity score.

In terms of extraction effect and evaluation, the current common method is to compare the popular words extracted by the algorithm with the annual ten popular words released by the National Language Resources Monitoring and Research Center, and verify the algorithm extraction effect by analyzing the algorithm extraction results with the annual buzzwords coincidence rate released by authoritative institutions.

Corpus linguistics studies data engineering problems from the perspective of corpus linguistics, which can perform grammatical and syntactic analysis of natural languages and its relationship with other languages. Corpus linguistics first started to develop abroad, and the development is relatively mature and advanced. In comparison, I Chinese the beginning and development of library linguistics later. Today, corpus linguistics has made great progress in theory and methodology, and in recent years, corpus linguistics has shown interdisciplinary characteristics and has been widely accepted by the linguistic community. Keeping abreast of the latest developments in corpus linguistics can fill the research gaps in this field, while capturing research hotspots and conducting in-depth research. As an important research tool in the field of linguistics, corpus profoundly promotes the development of linguistics, especially applied linguistics, specifically studying data capture, data cleaning and data storage.

There are two main aspects of data sources: online media data and microblog data. The time horizons are all 13 years from 2011 to 2023.

Online media data comes from representative online media in ethnic areas, such as local news networks, digital daily newspapers, radio and television stations, and media networks. The content crawled includes the release date, title, and content of each news, and the release date is convenient for data storage and traceability; The headline can summarize the content of the news and facilitate text classification, clustering and other work; The content is the focus of the analysis and mining of this article, which is expected to be 4 billion words.

Weibo comes from 5 autonomous regions and 30 autonomous prefectures, due to the difference in the number of people in these regions and the number of people using Weibo, the current capture of each municipality of microblog data is up to 50,000, the lowest is 13,000, in response to this situation, for areas with less data volume will continue to expand users, each region’s microblog data expanded to more than 30,000, is expected to be 1 billion words. During the data capture process, the content of Weibo and the time information of Weibo release are saved at the same time. Data cleaning refers to the process of re-examining and verifying data, discovering and correcting identifiable errors in data files, and washing out erroneous or conflicting data according to certain rules, including checking data consistency, handling invalid and missing values, etc. Data cleaning is carried out to solve data quality problems on the one hand, and to make data more suitable for mining, display and analysis on the other hand.

There is noise in the captured online media and microblog data, which is mainly divided into the following two aspects: 1) information that needs to be deleted: such as special symbols, meaningless HTML element tags, super phone names, user names, topic names, etc.; 2) Data information that needs to be standardized: such as time format,

date, repeat information, etc. Data cleaning is carried out on the above two aspects to improve the quality of the corpus.

The purpose of data storage is to save the collected data for subsequent viewing and recall. The data is saved locally in plain text TXT format, which is convenient for overall data management and migration. Store different network media and microblog data separately, and save them according to data type, year, month creation hierarchy and folder. On the basis of the basic corpus, a database is constructed to store the multi-dimensional attribute characteristics of the corpus.

IV. BUZZWORD EXTRACTION SCHEME

A. Technical route

The traditional new word discovery, hot word extraction and buzzword analysis technology rely on the words themselves, by considering the mutual information of words, TF-IDF and other characteristics of the analysis, in order to analyze the characteristics of words in more detail, consider the current mainstream word vector technology in academia to model words, combined with deep learning technology to study new words, hot words and buzzwords.

In linguistics, word embeddings are discussed in the field of research of distributed semantics, which aim to quantify and classify semantic similarities between linguistic items based on the properties of distributions in large samples of linguistic data. The technique of representing words as vectors originated in the 60s of the 20th century with the development of vector space models for information retrieval. The use of singular value decomposition to reduce dimensionality then led to the introduction of latent semantic analysis in the late 80s of the 20th century. In 2000, Bengio et al. provided a series of papers on "neural probabilistic language models" to reduce the high dimensionality of word representations in context by "learning distributed representations of words". There are two different styles of word embedding, one is a vector that represents a word as a common occurrence of the word, and the other is a vector that represents the word as the linguistic context in which the word appears; In 2013, the Google team led by Tomas Mikolov created word2vec, an embedded toolkit that trains vector space models faster than previous methods. Most neoword embedding techniques rely on neural network architectures rather than more traditional n-gram models and unsupervised learning.

In essence, Internet buzzwords are a special new word, automatic recognition of these words is the basis for further processing and analysis, the extraction of buzzwords is based on the use of buzzwords in the short term rapid increase and decline of this characteristic, through the analysis of a large amount of data of real Internet forums to characterize the use of words in the cross-year time period, so as to quantify and measure the popularity of words. According to the law that the usage distribution of Internet buzzwords increases rapidly in a short period of time, the frequency of use of candidate words at different time nodes is counted, and a probability model is established for candidate words by year, and the popularity of candidate network terms is measured and the word popularity score is measured by calculating the KL distance between models in adjacent time periods, and the Internet buzzwords are automatically obtained through sorting.

Internet buzzwords present the characteristics of simplification and regularization of grammar, popularization of rhetorical devices, diversified semantics, and complex causes.

Mutual information refers to the correlation between two collections of events and is used to measure the information shared by two random variables.

Left and right information entropy is calculated between the left and right entropy between a pair of words, the greater the entropy, the more it indicates that it is a new word. Because entropy represents uncertainty, the greater the entropy, the greater the uncertainty, that is, the richer the left and right combinations of the pair, and the more choices.

The overall process is shown in Fig. 1.

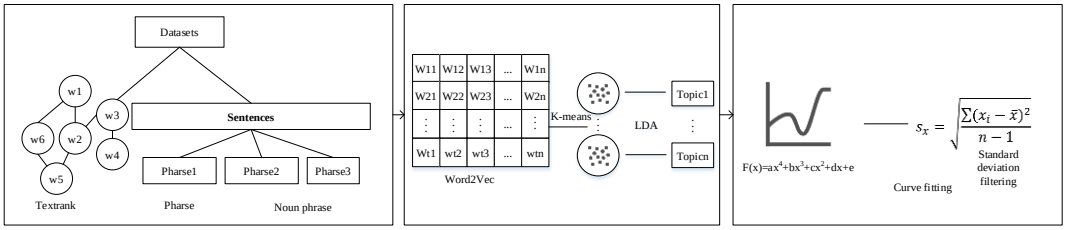


Fig. 1 The overall process.

B. Acquisition of noun phrases and key phrases

Construction of stop vocabulary: comprehensive Baidu, Harbin Institute of Technology, Sichuan University, Chinese stop word list to stop, the merged stop word list a total of 2340 words, the construction of official terms stop word list a total of 138 articles, the phrase containing the official stop word is removed, desensitization processing and some common words are removed, such as province, city, state, work, and other words. Text rank gets key phrases, stammering participles, and part-of-speech tags to get noun phrases. The process of obtaining nouns and phrases is shown in Fig. 2 as follows:

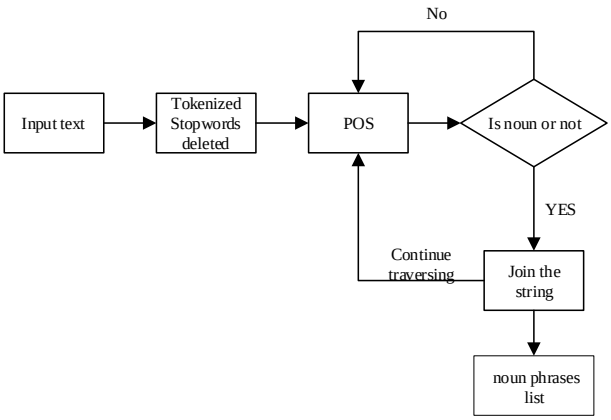


Fig. 2 noun phrase acquisition.

C. word2vec obtains phrase vectors, K-means clustering, and LDA extracts candidate phrases

To remove single words, Word2Vec obtains the phrase vector, uses K-means for phrase clustering, and some vocabulary clustering results are shown in Figure 3-6.

1) The phrase vector is the sum of the word vectors of the contained words, taking the mean

2) Put the trained phrase vector into K-means clustering, and cluster the phrases into different category clusters, so that the popular word extraction covers different categories of words as much as possible, and reduces the repeated extraction interference of the same category of words.

3) LDA was used to extract topics for clustered words, and the top 20% of the extracted subject words were used as candidate phrases

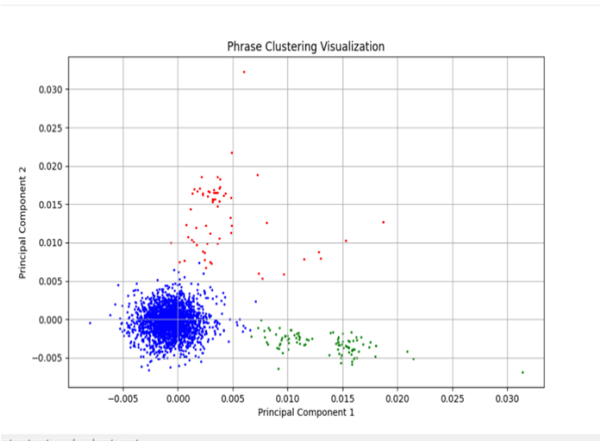


Fig. 3 Cluster Results

D. Quadruple polynomial curve fitting, standard deviation screening

The word frequency of candidate phrases was combined for 12 months, and the month-standardized cumulative string plot was used for quadruple polynomial curve fitting. Taking the month as the basic unit to count word frequency, remove low-frequency words, in several curve forms, according to the popular time attribute of the buzzword, it is considered that the curve with extreme value points has the characteristics of the buzzword, as shown in Figure 4, that is, it prevails for a while at a certain time, and then the frequency is low and stable, or stable for a period of time, and then the frequency increases. When the frequency of common words is stable, curve fitting also makes it difficult to identify common words. The standard deviation reflects the degree of dispersion of the data, so this section recalculates the monthly word frequency instead of using the cumulative word frequency to calculate the word frequency standard deviation. A larger standard deviation indicates that the monthly frequency

of use data for the vocabulary is more dispersed, the increase is larger, and the likelihood of being a buzzword is higher.

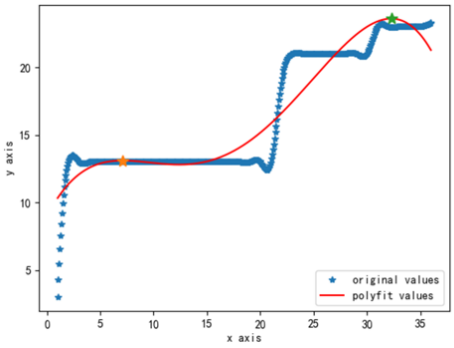
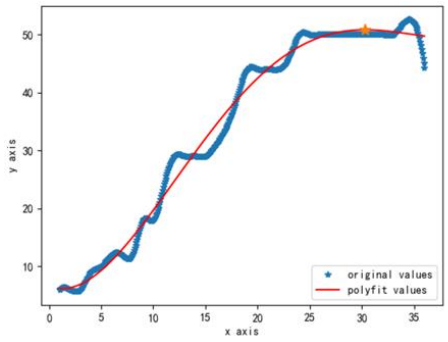


Fig. 4 Cross-infection Fig.



5 Long-term mechanism

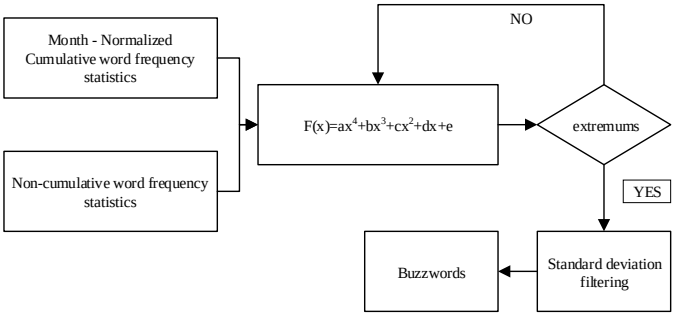


Fig. 6 Buzzword filtering

V. RESULTS

The extraction results of Internet buzzwords in ethnic areas from 2011 to 2023 are shown in Table. 1.

Table. 1 Extraction tool

annual	Systems	Basic ideas
OPENIE4	OpenIE system	Combining semantic role extraction SRL and nominal phrase RELNOUN
Stanford CoreNLP	A set of human language technology tools	Give the basic form of the word, the part of speech, and mark sentence structure with phrase and syntactic dependencies
Deepdive	An open source knowledge extraction system developed by Stanford University's Infolab Lab	Structured relational data is extracted from unstructured text through weakly supervised learning
ClausIE	OpenIE system	Use grammar knowledge (sentence recombination) to transform complex sentences and analyze sentence elements such as grammar and clauses
MinIE	Built on the ClausIE system	Focus on redundant items and whether they are trusted
ReVerb	For Web-scale information extraction	A program to automatically identify and extract binary relationships in English sentences

According to the 2020, 2021 and 2022 Chinese media ten popular words released by the Language Resources Monitoring and Research Center of Jia Jia Language Resources Monitoring and Research Center, statistics on the overlap between the corpus and popular words, it is found that many popular words do not appear frequently or have not appeared in the corpus, such as the frequency of the ten popular words in 2021 in the 12 months of the corpus in 2021 is shown in Figure 3-6 above.

VI. CONCLUSION

This paper uses the left-right information entropy method to calculate the semantic similarity of words, left and right information entropy is calculated between the left and right entropy between a pair of words, the greater the entropy, the more it indicates that it is a new word. Because entropy represents uncertainty, the greater the entropy, the greater the uncertainty, that is, the richer the left and right combinations of the pair, and the more choices. The next step in this article will be to combine big data models to extract and analyze buzzwords, the performance of buzzword extraction will develop to large model technology and general artificial intelligence technology in the future.

ACKNOWLEDGMENT

This work is supported by Gansu provincial university youth doctoral fund project (2022QB-016) and the Fundamental Research Funds for the Central Universities (NO. 31920240045).

REFERENCES

[1] Kishan S .US Watchdog Set to Clamp Down on Misleading Green Buzzwords. Environment & Energy Report,2022,(Jul.):1-2.

[2] Yan Zhigang, Li Chengcheng, Lin Min. Named entity recognition method based on knowledge graph information . Journal of Shanxi Normal University (Natural Science Edition),2021,35(01):51-58.

[3] Group O G I. OPINION: BUSINESS SCHOOL BUZZWORDS APPLIED TO MIDSTREAM. Oil and Gas Investor, 2023,43(03):77-77.

[4] Ma Dandan. Journal of Chongqing Second Normal University, 2018, 31(1):4.DOI:CNKI:SUN:XQJI.0.2018-01-013.

[5] Di Zihuan, Lü Mingchen. Research on network language information dissemination mechanism from the perspective of information ecology. Information Science, 2021, 39(1):7.

[6] XIE Xiaoming, CHENG Runfeng. Network language production factors and development of network language industry. Research in Institutional Economics, 2022(4):20.)

[7] Abbe. The Influence of Internet Language on Teaching Chinese as a Foreign Language, 2021,(19):342-343.DOI: 10.12217/j.issn.1009-5071.

[8] LIUY. Research on abbreviations in Internet buzzwords in the past five years. Modern Linguistics, 2023, 11(6):2516-2521.DOI:10.12677/ML.2023.116338.

[9] YANG Yixi, ZHOU Qiong. Classification and Genesis Analysis of Network Languages. 2021(2017-2):44-52.

[10] CHENG Runfeng. On the socialization of network language. Chinese Social Sciences Digest, 2022(8):2

[11] DU Zhenshuo, ZHOU Li. Research on the Propagation Mechanism of Network Language in the New Media Era[J]. Editorial Editor, 2021(10): 3.DOI: 10.3969/j.issn.1007-8177.2021.10.044.)

[12] SUN Liling. Research on the Relationship between the Development of Network Language and Social Value Choice. 2021(2014-6):32-36.

[13] Li Yanhong. The Formation of Internet Public Opinion Field from the Perspective of Collective Behavior: An Analysis Based on Internet Buzzwords. 2021(2014-11):41-43.

[14] Xie Teng, Yang Junan, Liu Hui. Chinese entity relation extraction using BERT. Application of Computer Systems,2021,30(05):253-261.

[15] Talati A. Insights into the beauty industry Historical trends, hero ingredients and buzz words. Euro Cosmetics, 2022(3):30.

[16] Melley L E, Sataloff R T. Beyond the Buzzwords: Artificial Intelligence in Laryngology. Journal of voice: official journal of the Voice Foundation, 2022,36(1):2-3.DOI: 10.1016/j.jvoice.2021.03.001.

[17] Lybarger Kevin,Ostendorf Mari,Thompson Matthew,Yetisgen Meliha. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. Journal of Biomedical Informatics,2021,117.

[18] Gao Yue. Overview of open domain information extraction. Modern Computers,2021(07):80-83+87.

[19] Fei Hao,Ren Yafeng,Zhang Yue,Ji Donghong,Liang Xiaohui. Enriching contextualized language model from knowledge graph for biomedical information extraction. Briefings in bioinformatics,2021,22(3).

[20] Liu C. The Interpretation of Chinese Youths Values Orientation in Post Epidemic Era Based on Buzzwords in Prototype Category Theory. English Literature and Language Review, 2021, 7.