COVID-19 Literature Mining and Analysis Research

Xie Luling[&]
School of Mathematics and Statistics
Northeastern University at Qinhuangdao
Qinhuangdao, China
x996706825@163.com
[&]These authors contributed equally to this work and should be considered co-first authors.

Wang Zixi[&]
School of Mathematics and Statistics
Northeastern University at Qinhuangdao
Qinhuangdao, China
1621901024@qq.com
[&]These authors contributed equally to this work and should be considered co-first authors.

Chen Yeqiu
School of Control Engineering
Northeastern University at Qinhuangdao
Qinhuangdao, China
805749576@qq.com

Wen Yichao
School of Mathematics and Statistics
Northeastern University at Qinhuangdao
Qinhuangdao, China
709844141@qq.com

Zhao Di
School of Mathematics and Statistics
Northeastern University at Qinhuangdao
Qinhuangdao, China
443345680@qq.com

2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA) | 978-1-6654-1606-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/EEBDA53927.2022.9744877

***Abstract*—By 2019 COVID-19, since the epidemic, the number of relevant documents exponentially level rise. Faced with a large amount of literature, this research provides convenience for exploring the connection between research topics and fields and quickly understanding relevant literature information. We pass on the data set after data cleansing using the LDA(Latent Dirichlet allocation) methods, and Berts and K-means modeling method extracting topic keywords. Use knowledge graph tools to output relevant visual graphics and systematically extract adequate information. Through text mining of biomedical research papers related to COVID-19, the improved model is used to analyze and make recommendations to respond to and prevent the COVID-19 pandemic. This research can support the rapid and in-depth analysis of a large number of relevant documents and can be used in future research to support real-time scientific disease research.**

Keywords—literature topic modeling, COVID-19, bibliometrics, machine learning

I. INTRODUCTION

Since the 2019 coronavirus disease epidemic, scientists have researched various treatment methods and transmission research topics. As the world pays attention to solving the COVID-19 disease, the number of relevant documents has also increased exponentially. In the face of such a brand-new disease, it was evident that it takes a lot of time and cost to conduct an in-depth analysis of a large number of related documents.

Therefore, it is very important to use scientific and technological methods to explore a large number of relevant documents.

The current literature analysis mainly uses bibliometric tools and R software to quantitatively evaluate and visualize knowledge in the field of COVID-19 research. The content of their study is also mainly focused on ranking and visualizing the number of related documents published in different countries and ranking sites (journals and preprint servers). At the same time, some studies use manual screening methods for topic recognition. However, these research themes only focus on the characteristics of the research publications and the pieces of the literature, which have substantial limitations. Part of the clustering algorithm is used to group published articles according to the similarity of abstracts to determine research hotspots and current research directions (Abd-Alrazaq et al., 2020).[4]Still, the LDA algorithm used in this study is more specific.

In this research, we conduct in-depth explorations of literature research on various topics to fill the gaps mentioned above. Using the methods in this study, we can conclude that COVID-19-related literature concentration areas and COVID-19 infection have a strong correlation. At the same time, the sales volume of searching related literature is increased. This research adds a literature search system, which can quickly select publications with more articles related to the epidemic. We can use these methods in future research to support real-time scientific disease research.

We summarize our contributions as follows:

-We select the required columns in the data set, such as title, publication time, abstract, etc., and perform preliminary document screening, abstract sorting, stop word processing, morphological restoration, and inverted index processing.

-For the selected data set, LDA is selected for topic modeling and T-SNE for visualization, and SentenceBert+K-means is used for clustering to draw word clouds. The performance through the LDA topic modeling method is even better.

-Based on the topic modeling results of LDA, we build a relevant knowledge map and conduct a map analysis of the topics covered by the literature.

-To better find a solution to COVID-19, we have

established a literature search and question-and-answer system. Through this system, you can quickly query and output related documents or results.

II. Related Work

This research mainly focuses on the COVID-19 literature data set. The current research methods on the literature data set are as follows:

1) Conducted a qualitative and quantitative analysis of publication outputs, journals, authors, institutions, countries, cited references, keywords, and terms according to bibliometric methods using VOS viewer c software packages.(Nasab et.al 2020)[5].

2) The data source was studied and indexed by bibliometric methodology. For visualized study, bibliographic coupling analysis, co-authorship analysis, co-citation analysis, co-occurrence analysis and the analysis of publication trends in coronavirus research were conducted by VOS (visualization of similarities) viewer and GraphPadPrism 6 software.(Mao Xingjia, 2020)[6]

The above articles all use software to analyze the literature, but also mainly statistics on the countries and topics that have published the most literature.

3) Use machine learning to process the research literature on SARS-CoV-2 together with SARS-CoV and MERS-CoV. (Jake Lever, 2021)[7]

4) Use a machine learning–based method to analyze the most relevant COVID-19–related articles and extracted the most prominent topics. Specifically, we used a clustering algorithm to group published articles based on the similarity of their abstracts to identify research hotspots and current research directions.

They use machine learning algorithms to analyze the literature, but the topics are still limited to the journals that publish the most related literature, the countries that publish the most related literature, and the main focus of the literature.(Abd-Alrazaq et.al,2021)[8]

We can notice that these articles basically analyze and visualize the journals that publish the most literature, the countries that publish the most related literature, and the main topics of the literature, which have relatively large limitations.

Therefore, if we want to make the COVID-19 literature data set play a greater role, we should conduct a more in-depth and accurate analysis of more aspects, such as the use of a more accurate algorithm, the LDA algorithm. At the same time, increase the analysis of the related factors of infection with COIVD-19 and increase the literature question and answer system to provide convenience for rapid literature analysis.

III. Method

In this section, we will introduce the details of our experimental process in detail.

We used Kaggle, one of the COVID-19 paper document data set, the data set includes 497798 Row and 19 columns. The data set contains a CSV (Comma Separated Value) file, which includes the metadata of all articles in the data set, such as article ID, title, abstract, and publication date. Dataset article represents a single JSON (JavaScript Object Notation) file, the article ID, source, title, abstract, text, author, journal, and URL(Uniform Resource Locator) composition.

A. Data Preprocessing

We use Python to analyze the metadata of the dataset in the Jupyter Notebook environment. First, read the data in the data set, select the required columns, such as title, publication time, abstract, etc., and then filter the publication year of the literature. We can see 2020 Total years 403,020 Article 25. After screening abstracts, remove the summary missing and duplicate posts, 120,254 papers don't have abstracts, 38,581 papers have repeat abstract of documents that need to be deleted. Then convert it to lowercase text. After that, Python of NLTK(Natural Language Toolkit) library useless words, such non-English, stop words, and only one letter of the word removal can save time and space. After that, we use the spacy model to perform lemmatization and convert the word to its root. Finally, we use the inverted index technique, assign a file for each word id list for quick retrieval keyword matching articles in the inverted index in a total of 225,549 keywords.

Then, article matching and deep cleaning are performed on the preliminarily processed data set. Firstly, we searched for words related to risk and found a total of 4,040 related papers. For topic modeling and visualization, lemmatization is required. Part-of-speech tagging technology is also used to filter meaningful words. Then we drew a word cloud for it, as shown in figure 1.

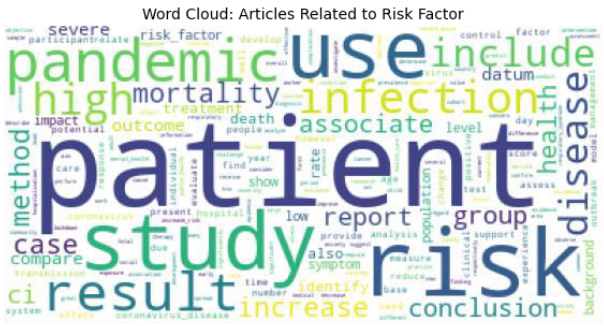

Fig. 1 Word Cloud:Articles Related to Rish Factor

It can be seen that words or phrases like "risk", "outcome", "patient", "infection", "disease", and "research" often appear in this corpus.

B. The Latent Dirichlet Allocation (LDA)

LDA is a supervised machine learning method used to separate two groups and classes. The main idea of linear discriminant analysis. LDA is to maximize the separability between two groups to make the best decision to classify them. LDA maximizes the separability between known categories by creating a new linear axis and projecting data points onto that axis.

The LDA topic modeling method is to treat each document as a certain proportion of the topic collection and automatically output the topic of the data set. Automatically extract a related subject (expressed by keywords) from a large number of documents. Determine the number of relevant issues, and then use LdaMulticore to build a model. Calculate the topic consistency through CoherenceModel to determine the number of topics. The higher the CV coherence score, the more appropriate the question number.

However, if we run iterations at different times, the coherence score will also be different.

C. Berts+K-means

Berts and its improved version are now performing well in various natural language processing tasks. SentenceBert is proposed to solve the huge time cost of Bert semantic similarity retrieval and its sentence representation is not suitable for unsupervised tasks such as clustering, sentence similarity calculation, etc. It can represent documents well, and the topic is more specific. However, the decision boundary of document clustering based on this is not clear.

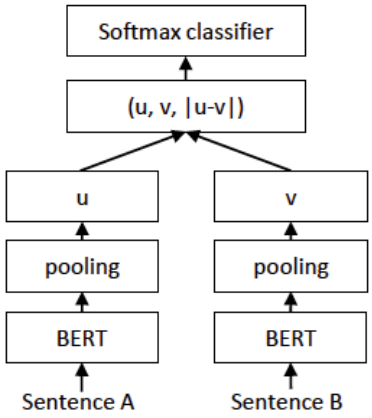

Fig. 2 Flow chart of SBERT

Sentence-BERT(SBERT), modify the pre-trained BERT: use a three-level network structure to obtain semantically meaningful sentence embedding can generate fixed-length sentence embedding, use cosine similarity or Manhattan Euclidean distance, etc. Compare and find sentences with similar semantics as shown in figure 2.

D. T-distributed Stochastic Neighbor Embedding(T-SNE)

T-SNE is an unsupervised machine learning method used to visualize high-dimensional data in low dimensions. T-SNE, for design implementation, can reduce any number to two feature space dimensional feature space. Both PCA and LDA are used for visualization and dimensionality reduction, but T-SNE is used exclusively for visualization purposes. It is very suitable for the visualization of high-dimensional data sets.

With the T-SNE dimensionality reduction visual method, the COIVD-19 data set within the data into data to be visualized as a two-dimensional. The respective distributed fashion classes and the existence of cross therebetween.

E. Knowledge Graph

Knowledge graph (KG) embedding is to embed components of a KG including entities and relations into continuous vector spaces, so as to simplify the manipulation while preserving the inherent structure of the KG. It can benefit a variety of downstream tasks such as KG completion and relation extraction, and hence has quickly gained massive attention.[1]This study uses knowledge graphs to visualize the results of LDA topic modeling, taking the topic and article title as two types of entities, author and abstract as entity attributes, and the related attributes as belong to. It is concluded that the factors related to COVID-19 infection and COVID-19 standard models

F. TF-IDF principle

The rapid growth of new literature on COVID-19 makes it difficult for the medical research community to keep up with recent updates. The previous work has already obtained some practical conclusions, but this is not enough.

To help scientific research and the government better solve COVID-19, we have built a literature search and question-and-answer system to help obtain information more quickly and obtain desired conclusions based on needs.

For document clustering, we first convert each document into a feature vector, where the feature is defined by the term frequency-inverse document frequency (TF-IDF) weight. TF-IDF represents the importance of words relative to documents in the corpus. This importance is proportional to the number of times the word appears in the document, but it will be offset by the frequency of the expression in the corpus. This ensures that the similarity measurement between documents based on TF-IDF is mainly affected by discriminants with relatively low frequencies in the canon. For the TF-IDF representation of the abstract, we used the TfidfVectorizer module of the Python scikit library.

Word Frenquency:

$$TF = \frac{number\ of\ word}{category\ word} \tag{1}$$

Reverse file frequence:

$$IDF = \log\left(\frac{number\ of\ corpus}{number\ of\ term\ word + 1}\right) \tag{2}$$

IV. Results

A. LDA topic modeling

Automatically extract a related topic (expressed by keywords) from the data set. Determine the number of suitable subjects, and then use LdaMulticore to build a model. Calculate the topic consistency through CoherenceModel to determine the number of issues. The higher the CV coherence score, the more appropriate the question number. However, if we run iterations at different times, the coherence score will also be different. After many experiments, it can be seen that for the processed data set, topic selection 6 is appropriate as shown in figure 3.

Set num_topics = 6, calculate the model perplexity to measure the model, the score is -8.43, the coherence score is 0.465, indicating that the model effect is not bad.

As shown in figure 4 and figure 5, the first topic is about treatment. You can see words like ``vaccine", ``drug", ``treatment", and ``procedure".

The second topic talks about people's mental health during the epidemic. You can see words like ``mental health" and ``pandemic".

The third topic talks about the severity of the disease. You can see words like ``severe", ``patient" and ``infected".

The fourth topic talks about the spread of the virus. We can see words like ``transmission",``country" and ``population".

The fifth topic talks about public services and work during the epidemic. You can see words like ``care", ``service", ``support", and ``work".

The sixth topic is about death from the epidemic. You can see words like ``death" and ``conclusion".

And the subject matter corresponding to each keyword relating to the use of pyLDAvis visual figure 6-9. We can see that these six themes have almost no overlap, and the themes are better obtained.

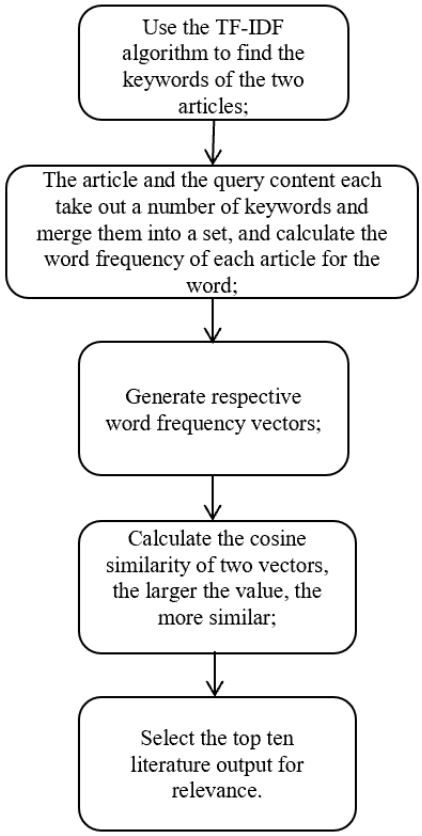

Fig. 3 Flow chart of LDA topic modeling

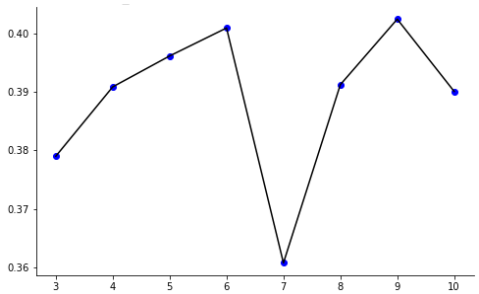

Fig. 4 Score for different number of topics

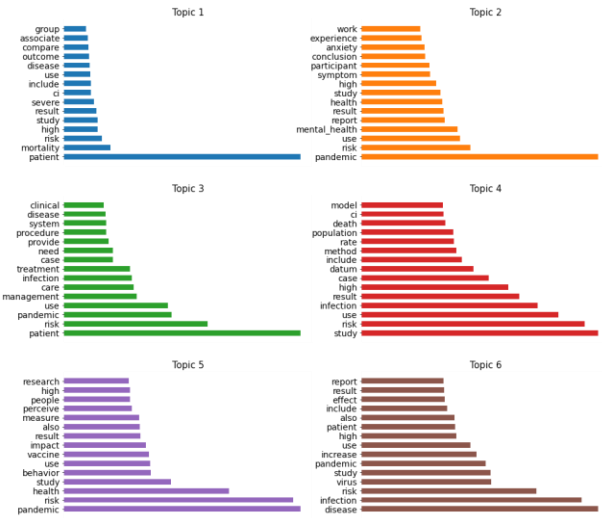

Fig. 5 Different topics and the words it contains

Fig. 6 Intertopic Distance Map

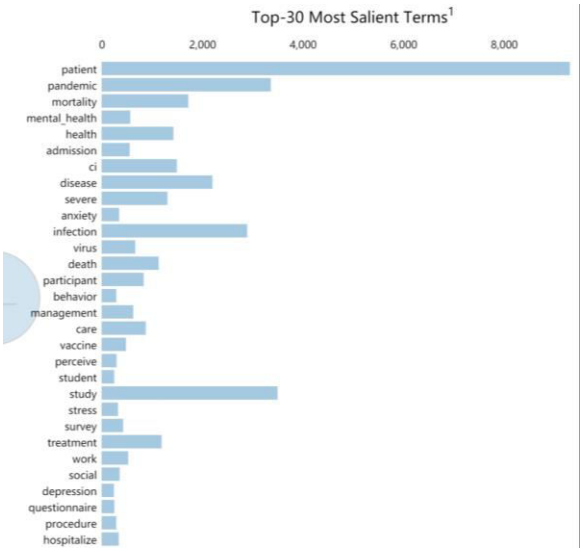

Fig. 7 Top-30 Most Salient Terms

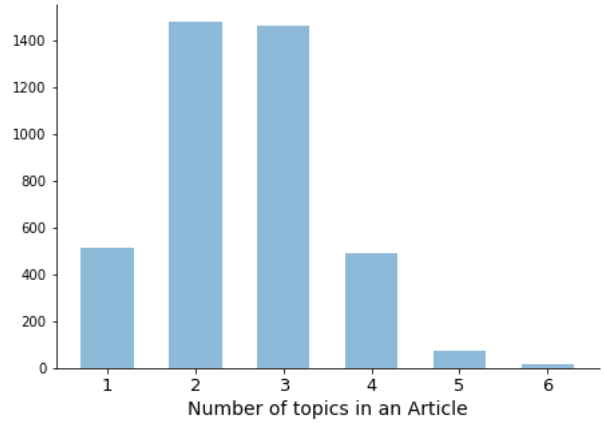

Fig. 8 Number of topics in an Article

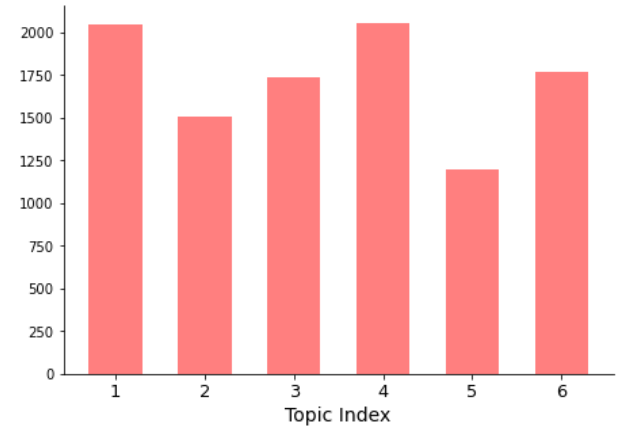

Fig. 9 Different topics and scores

Next, visualize the number of topics in the paper and the number of articles discussing each topic.

B. Modeling by Berts + K-means method

Get embedding sentence, a total of seven 68 Tiao, using the K-Means clustering, and draw word cloud.

Fig. 10 wordclouds by Berts + K-means method

This principle is used to compare the similarity between different topics. The first topic is similar to the subject generated by LDA and is about the spread of the virus. The second topic is similar to topic 6 generated by LDA and talks about death. The third topic is similar to topic 3 generated by LDA, but it seems more specific here, discussing a severe symptom: breathing. The fourth topic is similar to topic 5 generated by LDA, but it seems more specific here, consulting public services: health care services. The fifth topic is similar to topic 2 generated by LDA, but it seems more specific here, talking about negative emotions such as anxiety and disgust. The sixth topic may be similar to topic 1 generated by LDA and is about treatment methods.

First, use the T-SNE dimensionality reduction visualization method to transform the data into two dimensions to visualize the data and obtain the distribution of each class and whether there is an intersection between them.

The topic determines the category of each article with the highest probability. As can be seen from the figure 10, there is not much overlap between the documents of different clusters (topics), and the distribution is reasonable.

We can see from the figure that K-Means clustering overlaps the documents of different clusters (topics) in this task.

C. Comparison of two topic modeling methods

Get embedding sentence, a total of seven 68 Tiao, using the K-Means clustering, and draw word cloud.

Through these two methods, We can see that the topics extracted by the two methods are similar, but the issues extracted by the second method are more specific. LDA is the first choice for the thematic analysis of this task. You can use this algorithm to remove topics and cluster documents based on issues. LDA clustering can get a more explicit boundary and less topic overlap.

As shown in figure 11 and figure 12, Based on the topic modeling results of LDA, the category of each article is determined by the topic with the highest probability. Visualization of six treatment themes, transmission, support, severe, mental, death, and corresponding literature atlas. The subject and the article title are two types of entities, the author and abstract are the entity attributes, and the relationship attribute is belong_to.

Set weight, smoke, age, gender these four keywords, the summary extracted keywords and compare these four words, and this will be mentioned in the literature classify these factors. The document title and factors are used as entities to establish the relationship between the two types of entities, and the attribute is related to.

D. Knowledge Graph

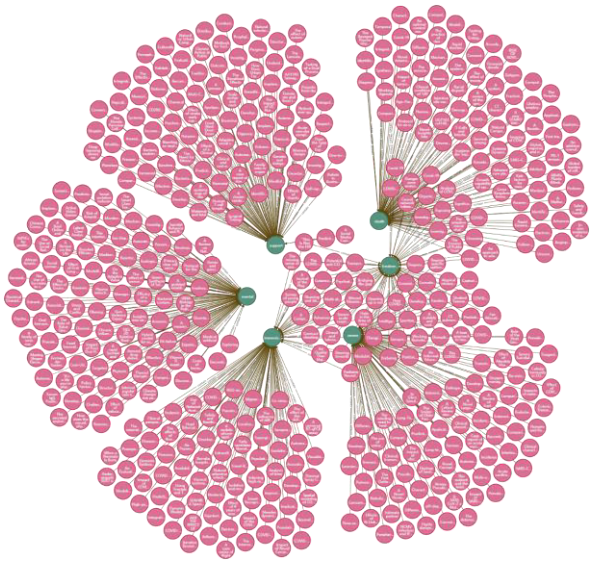

Fig. 11 Knowledge Graph 1

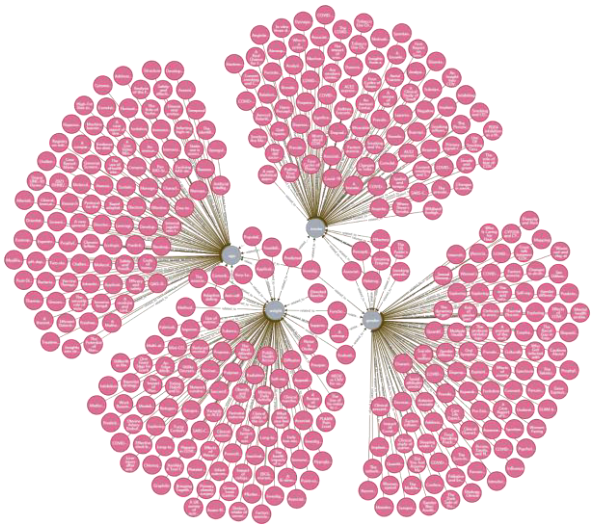

Fig. 12 Knowledge Gragh 2

We can see from the graph that the study of smoking is more related to gender among the four factors as shown in figure 13-16.

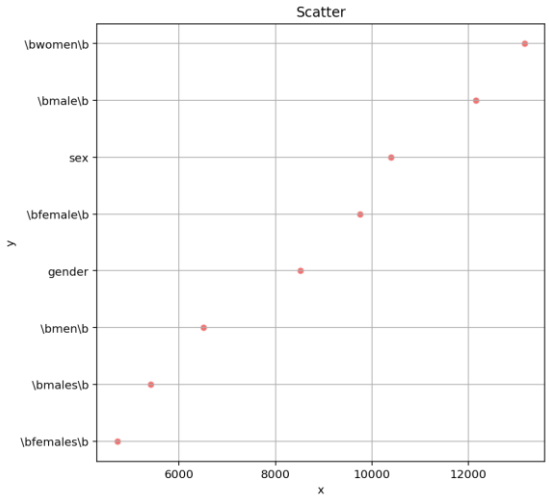

Fig. 13 Keyword: gender

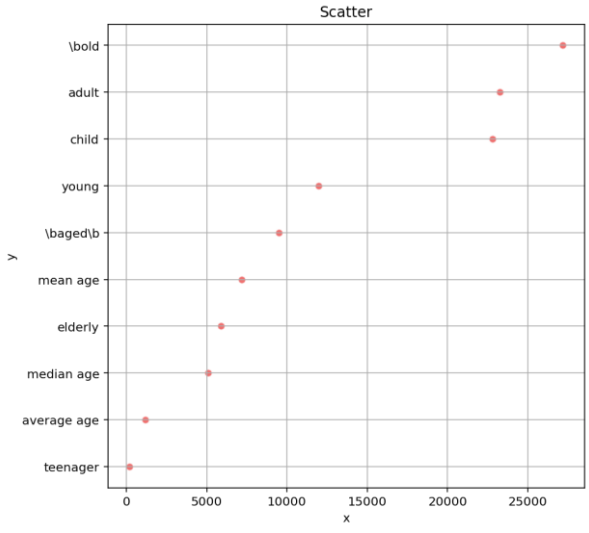

Fig. 14 Keyword: Age

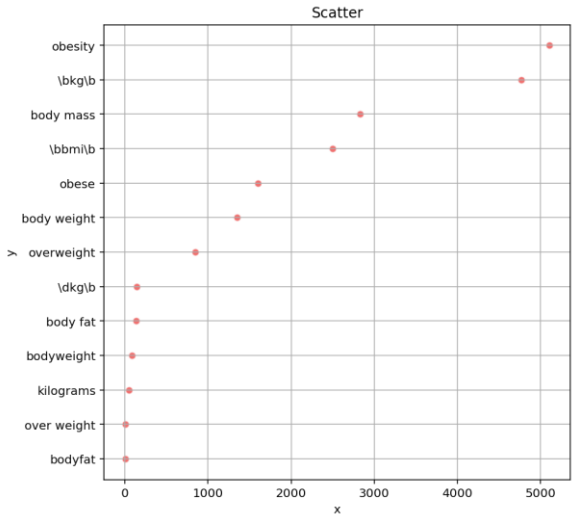

Fig. 15 Keyword: Weight

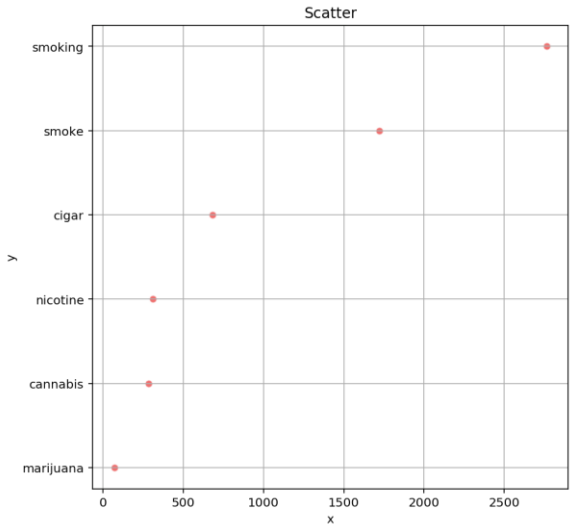

Fig. 16 Keyword: smoke

Use cipher of the count statement calculated for each factor off the number of documents linked by seeing a majority of the literature referred to the age factor, age and COVID-19 infection have a strong correlation.

Use cipher of factors calculated for each count statement off the number of documents associated with, seen from the results of the majority of the literature mentioned in weight due to factors, importance, and COVID-19 infection have a strong correlation. The fatality rate of obese people is higher than that of others. Obese people are more likely to develop severe illness after being infected with COVID-19.

There are few studies on the relationship between smoking factors and COVID-19. On the other hand, it reflects that smoking has a particular weak correlation with infection with COVID-19, and We can take specific intervention measures in this regard.

Epidemic models can predict how COVID-19 will develop to show the possible outcomes of the epidemic and help provide information for public health interventions As shown in figure 17.

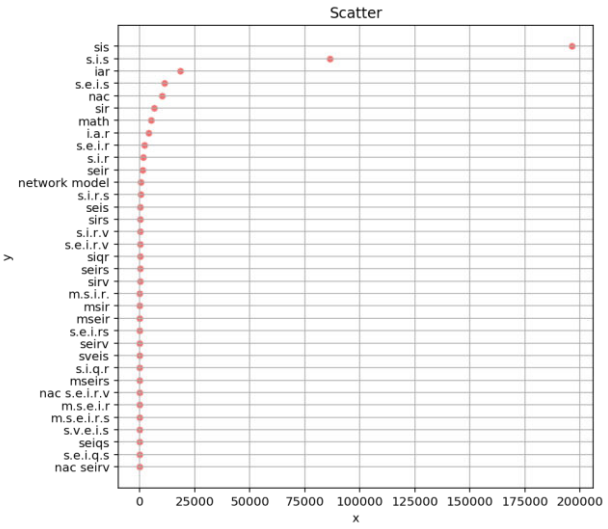

Fig. 17 Keyword: model

From the literature data, we can obtain that the mainstream epidemic model of COVID-19 is SIS, which can help decide which interventions to try and which interventions or predict the future growth pattern.

SIS can be used as one of the main methods, mainly because its method is simple and can be significantly adjusted by adjusting the modeling method and parameters.

On the other hand, the SEIR model (which phase is susceptible, exposed, sick, and rehabilitation) has been widely used COVID-19. However, if the disease is prolonged, they may be infected again, so the R (recovery) stage should be used with caution. The SEIRV model was already well used during the outbreak in Wuhan.

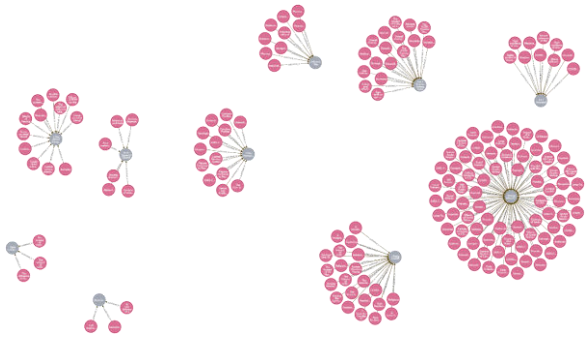

Fig. 18 Knowledge Graph of model

As shown in figure 18, The document and the published journal are used as entities to establish a corresponding relationship, and the attribute is published in.

We can see that the publications that publish more articles related to the epidemic, they can be given priority when looking for information.

Two thousand twenty-one years later, the data set contains 69514 articles. Among them, we excluded 35057 articles. Therefore, we have included in the analysis in this study 34457 article abstracts.

The en_core_sci_lg and TfidfVectorizer algorithms in the Spacy model transform the original text into a feature matrix of TF-IDF, which has 34457 rows 53756 columns for subsequent text similarity calculations.

Count the words appearing in the literature using the principle of TF-IDF, as shown in Figure 19.

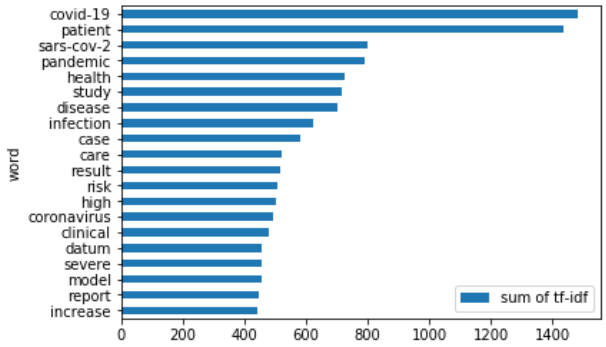

Fig. 19 Diffenert words and its sum of tf-idf

We obtained the definition of cosine similarity before the nth document. Define a method to get the first ten documents of the search query and define a strategy to answer any question. Define the way to display the results in the table. Our literature search system can enter the relevant content to be queried to find the most pertinent and output the first ten documents, enter the problem, and then from the robust correlation integration literature after the output of the relevant answers.

V. Discussion and Conclusion

From the results of this study, we can see that the relevant literature topics of COVID-19 are mainly six categories: treatment methods, people's mental health during the epidemic, the severity of the disease, the dynamics of virus transmission, public services, and work during the epidemic, and death.

We can see from the number of documents that the COVID-19 infection group strongly correlates with pregnancy, age, smoking, gender, and weight.

In infected COVID-19 in pregnant women, diabetes and high blood pressure is a common complication of the disease, there is the risk of preterm birth. Respiratory syndrome and pneumonia are common among newborns born to mothers with COVID-19. The SARS-CoV-2 infection spread among COVID-19 women.

There is a strong correlation between age and COVID-19 infection. Because of immunity and complications, the elderly are a vulnerable group in COVID-19.

Smoking is a risk factor that affects the progression of COVID-19. Smokers are more likely to develop COVID-19 progression than those who have never smoked. Doctors and public health professionals receivables related data on smoking as part of clinical management and smoking cessation added to the list COVID-19 to deal with the pandemic.

We can see that gender highly correlates with COVID-19 infection, and male patients are vulnerable in COVID-19. It may relate that the concentration of ACE2 in male plasma is higher than that in females.

The fatality rate of obese people is higher than that of others. Obese people are more likely to develop severe illness after being infected with COVID-19. The dysregulation of lipogenesis in obese patients and the

subsequent high expression of ACE2 may be the mechanism by which these patients are infected with SARS-CoV-2 and have an increased risk of serious complications.

The primary epidemic model used by COVID-19 is SIS. The main reason may be that the method is simple and can be significantly adjusted in modeling methods and parameters. On the other hand, the SEIR model (whose stages are susceptibility, exposure, illness, and recovery) has been widely used for COVID-19. However, if the disease is prolonged, they may be infected again, so we should use the R (recovery) stage with caution. The SEIRV model was already well used during the outbreak in Wuhan.

This study uses the LDA clustering algorithm and the Berts + K-means algorithm. We can see that the topics extracted by the two methods are similar, but the issues extracted by the second method are more specific. LDA is the first choice for the thematic analysis of this task. You can use this algorithm to remove topics and cluster documents based on issues. LDA clustering can get a more transparent boundary and less topic overlap.

We can use the literature search and Q&A system of this study to select journals that have published more articles related to the epidemic when submitting articles. Compared with conventional methods, the system integrates massive literature and can quickly search for relevant information, proving the feasibility of using artificial intelligence-based data mining in more fields. However, because we only query the keywords in the title summary to get the desired results. Given that only abstracts and titles are filtered, there may still be room for improvement in the accuracy of our results.

Application of this research in algorithms and literature inquiry and question and answer system, doctors in the treatment method, provides real-time scientific assistance aspects, how to psychiatrists and other related professionals epidemic of safeguard people's mental health recommendations for people to understand the disease The severity of the virus, the dynamics of the spread of the virus, public services, and jobs during the epidemic, and the number of deaths have provided quick methods. The methods involved in this study can not only be used in COVID-19 related treatments and provide related recommendations. At the same time, We can also use it in other fields where the research literature is growing exponentially.

REFERENCES

[1] Wang Q , Mao Z , Wang B , et al. Knowledge Graph Embedding: A Survey of Approaches and Applications[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(12):2724-2743.

[2] Blei D M , Ng A Y , Jordan M I . Latent Dirichlet Allocation[C]// Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]. 2001.

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

[4] Abd-Alrazaq A , Schneider J , Mifsud B , et al. A Comprehensive Overview of the COVID-19 Literature: Machine Learning–Based Bibliometric Analysis (Preprint)[J]. Journal of Medical Internet Research, 2020, 23(3).

[5] Nasab F R , Rahim F . Bibliometric Analysis of Global Scientific Research on SARSCoV-2 (COVID-19). 2020.

[6] Mao X , Guo L , Fu P , et al. The status and trends of coronavirus research: A global bibliometric and visualized analysis[J]. Medicine, 2020, 99.

[7] Lever J , Altman R B . Analyzing the vast coronavirus literature with CoronaCentral[J]. Proceedings of the National Academy of Sciences, 2021, 118(23):e2100766118.

[8] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need.” In Advances in neural information processing systems, pp. 5998–6008. 2017.

[9] Paulheim, Heiko. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods[J]. Semantic Web, 2017.

[10] Wang C , Ming G , He X , et al. Challenges in Chinese knowledge graph construction[C]// IEEE International Conference on Data Engineering Workshops. IEEE, 2015.

[11] Jing L P , Huang H K , Shi H B . Improved feature selection approach TFIDF in text mining[C]// 2002.

[12] Kuang Q , Xu X . Improvement and Application of TF-IDF Method Based on Text Classification[C]// International Conference on Internet Technology & Applications. IEEE, 2010.

[13] Wang H , Song M . Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming[J]. R Journal, 2011, 3(2):29.