Visual Analysis System for rumor propagation under epidemic

Shen haowen

July 25, 2023

Abstract

In this project I design a rumor visualization system for visualizing rumors in social networks during the epidemic. I conduct topic modeling of the rumor dataset and explore the changes in topics over time and region. Geographical rumor distribution is also performed and visualized in the form of China map. At last, I build a user network and analyse the popular rumor data. Visualization and reasonable explanation of rumors are also given along with the analysis.

1 Introduction

Since 2019, there has been a global outbreak of New Crown Pneumonia. In addition to bringing physical harm to the epidemic prevention and control, the spread of the epidemic has also triggered the spread of a large number of rumors, posing a huge challenge. This project aims to explore the rumor spreading phenomenon under the epidemic by analyzing the data of 366 epidemic rumors collected by Today's Headlines from November 7, 2021 to February 18, 2022.

1.1 Data description

The rumor data involved in this project came from Today's Headline, and a total of 366 rumors in chinese were collected from November 7, 2021 to February 18, 2022, containing information related to rumor release time, rumor source, main content of rumors, provinces involved in rumors, and users' browsing records of rumors. Here are the detailed features and descriptions of the dataset:

Features:
date - date of the rumor, recorded in str

source - the source of the rumor, recorded in str
content - text content of the rumor in chinese, recorded with str
province - the province in which the rumor is about, recorded in str, a null value means that the rumor does not involve geographical information
user_0 to user_17 - the number of rumors viewed by the user, recorded as int, 1 means that the user viewed the rumor, otherwise 0
like - number of likes

2 Data Preprocessing

Before performing a deep analysis of the dataset, first I conduct a basic data preprocessing. Considering that the dataset has no "id" column, I append the "id" list with range 0 to 365. Then I check whether this dataset has abnormal values. Except for the "source" and "province" column, all the data values are non-null. However, it is acceptable that some rumor records may not have specific source and province. So I did not conduct special process towards these null values in source and province. Also, there are no duplicate or infinite values in the dataset. For the number of likes, I checked the data of the top rumors with the highest number of likes. The first place is a total of ten social hot rumors included, with 2495 likes. Other likes data records are generally around or below 1000. As a result the first place rumor record can be considered as an outlier in the case of likes.

3 Topic Modeling

Topic Modeling is a process of detecting themes in a text corpus. The main idea behind this task is to produce a concise summary, highlighting the most common topics from a corpus of many texts. Rumor topic modeling is vital in addressing misinformation during the COVID-19 pandemic. By analyzing prevalent themes in rumor datasets, it provides valuable insights for combating false information effectively. Here I implemented Latent Dirichlet Allocation (LDA) and a deep learning method, BERTopic to conduct topic modeling and visualize the results.

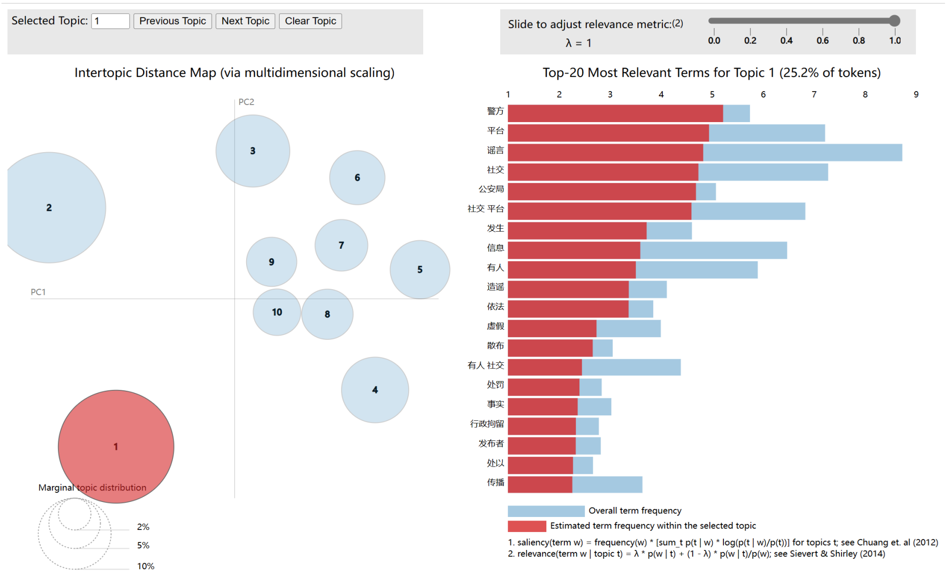
3.1 LDA

3.1.1 Brief Introduction

LDA is based on the idea that textual documents are made up of topics, which are made up of words from a lexicon. The hidden topics are a recurring pattern of co-occurring words. Every corpus that contains a collection of documents can be converted to a document word/document term matrix (DTM). LDA converts the documents into DTM, a statistical representation describing the frequency of terms that occur within a collection of documents. The DTM gets separated into two sub-matrices: the document-topic-matrix: which contains the possible topics, and the topic-term-matrix: which includes the words that the potential topics contain.

3.1.2 Training Details and Visualization

Considering that our rumor dataset contains chinese corpus, I used "jieba" module to tokenize the contents and remove stop words. Then I converted the text data into TF-IDF feature vectors using TfidfVectorizer. In the end I created the LDA model using LatentDirichletAllocation and used the TF-IDF feature vector as input. This approach applies LDA to the TF-IDF feature vector instead of using raw text data. This is because LDA is based on word frequency statistics for topic modeling. Therefore, TF-IDF feature vectors can provide better results because they take into account word frequency and word importance.



Selected Topic: 1

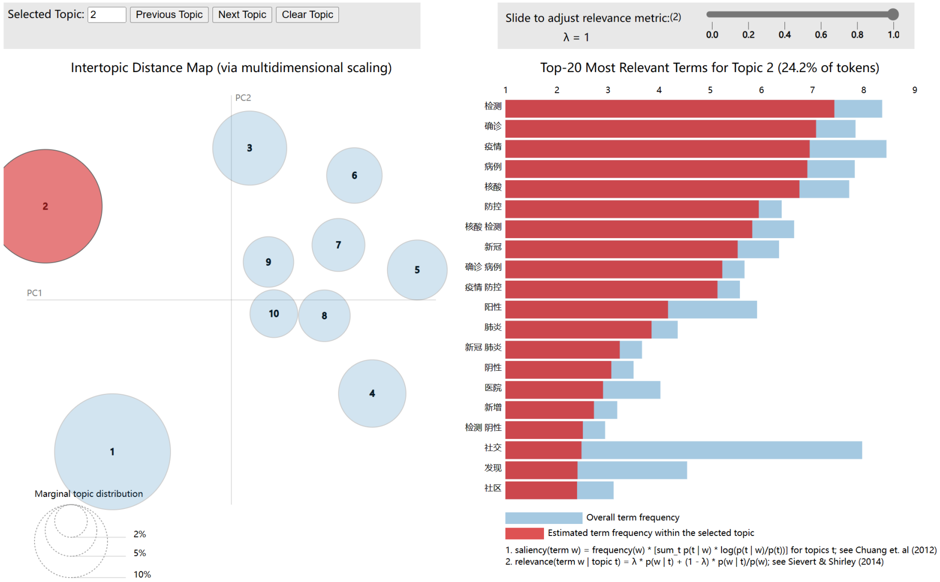
Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 1: Topic1 using LDA

In the above I visualize the overall topic distribution and top2 topic analysis from LDA. As we can see from the left part of 1, ten topics distributed independently from each other, indicating that LDA has a good discrimination for topic classification. Specifically for the first topic, it contains 25.2% proportion of all tokens. It can be seen from the right part of 1 that this topic mainly contains subjects like " 警方、平台、谣言、公安局、处罚..."

Also, the LDA analysis for topic2 is shown below. In this case we can clear see that subjects about epidemic appear frequently. Top most relevant terms contains " 检测、确诊、疫情、核酸、封控..." This topic contains 24.2% of tokens. As we can see from the LDA results, in general, our rumor dataset contains records mainly about "Public Security" and "Epidemic". There is good differentiation between different topics.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 2: Topic2 using LDA

3.2 BERTopic

3.2.1 Brief Introduction

BERTopic was developed in 2020 by Grootendorst (2020) [1] and is a combination of techniques that uses transformers and class TF-IDF to produce dense clusters that are easy to understand while maintaining significant words in the topic description. More specifically, first each document is converted to its embedding representation using a pre-trained language model(sentence_transformer by default). Then, before clustering these embeddings, the dimensionality of the resulting embeddings is reduced to optimize the clustering process. Lastly, from the clusters of documents, topic representations are extracted using a custom class-based variation of TF-IDF.

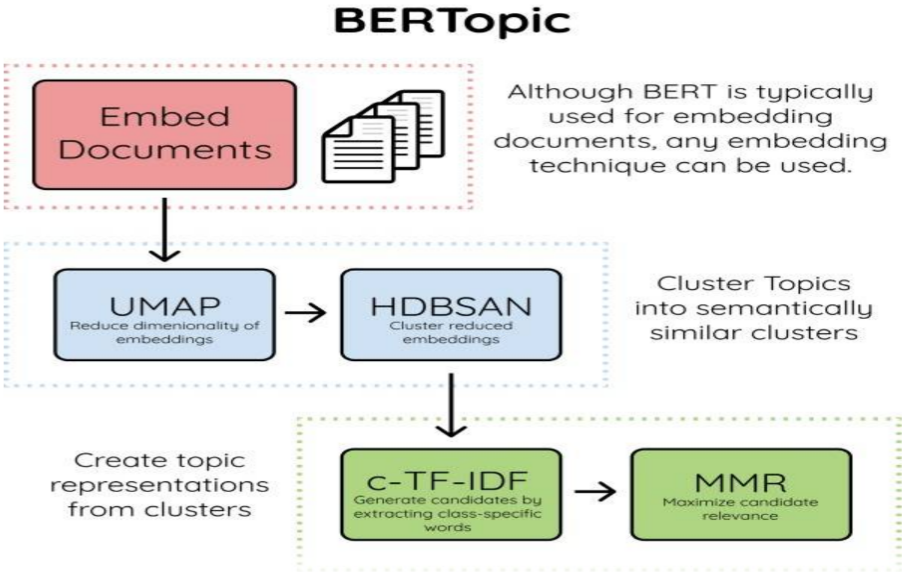

BERTopic
Embed Documents
Although BERT is typically used for embedding documents, any embedding technique can be used.
UMAP
Reduce dimenionality of embeddings
HDBSAN
Cluster reduced embeddings
Cluster Topics into semantically similar clusters
Create topic representations from clusters
c-TF-IDF
Generate candidates by extracting class-specific words
MMR
Maximize candidate relevance

Figure 3: Procedures of BERTopic

So why choose BERTopic? In LDA, each document is represented using a bag-of-words model (BOW). Based on this approach, topic models are adequate for learning hidden themes but do not account for a document’ s deeper semantic understanding. The semantic representation of a word can be an essential element in this procedure. In contrast, BERTopic can take into consideration the semantic understanding of the text with the use of an embedding model, and generate meaningful topics of which the content is semantically correlated. Also, it is possible for users to choose from a wide variety of embedding models, giving the model more flexible representation capabilities.

3.2.2 Training Details and Visualization

Here I conduct some topic modeling and visualization based on BERTopic. Considering that we are conducting topic modeling on a Chinese news corpus, I implement the sgns.sogou which is trianed on the Sogou News corpus [2] as the Chinese Word Embedding. As we can see in figure 4, BERTopic also performs good differentiation of different topics. In figure 5, we downscale the high-dimensional word embedding to 2 dimensions and conduct visualization. The documents in rumor dataset are mainly downscaled to three categories, in which "疫情、社交、谣言" occupies the biggest proportion. This is also in line with the theme of our project. The second category contains information about " 信息、人员、内容" which refers to the non-epidemic rumor records. The third category accounts for a small percentage, mainly belonging to some less related information.

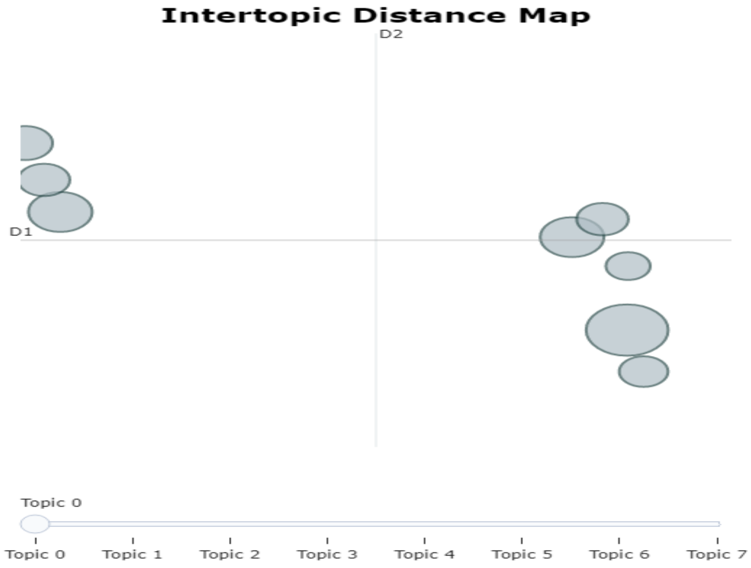


Figure 4: Intertopic Distance Map of BERTopic

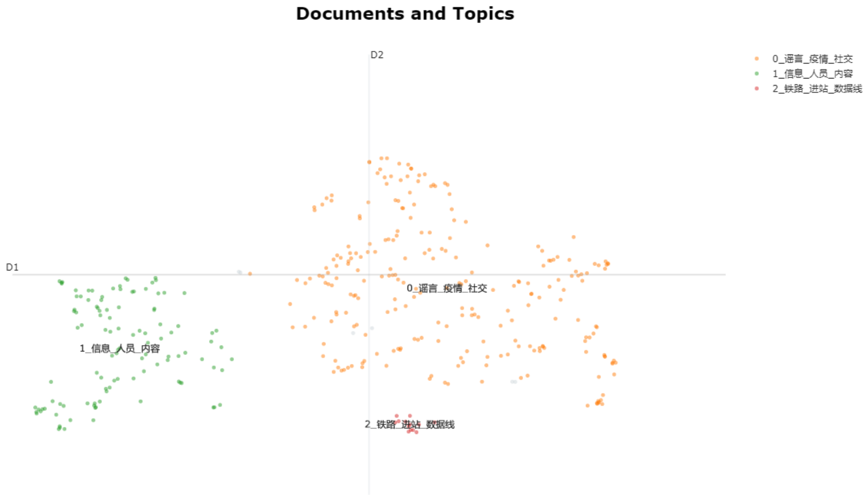


Figure 5: Document Downscaling and Embedding

As we can see in 6, BERTopic distributes the documents into eight topics. Compared to LDA, BERTopic gives a more specific and clear classification, while the differences between categories are also greater. Concrete and special topics like " 车辆、驾驶", " 学校、教育局", " 北京、冬奥会", " 坚果、食物" also appear in BERTopic, which give us a more comprehensive understanding of the rumor data.

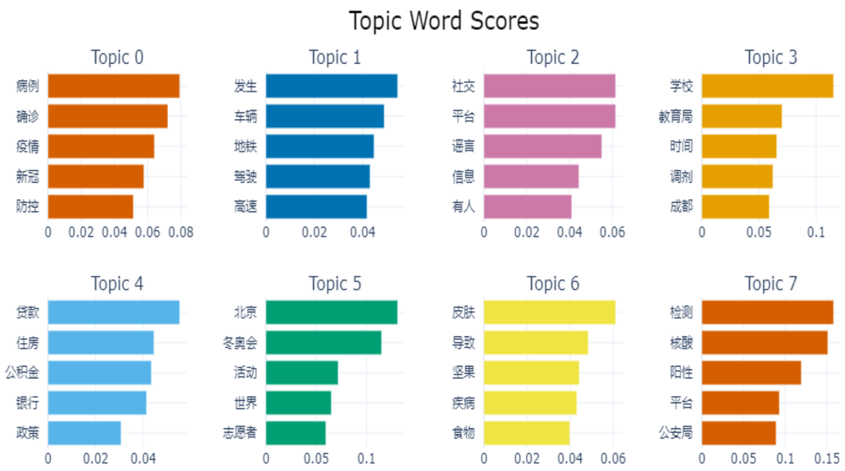


Figure 6: Topic word scores of each topic

I also conducted a hierarchical clustering of topics. It can be clear seen that topics like " 核酸、确诊、社交" are in the same cluster while " 教育局、公积金、冬奥会" are in the other cluster.

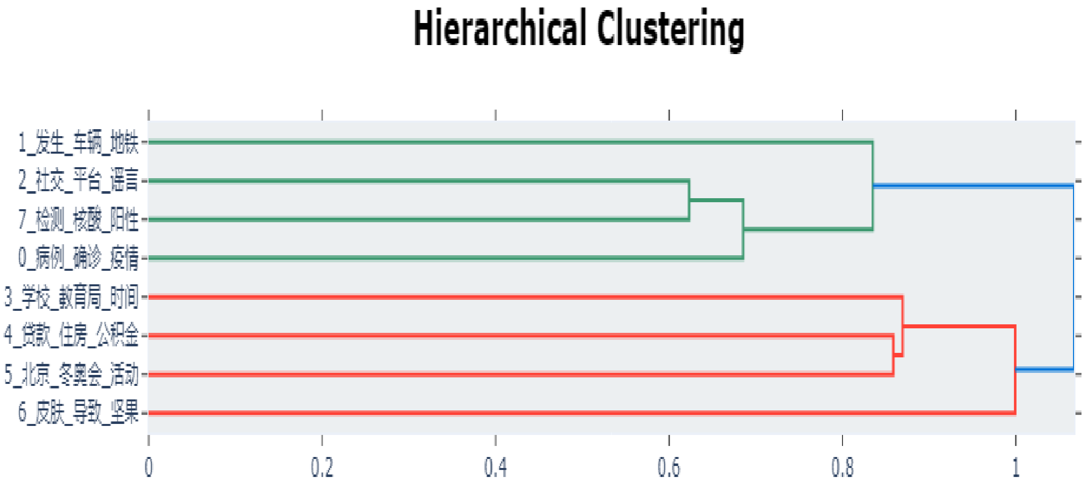


Figure 7: Hierarchical Clustering of each topic

Also the correlation heatmap is shown below. The top left corner is much darker than the other parts, indicating the strong inner relation between topics like " 确诊，核酸，谣言".

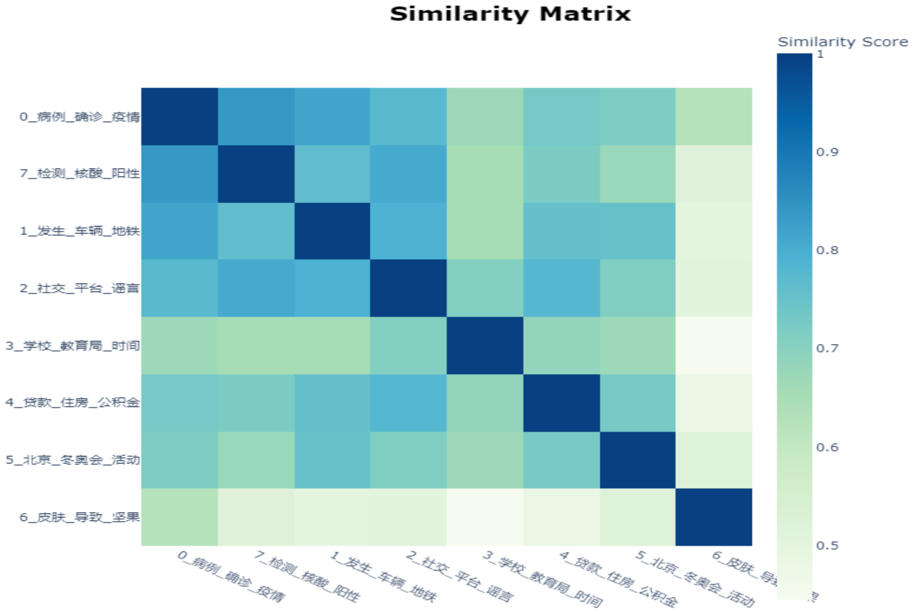


Figure 8: Correlation Heatmap of each topic

4 Changes in Topics over time and region

Here I implement the dynamic topic modeling (DTM) method [3] to explore the change of rumor topics over time and region. Topics over time is displayed in figure 9. Topic0 occupies the largest proportion of documents. Since we have only 366 rumor data, the frequency fluctuation of rumor topics is not very obvious. However, we can still deduce some interesting results. As we can see in this figure, rumor topics can be roughly divided into four stages according to time.

1. Nov 7 to DEC 5:
   In this period, topics like " 贷款、住房"(light blue), " 皮肤、坚果"(pink)，" 驾驶、交通事故" (dark blue)appear frequently. These rumors were on a wide range of topics and not too related to the epidemic outbreak, suggesting a relatively stable situation of the epidemic at the time.

2. DEC 5 to DEC 19:
   Rumor data related to the epidemic outbreak frequently appeared during this time period. There were two major outbreaks of epidemic rumors around December 10 and January 15(light yellow), which may serve as an indicator of possible epidemic outbreaks like Xi’an’s city closure.

3. DEC 19 to Jan 16:
   Rumors related to " 学校、教育局"(green) have a high rate of occurrence in this period. I checked some of the rumors, they mainly related to the start time of school after winter break. It is reasonable to have such rumor topics near the end of the winter break.

4. Jan 16 to Feb 13:
   This is a period of rumor data mainly related to the Beijing Winter Olympic Games(dark yellow). These time-series rumor data can also give us some insight into the events of the time.

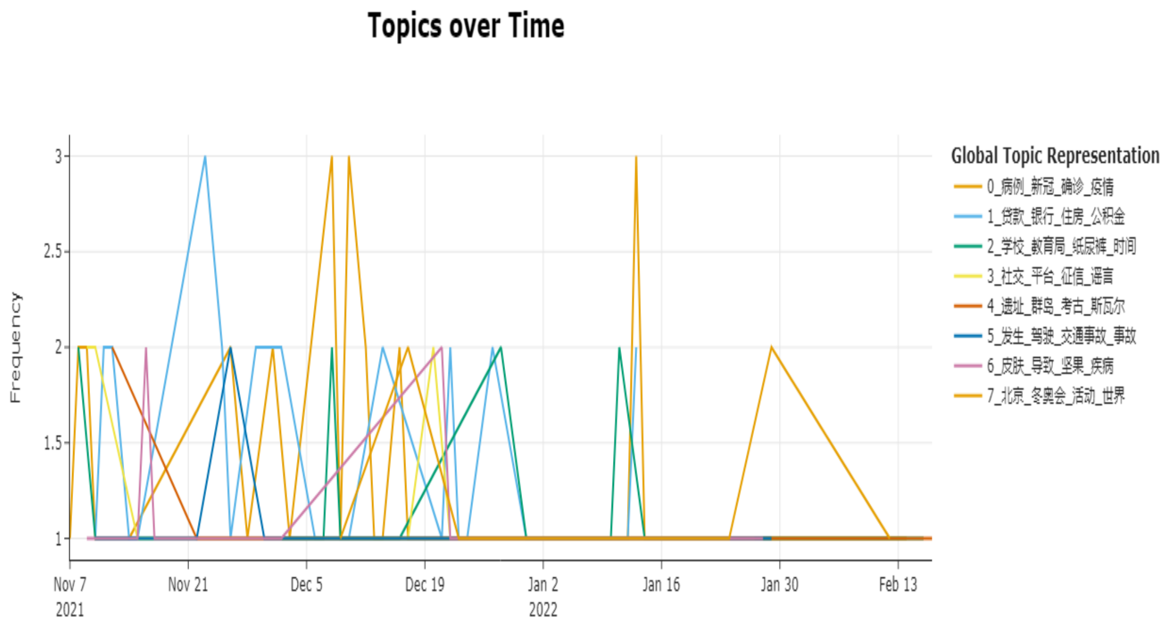


Figure 9: Topics over time

To specific the rumor data changes over time in a province, here I choose Zhejiang province since it has the most rumor data. Its' epidemic-related rumor data begins at Dec 5. During Dec 19 to Jan 5, topics like " 阳性、检测" appear frequently and constantly, indicating that Zhejiang was experiencing a hard time of epidemic. Also, rumor data like " 警方、社交" appears around Dec 19, which is also the beginning of epidemic-related rumor data. This can be the proof that the outbreak of a public health crisis can trigger a series of social security rumors and unrest.

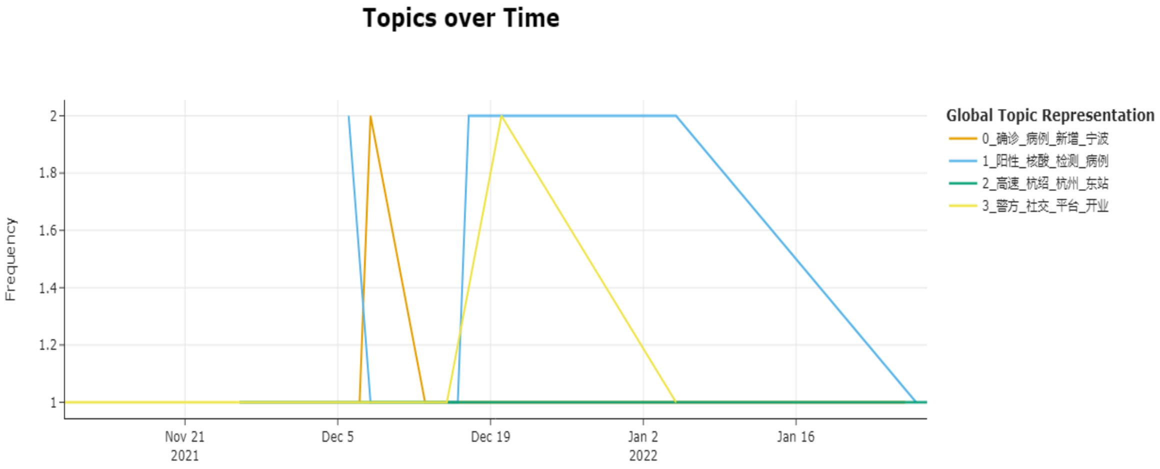


Figure 10: Topics over time in Zhejiang province

5 Geographical Rumor Distribution

Base on the pyecharts module, I conducted the rumor area distribution mining using geographic data information and visualized the results in the form of maps. The regional distribution of the total rumor records are displayed in figure 11. As we can see, 浙江、陕西、四川 accounte for the largest number of rumor data. To be specific, the number of rumors in Zhejiang significantly exceeds that of other provinces, with a total number of 52. Most provinces contain rumor records less than ten.

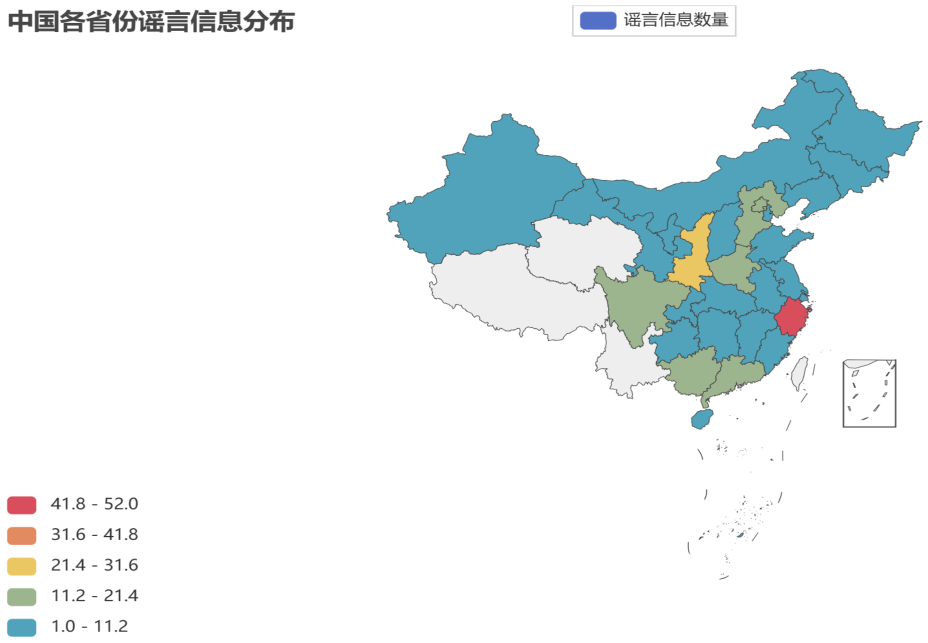


Figure 11: Total Geographical Rumor Distribution

Also, to figure out the geographical distribution of specific rumor topics, I choose two rumor topics： epidemic and loan based on the BERTopic and visualize them below. As we can see in the epidemic distribution(left), 浙江、陕西 still accounts for the largest proportion of epidemic rumors. According to the time series analysis in figure 9, I searched for possible epidemic issues in Zhejiang and Shanxi around 2021 Dec 10 and 2022 Jan 15. Zhejiang province issued 《关于全面从严从紧加强疫情防控工作的紧急通知》 in 2021 Dec 11, which may explain the outbreak of rumors since the issuance of emergency documents often leads to rumors and misinformation. In late December 2021 and early January 2022, Xi'an experienced a massive outbreak of a local epidemic and a city closure. As we all know, the policy of sealing the city is easy to start rumors.

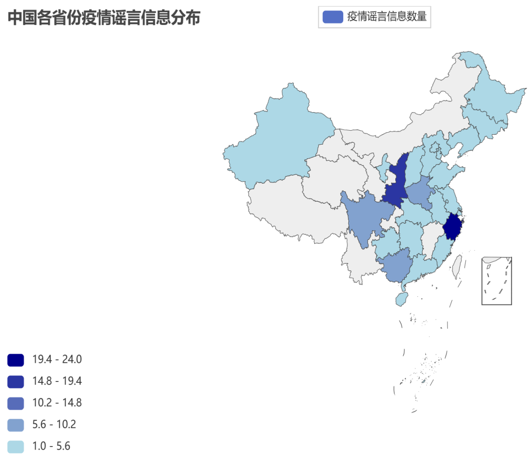


Figure 12: Geographical Epidemic Rumor Distribution

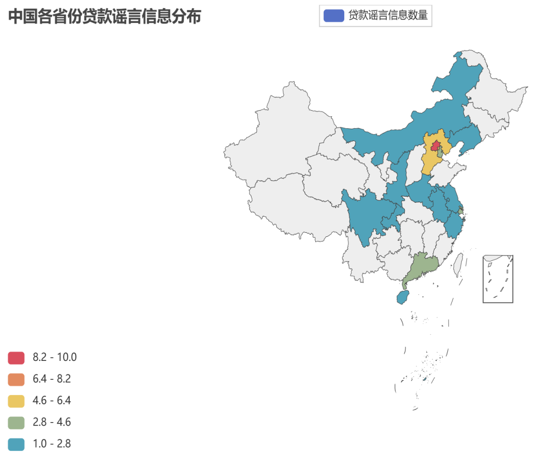


Figure 13: Geographical Loan Rumor Distribution

Through the time and geographical location of the rumor data analysis, combined with the specific events of that time, we successfully come up with some reasonable explanations.

6 User Network Analysis

Here I build a user network based on the users' browsing history for different rumors. By implementing the networkx module, I visualize the user network with nodes and edges within the network.

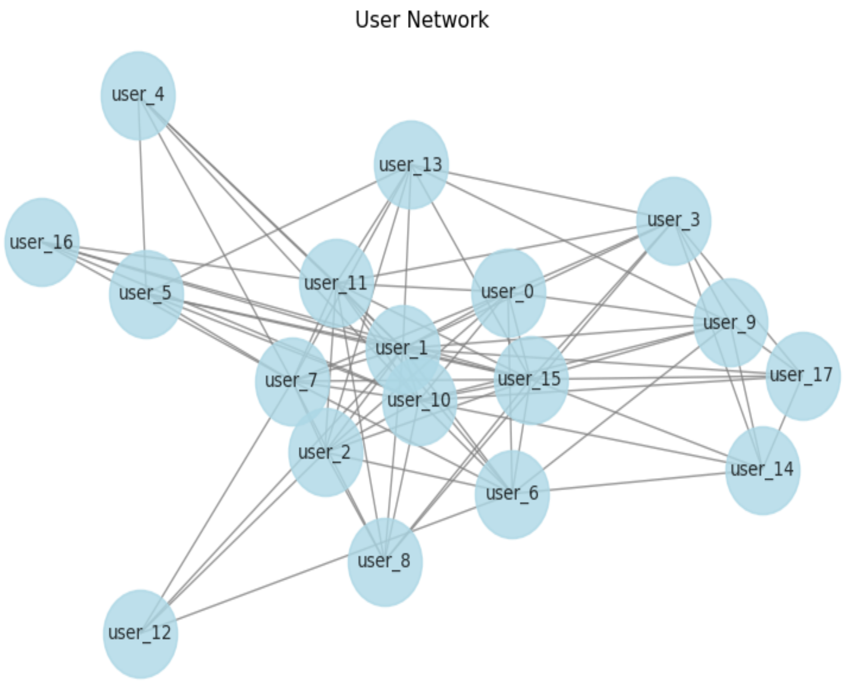


Figure 14: User Network Structure

It is not very straightforward to see the centrality of different users and their preferred characteristics directly from the figure above. Therefore, I compute the Degree Centrality, Betweenness Centrality and Rumor Preference of each user in the user network.

1. Degree Centrality: Degree centrality is a measure in network analysis that quantifies the importance or centrality of a node (or user) based on the number of edges (or connections) it has. In the context of a user network, degree centrality indicates how well-connected a user is to other users.

2. Betweenness Centrality: Betweenness centrality quantifies the extent to which a node (or user) lies on the shortest paths between other nodes in the network. Users with high betweenness centrality act as bridges or intermediaries between different parts of the network. They have the potential to control or influence the flow of information between users.

3. Rumor Preference: Rumor preference refers to the extent to which users are inclined or interested in engaging with rumors or pieces of information. By examining whether users have viewed or interacted with specific rumors, we can determine their preference for or engagement with those rumors. Understanding users' rumor preferences can provide insights into their interests, beliefs, and potential influence in the spread of rumors within a network.

User_id	Degree Centrality	Betweenness Centrality	Rumor Preference
user_0	0.5294	0.0081	0.0410
user_1	**0.9412**	**0.1272**	**0.1885**
user_2	0.5882	0.0191	0.0492
user_3	0.5294	0.0226	0.0301
user_4	0.2941	0.0012	0.0109
user_5	0.3529	0.0044	0.0383
user_6	0.5882	0.0313	0.0464
user_7	0.7647	0.0604	**0.1202**
user_8	0.4118	0.0034	0.0383
user_9	0.5294	0.0154	0.0383
user_10	**0.8235**	**0.0889**	0.0874
user_11	0.7059	0.0378	0.0956
user_12	0.2353	0.0000	0.0273
user_13	0.4706	0.0129	0.0246
user_14	0.3529	0.0036	0.0219
user_15	**0.8824**	**0.0837**	0.0956
user_16	0.2941	0.0000	0.0191
user_17	0.3529	0.0021	0.0219

As we can see in the table above, users with high Degree Centrality also tend to have high Betweenness Centrality and Rumor Preference, making them in the center of the user network 14. Also, to better group users with similar features, I implement a modularity-maximization algorithm called the Louvain algorithm to perform community detection of the user network. The algorithm divides users into three different categories, which also fit well with their own clustering in the user network.

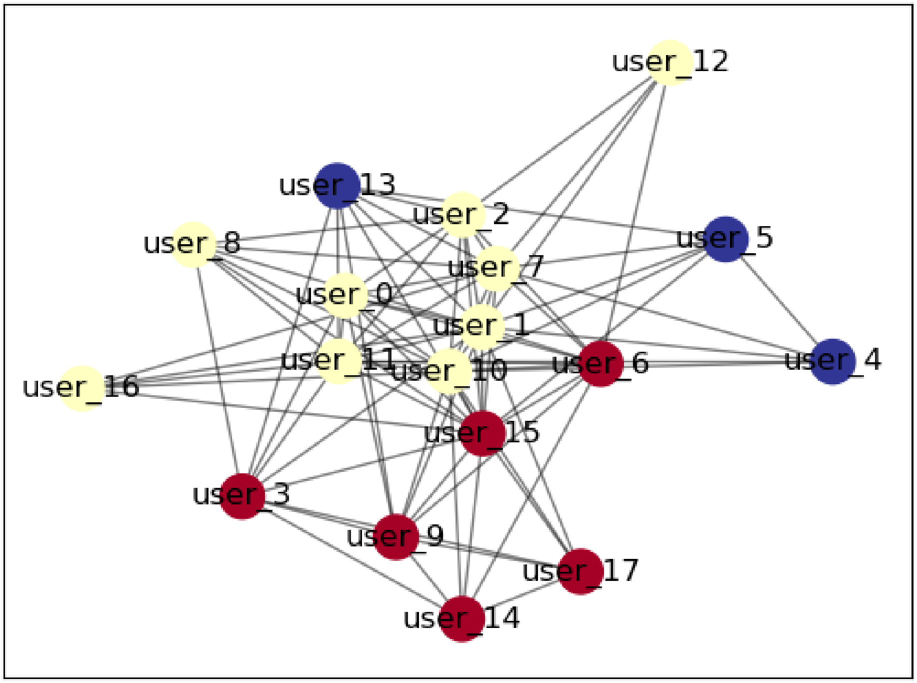


Figure 15: Community detection with Louvain algorithm

7 Popular Rumor Data Analysis

In the period of an epidemic outbreak, we want to investigate the type of rumor data that would attract the highest number of likes and achieve the widest distribution. Consequently, we have selected the rumor data with the most likes and edges with others, intending to perform topic modeling and other analysis on this selected set of data.

First, I choose the rumor data with number of likes more than 50 as rumor_hot(137 data in total). The number of likes can reflect the spread and popularity of rumor information to a certain extent. In this case, rumor_hot can represent some of the rumors that were widely circulated during the epidemic. The topic modeling conducted by LDA is shown below. As we can see, subjects like " 确诊、新增、阳性、核酸检测" still accounts for the largest proportion, indicating that during the epidemic prevention and control period, rumors about new positive cases are the most concerned and most widely spread.

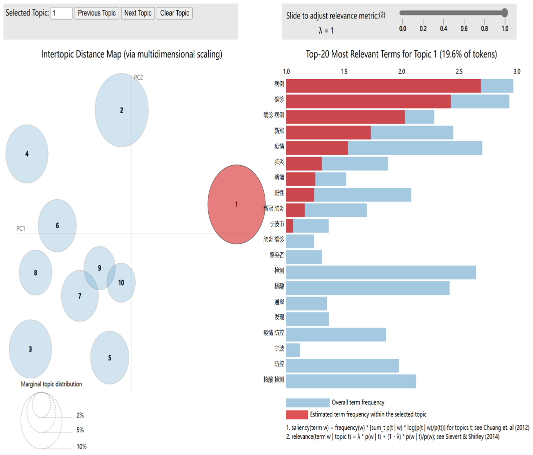

Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-20 Most Relevant Terms for Topic 1 (19.6% of tokens)
1.0
1.5
2.0
2.5
3.0
病例
确诊
确诊 病例
新冠
疫情
肺炎
新增
阳性
新冠 肺炎
宁波市
肺炎 确诊
感染者
检测
核酸
通报
发现
疫情 防控
宁波
防控
核酸 检测
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 16: Topic1 of Popular Rumor Data

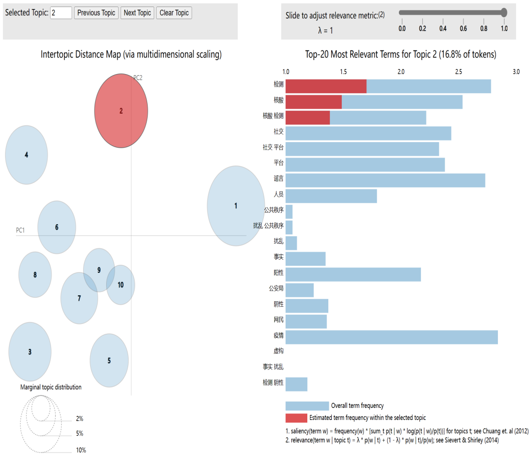

Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-20 Most Relevant Terms for Topic 2 (16.8% of tokens)
1.0
1.5
2.0
2.5
3.0
检测
核酸
核酸 检测
社交
社交 平台
平台
谣言
人员
公共秩序
扰乱 公共秩序
扰乱
事实
阳性
公安局
阴性
网民
疫情
虚构
事实 扰乱
检测 阴性
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 17: Topic2 of Popular Rumor Data

The wordcloud for popular rumor data is also displayed for more straightforward understanding.

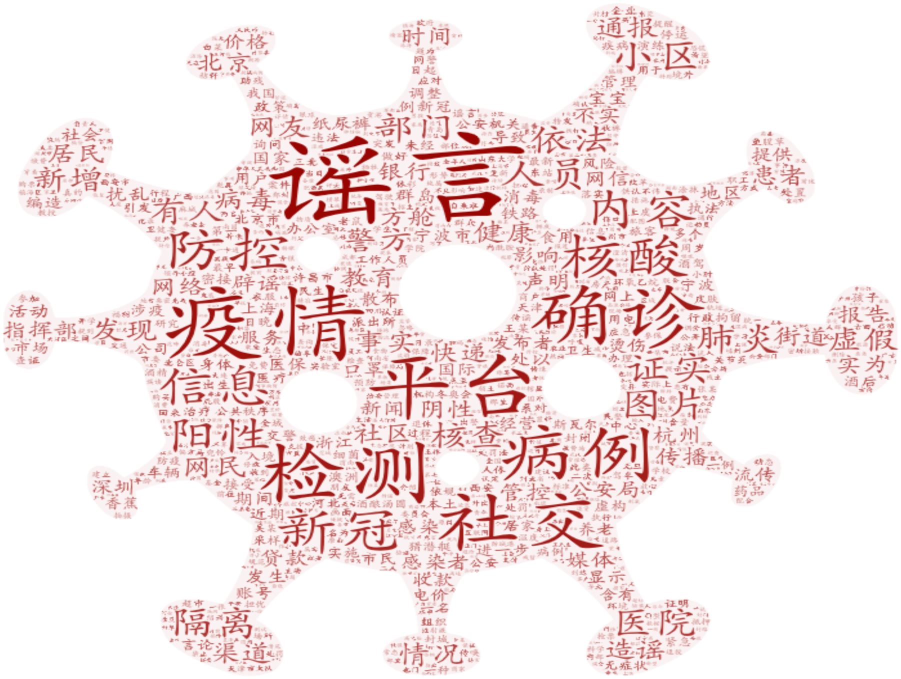


Figure 18: WordCloud of Popular Rumor Data

Social media also plays an important role in the spread of the rumor. There are 306 different social medias altogether in our rumor dataset. Here I count the number of rumor records of the top 20 social medias. Histogram of the number of rumors is shown below. I check the content of rumor data in "中国新闻网", the site mainly contains fake news from the Internet and disinformation of it. Social media with high counts can be a good place for us to find official disinformation.

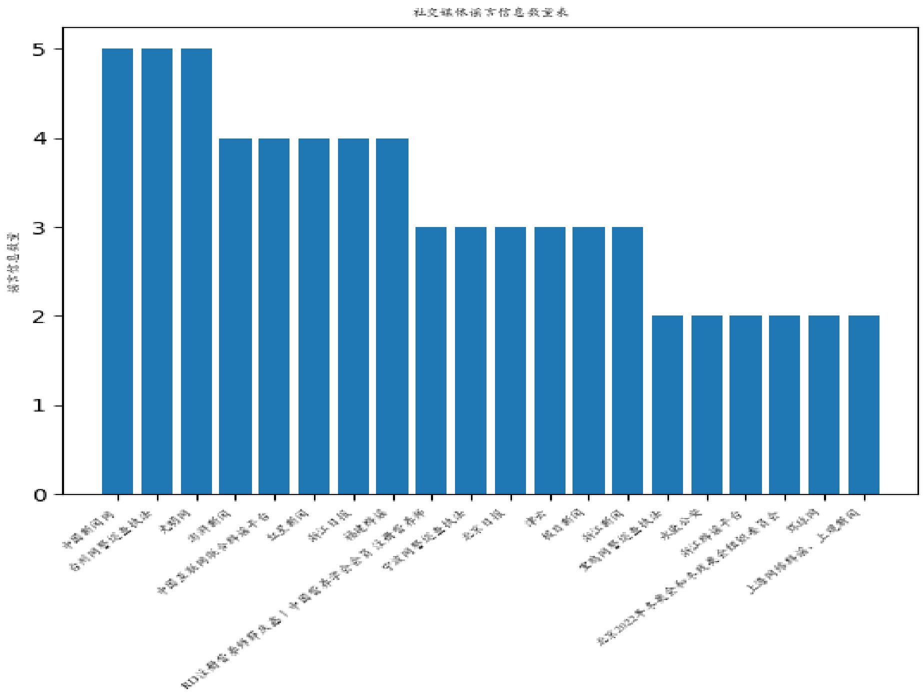


Figure 19: Histogram of Rumors in Different Media

8 Conclusion

In this project I design a rumor visualization system for visualizing rumors in social networks during the epidemic. I conduct Topic Modeling of the rumor dataset and compare different method of LDA and BERTopic. LDA is a convenient and intuitive way for topic modeling while BERTopic has better word embedding and can understand the hidden semantics of the document. Topics like " 疫情、核酸、封控" and " 警方、平台、处罚 “appear most frequently in our rumor dataset, which is reasonable in the period of epidemic.

Using dynamic topic modeling (DTM), I explore the change of rumor topics over time and region. I divide the rumor topics into four time periods and give reasonable explanations of the changes in topics overtime, like Xi’an’s city closure and Zhejiang’s epidemic policy adjustment. Geographical Rumor Distribution is also conducted in this project. Zhejiang and Shanxi province occupy the largest proportion of rumor data. Meanwhile, I perform a geographic analysis based on different rumor topics like " 疫情", " 贷款" and visualize them in the form of China map.

Based on the 18 users’ browse records, I build a user network and compute the degree centrality, betweenness centrality and rumor preference of each user. Community detection using the Louvain algorithm is also performed to demonstrate deeper user connections. Popular rumor data is specially listed according to the number of likes. Through topical modeling of popular rumor data and media analysis, we have a better understanding of the spread of rumors and features of popular rumor.

References

[1] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[2] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics, 2018.

[3] Hao Sha, Mohammad Al Hasan, George Mohler, and P. Jeffrey Brantingham. Dynamic topic modeling of the covid-19 twitter narrative among u.s. governors and cabinet executives, 2020.