Social Media and Epidemic Modeling

(Corona)

Team 3 Members:

ZHANG, Cao: 20803018

LU, Zijun: 20784286

HUANG, Xi: 20785694

1. Introduction

Since the outbreak of COVID-19 globally in 2019, the pandemic has had a huge impact on human society and people's daily life. The governments of many countries take action to stop the spread of disease, including lockdown, quarantine, and promotion of vaccines. As the number of COVID-19 confirmed cases increases rapidly, people are asked to keep social distance and work from home. This measure reduces people's social interaction with each other and makes people shift their life online. Twitter(Weibo in China) is a popular social media platform, its user groups express various opinions on some political values including COVID-19 related topics. Thus social media provides a powerful data resource, which helps us to gain insights into public knowledge and attitudes towards COVID-19. With a better understanding of public attitude, society is able to solve some issues like vaccine hesitancy effectively.

Our study is aiming to analyze and evaluate the change of engagement of COVID-19 related discussion over time and in different locations on social media, including the sentiment analysis,

etc. under some machine learning models.

This proposal includes three literature reviews in which we gained inspiration and motivations that enable us to come up with research questions.

2. Motivation behind the proposal

The epidemic situation is different in different regions, which may be related to beliefs, cultural background, and government management. For instance, some users refuse to be vaccinated, since they have little knowledge of it. We hope to analyze the potential intentions and real thoughts of users through social networks. This allows the government to respond positively to what people really care about, so that people have a positive attitude about prevention measures, like masks and vaccines, which becomes a leading cause of reducing the spread of the epidemic.

On the other hand, time is also an important role in our project. As time goes on, different regions may have different degrees of involvement in the discussion. For example, at the beginning of the COVID-19, the degree of involvement among Chinese users may be relatively high, while the degree in other countries may be relatively low, and vice versa after a while. We want to correlate the COVID-19 topic participation with the epidemic situation in the region.

In this project, we expect to learn how different semantic machine learning models work, and what the differences between them are. In addition, it is also a learning objective on how to combine these models with social networks to achieve good results.

3. Literature review

Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China

Summary:

Xuehua Han and Juanle Wang et al.(2020)[1] conducted a study on exploring the public perception of the COVID-19 at its early stage in China. They aimed to extract and identify public opinion from the social media, Weibo, and analyze the spatial-temporal characteristics based on Weibo texts.

To get started, they firstly used Weibo Application Programming Interfaces (APIs) to retrieve related texts with specific keywords and timestamps and filter interfering information to eliminate noise and improve efficiency. After data collection with the result of 105,330 texts in the dataset, they built a topic extraction and classification model based on the LDA model and the random forest (RF) algorithm to classify the topics of Weibo texts relating to the COVID-19 and mine the sub-topics through those related texts. Within the timestamps and location information, they conducted the time-series analysis using the Seasonal-Trend decomposition procedure based on Loess (STL) and the spatial analysis based on the distribution of Weibo texts using statistic solutions(Kernel Density Estimation and Spearman Rank Correlation).

In their result, they obtained 7 topics and 13 sub-topics related to the disease from Weibo texts and discussed the time trend and spatial distribution of the epidemic-related Weibo texts.

The contribution of this study to the existing research is that it has built a reliable model to mine public latent opinions about COVID-19. And the findings of this study can assist the decision-makers to have a better understanding of public opinions toward epidemics and support analysis in planning and executing appropriate resource allocation.

Although the result is accurate, there are some limitations. In the paper, they only analyzed texts but ignored other contents such as pictures. And they just analyzed the time trend of the topic and spatial distribution of the topic respectively without combining them together.

Critique:

We found something that may be helpful for our project from this paper. The LDA model and RF algorithm can be used to identify the topic of the texts. And the statistical solutions can be used for time and spatial analysis.

We also get some new ideas from this paper. We can combine the time and spatial analysis with contents together to see the evolution of public perception clearly instead of discussing them separately. Besides, we can implement other machine learning models such as SVM[2] or VADER to further identify the sentiments of the public rather than just extract sub-topics.

Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning

Summary:

Over a year since the coronavirus disease (COVID-19) first appeared, the researchers conducted a data monitoring study based on Reddit social media platform in North Carolina [3].

Considering the title, label, comments, and the body of a specific post, they used both Reddit Application Programming Interface and the Python Reddit API Wrapper to collect the data from 18 location-specific subreddits, and further preprocessed the data by several steps, such as tokenization, part of speech tagging and etc. Moreover, the authors used the Bidirectional Encoder Representations from Transformers (BERT) language model to capture the context in which sentences appear within Reddit posts, and identified 5 main categories with

high-frequency words. Overall, this study observed people's thoughts and concerns. In the early days, people may focus on the discussion of how to reduce the spread of COVID-19, wherein the mid-term, people pay more attention to preventative measures, and there is a significantly positive change in attitudes towards masks for residents in North Carolina. Another finding is that the representative posts or data on social media can be utilized to surveil the local epidemic situation in a specific area.

The strength of this article is that the authors propose a machine learning method that builds the relationship between social media and the epidemic situation, thereby reflecting people's opinions and attitudes. However, the weakness is also obvious. Not every region in North Carolina has a subreddit community, and they either cannot guarantee everyone posting in the subreddit still lives in North Carolina. These two limitations may lead to the conclusion inaccuracy. Since the data set used in our group is relatively large, covering all the information in most regions, it is guaranteed that these twitter comments are the main opinions of most people. Additionally, our analytical approach is based on time, which can avoid the second weakness mentioned before. For example, if someone posted a negative statement in the early days, and then left the city in the medium term, it doesn't affect our conclusion that people have a negative attitude in that city at the beginning.

Critique:

The specific objective of this thesis was to monitor the COVID-19 pandemic overtime on social media by using Bidirectional Encoder Representations from the Transformers language model, which is strongly related to our project. The BERT model might become one of the options we use to analyze Twitter plain-text data. In addition, the data preprocessing and high-frequency words analysis method mentioned in this paper will also be one of the references.

Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis

Summary:

The study aims to analyze the public sentiment in regards to COVID-19 vaccines on Twitter[4].

COVID-19 vaccine has been considered as the key in order to achieve herd immunity by many countries since the outbreak of the pandemic. Many studies have conducted tweets sentiment analysis only on COVID-19 disease. However, the research team considers the public attitude towards COVID-19 vaccines is important to promote citizens getting vaccinated. Twitter as an important social media platform provides tones of public discussion that reflect different attitudes and opinions in different locations. Also the user group of Twitter across different gender, race, and age.

The research questions focus on the sentiment variation trend before and after the announcement of Pfizer's vaccine and the difference in people's sentiment changes according to different geographic locations. The research team collects tweets on the topic of COVID-19 vaccines in English from November 1st, 2020 to January 31th, 2021. The Valence Aware Dictionary and sEntiment Reasoner tool is applied to compute the compound score for sentiment analysis. The main topics for tweets are extracted via latent Dirichlet allocation analysis. For geographic analysis, the research team used the python package Twitter_Geoloaction to map US users in tweets to state-value pairs, then conduct geographic analysis.

The results of the research show us most of the tweets have positive sentiments (42.8%) among a total of 2678,372 related tweets. Brazil has the lowest sentiment score whereas the US

has the highest sentiment score. The public attitude on vaccines varies rapidly over time and geography.

The strengths of this paper include the thesis provides a two-dimensional analysis of both time and geography and the topic is focused on the attitude of the public towards COVID-vaccine. The methodology section clearly shows the data collection process and data analysis methods discussion.

However, the analysis conducted by the study is only based on tweets in English, but for some regions like Brazil, people mainly speak Portuguese, which will limit the representation of the conclusion in some areas.

Critique:

The research questions proposed by this study inspired us in terms of sentiment analysis on tweets and the sentiment changes in relation to time and geography. The analysis tools and methods applied in this study are also feasible and suitable for our research, such as the Valence Aware Dictionary and Sentiment Reasoner to compute the sentiment scores. In terms of temporal analysis, we could also apply the Pruned Exact Linear Time to plot the distribution of sentiment scores changes over time.

4. Brainstorming

First, we would like to analyze public opinion on some specific aspects over time, for example, the attitude on social distance overtime or the attitude on vaccines or the attitude on wearing masks, etc. But, we consider that the amount of data on those specific topics may not be adequate for us to analyze since we are not sure what will be contained in the datasets. Also, we

think those cannot show the evolution of public perceptions about the epidemic comprehensively since they are too specific.

After we know the structure of the data we can obtain, we decide to study and analyze the data combining 2 dimensions together -- time and space. And since we need to study public perception, we decide to focus on the topic of tweets and Weibo texts and further analyze the latent sentiments from the contents of tweets and Weibo texts. And it will be valuable for us to show and discuss the results of the topic and public attitude along with time series and spatial distribution together.

5. Proposal

5.1. Research questions

RQ1: Which region is most engaged with the epidemic?

RQ2: How do people's attitudes to the epidemic change over time in different regions?

RQ3: What are people's favorite main topics when discussing the epidemic in different regions?

5.2. Dataset

In this project, we intend to use the data extracted from Twitter and Weibo since large-scale datasets related to the recent CoronaVirus have already been collected before. And to further alleviate the workload, we would use tweets written in English only. In order to make the results more accurate, we may not only deal with texts but also deal with emojis in the analysis of public attitude.

There are some typical data types in the dataset that contain the most valuable information for our project.

- Create_at: The timestamps of the tweet/Weibo text when created.
- User_location: The country where the tweet/Weibo text was posted.
- Text: The contents of the tweet/Weibo text.
- Hash_tags: The tags of the tweet/Weibo text may be helpful for us to distinguish the topic.
- Lang: The lang can be used to select the English texts for us.

5.3. Methodology

To answer our research questions, we would process tweets in the dataset by classifying with different timestamps and further with different location information, transferring emojis to readable texts, and removing duplicate tweets. And then try to use machine learning models such as the LDA model or BERT model to extract and identify the topic of processed data to specify the change in the content of public discussions. After that, we would use other machine learning models such as SVM or VADER to distinguish the public attitude towards the COVID-19 over time. At last, we would use some statistical methods to present and discuss the result.

6. Evaluation

To evaluate the reliability of the analysis results of Twitter data, we divided the dataset into the training set and testing set with a ratio of 9:1, among which, we will continue to divide the training set into a smaller training set and validation set with the same ratio of 9:1 for early-stop in machine learning, which is used to prevent overfitting. The possible evaluation

measure method would be the Mean Squared Loss function, Cross-Entropy Loss function, or any other useful function that we find during the experiment. We will judge the accuracy of our methodology by the results on the testing set, and adjust the model parameters as much as possible to achieve at least 70% accuracy.

7. Explanations about differences between others

In this project, our groups will mainly focus on two dimensions of social media: Time and Space. We will use the Machine Learning model to analyze the people's attitude, degree of participation, and high-frequency words based on these two dimensions.

Group 11 will only consider the time dimension, which is different from our group. Same as our group, group 6 will also consider these two dimensions. However, they may pay more attention to the Twitter posts about “work from home”, where our group may evaluate all the relevant and representative posts.

8. Reference

1. Han X, Wang J, Zhang M, Wang X. Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China. International Journal of Environmental Research and Public Health. 2020; 17(8):2788. https://doi.org/10.3390/ijerph17082788
2. Schuldt, C., Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 3, pp. 32-36). IEEE.

3. Liu Y, Whitfield C, Zhang T, Hauser A, Reynolds T, Anwar M. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. Health Inf Sci Syst. 2021 Jun 25;9(1):25. doi: 10.1007/s13755-021-00158-4. PMID: 34188896; PMCID: PMC8226148.
4. Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. Vaccine. 2021 Sep 15;39(39):5499-5505. doi: 10.1016/j.vaccine.2021.08.058. Epub 2021 Aug 17. PMID: 34452774; PMCID: PMC8439574.