University of Nottingham Ningbo China

Business School

Academic Year 2011/22 Spring Semester

Analytics Specializations and Applications

Prof. Xi CHEN

Tweet Analysis

Linfang Wu (Student ID: 20415839)

Word count:2883

Contents

1. Executive Summary

The Public's view of certain social issues reflects their values and attitudes. In this project, 2 datasets included the keywords "covid victim" and "covid survivor" have been provided to understand whether and how people's attitudes, emotions, or opinions may be different when they include these two different keywords. Both datasets have around 3000 and 4000 records respectively with 23 features, such as tweet text, create time, user location, and other users mentioned. Detailed information has been presented in the description section.

To achieve the purpose, the whole analyzing process could be divided into 5 parts, which are data cleaning, sentiment analysis, topic modeling, influencer, and geographical analysis. Techniques such as stemming, and lemmatize have been used in text cleaning; LDA for topic modeling, and newtworkx, pyecharts for influencer and geographical analysis.

There are several differences between the two datasets. First, there is a greater sentiment score variation in the victim group, and the proportion of users holding a negative attitude is mostly the highest; while users in the survivor group generally are more optimistic and emotionally stable. In terms of the topics, though both groups mentioned victims of the pandemic, the victim group mentioned more about economic destruction while the survivor group cared more about the medical treatment. Finally, there are no big differences in terms of geographical sentiment change, but users from Canada are more optimistic than those from USA.

Overall, as the keywords included, users in the victim group focused more on the downside of the covid, and therefore, attitudes are generally more negative; while users in the survivor group paid more attention to the survivor and medical care, and their attitudes are more optimistic and stable. Influencer text sentiment analysis and topic-changing tracking have been suggested to conduct in the future to better understand the reasons for the phenomenon.

2. Approach breakdown

The whole project analysis could be broken into 5 parts: basic data cleaning, sentiment analysis, geographic analysis, topic modeling, and influencer analysis. More detailed sub-breakdown parts in each part have been presented in table 1 below.

Breakdown Parts	Analytic Process	Technique Applied
Basic Data Cleaning	Text & source format cleaning, Stemming, Lemmatize	Neattext, Stemming, WordNetLemmetizer
Sentiment Analysis	Keywords could, text sentiment scoring, temporal sentiment changing	TextBlob, wordclouds, time aggregation
Geographic Analysis	Geographical sentiment aggregation analyzing	Pyecharts
LDA Analysis	Topic modeling	LDA
Influencer Analyzing	/	networkx

(Table 1. Approach breakdown)

2.1 Basic Data Cleaning

URL links, user mentions, emojis, hashtags, and some special characters have been removed from tweet text. In addition,' covid' or 'covid19' have also all been removed. After removing, these tweet texts have been split into words to conduct stemming and net lemmatize, and all the words have been converted into original form before analyzing. In addition, tweet-create-time has also been converted into DateTime format.

2.2 Sentiment Analysis

By using the TextBlob package, each tweet text got a sentiment score and has been categorized into the positive, negative, and neutral groups. By calculating the mean and sum of sentiment scores on a daily or hourly basis, the trends of users' attitudes changing during the period could be observed.

2.3 LDA Analysis

To understand the topics of each dataset, topic modeling technique has also been used. The number of topics has been decided according to perplexity and coherence scores. By doing this, topics of each dataset could be spotted, which is helpful for understanding the differences between the attitudes when including two different keywords.

2.4 Influencer Analysis

In order to understand the correlation between influencers and the attitudes of users, the sum of the sentiment score has been calculated for each influencer. By comparing the sentiment scores of the top 10 influencers in both datasets, it is possible to understand the influences of these opinion leaders.

2.3 Geographic Analysis

Since user location is self-defined, a database from Kaggle called "world cities", containing global cities and corresponding countries, has been used for reference. After uniform, the mean sentiment score has been calculated and projected to pyecharts map to examine the geographical sentiment differences.

3. Data Description section

An exploratory data analysis includes the number of tweets, users, time range, geographical area, as well as mentions, have been quickly detailed as the basis of further analysis.

3.1 Tweets Number & Item Analysis

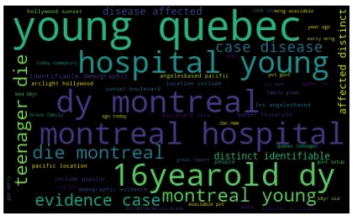
After removing the duplicates, information from both datasets have been presented in table 1. Both datasets have 23 features, which are self-explainable.

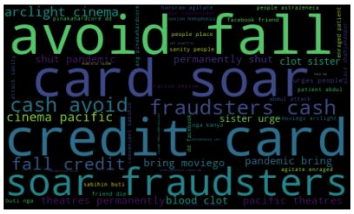
Properties	Victim	Survivor
Users Number	3046	4408
Tweets Number	3289	4725
Time Range	2021/4/5 -2021/4/13	2021/4/5 -2021/4/13
Tweet features	tweet_id, user_id, User_name, User_screen_name, User_location, User_description, User_url, User_followers_count, User_following_count User_favourites_count, User_status_counts, User_created_at, User_verified, Tweet_created_at, Tweet_soure, Tweet_text, Tweet_lang, tweet_favorite_count, tweet_retweet_count, hashtags, mentions, urls	

(Table 1. Datasets information)

3.2 Tweets Keywords

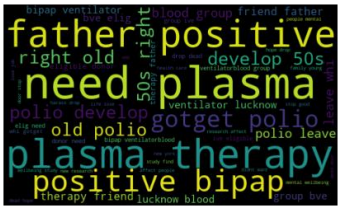
Figure 1 to 3 illustrates the keywords of tweet text in the victim dataset with different sentiments. Several keywords have been mentioned very often, such as young, hospital, police, crime, card, avoid, soar and fall.

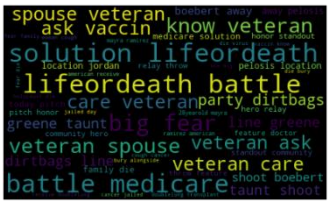




(Figure 1. Positive keywords of victim) (Figure 2. Negative keywords of victim) (Figure 3. Neutral keywords of victim)

Figure 4 to 6 illustrates the keywords of the survivor dataset with different attitudes. Different from the victim, keywords in the survivor group are mainly about medical treatment, such as plasma, medicare, and vaccine. Certain groups of people have also appeared often, such as single mothers and veterans.



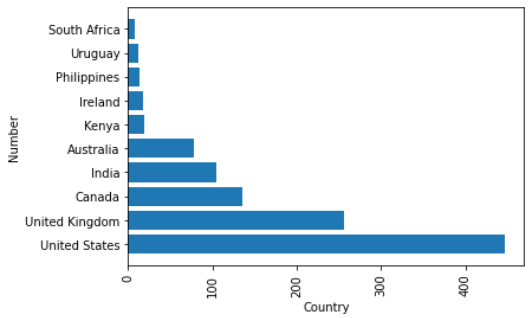


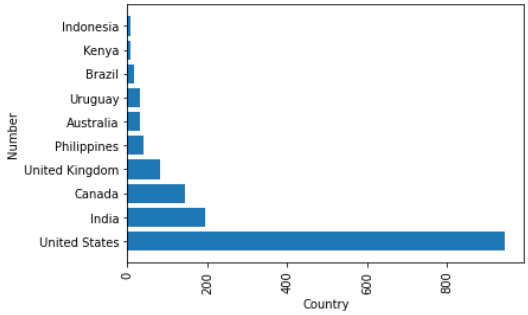
(Figure 3. Positive keywords of survivor) (Figure 4. Negative keywords of survivor) (Figure 5. Neutral keywords of survivor)

Overall, though users both mentioned victims of covid-19, users in the victim datasets focused more on the teenagers, while users in another group paid more attention to mothers and elders. In addition, victim groups mentioned more economic issues, while users in the survivor group care more about therapy. The following topic modeling analysis will provide a more detailed topic analysis of two datasets.

3.2 User & Geographic

Figures 7 and 8 present the global user distribution of datasets. Obviously, users in both datasets are mainly from the USA, UK, Canada, Australia, and India, and therefore following geographical analysis will mainly focus on these 5 countries.





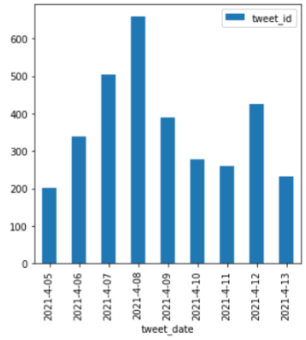
(Figure 7. User geographical distribution (Victim)) (Figure 7. User geographical distribution(survivor))

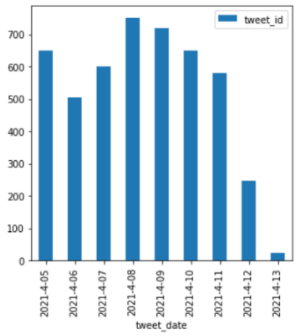
3.3 Tweet Time

Time analysis could be divided into two parts: date and hour.

3.3.1 Tweet Date

Figure 3 and 4 present the number of tweets posted each day. The number of tweets in the victim group increased from 200 to over 600 in the first 4 days and finally end with 250; while that of the survivor group is more stable in the first week, but decreased significantly in the last two days.

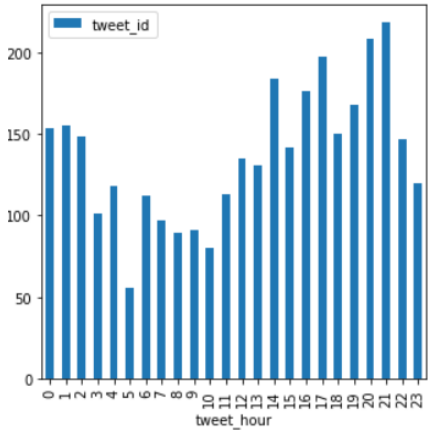


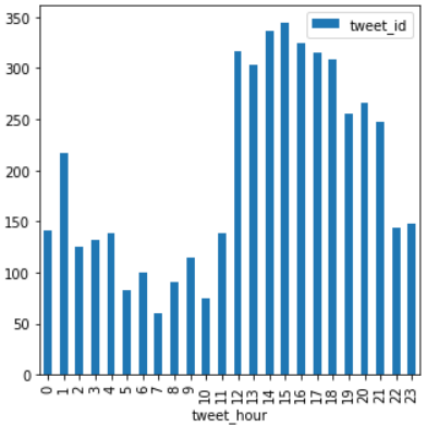


(Figure 9. Tweets number & Date(victim) (Figure 10. Tweets number & Date(survivor)

3.3.2 Tweet Hour

By aggregating the number of tweets on an hourly basis, it has been found that the number of tweets in the victim database reached its peak at the night (20-21:00); while users tend to post more tweets during the afternoon (13-14:00) in the survivor group. It is also reasonable as people tend to be more emotional at night while more rational during the daytime.





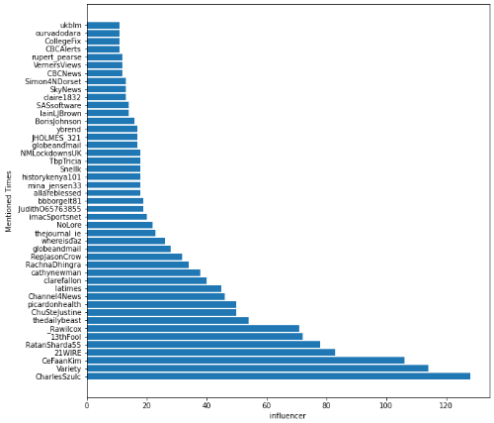
(Figure 11. Tweets number & Hour (Victim)) (Figure 12. Tweets number & Hour(survivor))

The proportion of tweets with different attitudes on daily basis will be further analyzed to observe the attitude-changing trends in the analysis section based on this.

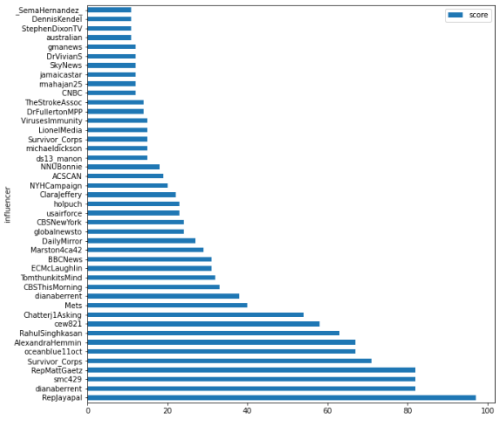
3.3 Mentions

Reasons behind mentioning other users could either be agreed or disagree with certain tweet contents. This implies that tweets from an influencer can subconsciously influence the views of others. In this case, users that have been mentioned frequently, so-called influencers, are worth to be discussed.

Figure 7 and 8 illustrate the mentioned number of different users in both groups. It is obvious that *CharlesSzulc* has been mentioned most often (128) in the victim group, followed by *Variety* and *CeFaanKim*. Differently, though *RepJayapal* has been mentioned most often in the survivor dataset, the number of mentions is lower than 100. Further tweets sentiments analysis will mainly focus on the top 10 influencers, to see if there are certain correlations between influencers and attitudes.



(Figure 13. Tweets number & Hour (Victim))



(Figure 14. Tweets number & Hour(survivor))

4. Analysis section

Based on the description section, a more detailed analysis includeing temporal sentiment analysis, topic modeling, influencer and geographical analysis have been conducted.

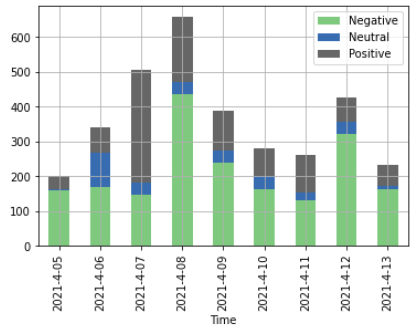
4.1 Sentiment Analysis

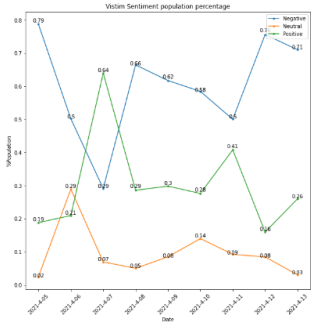
4.1.1 Time Series Sentiment Analysis

4.1.1.1 Sentiment & population

Figure 14 to 17 present the number and percentage of users in different sentiment groups from two datasets.

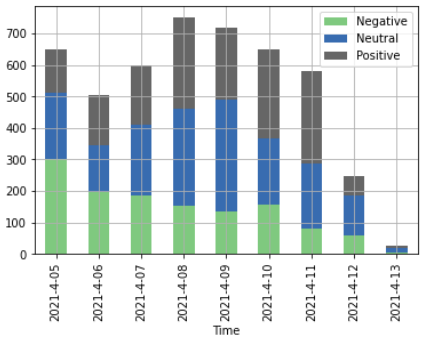
In the victim group, both the number and percentage of users that posted tweets with negative attitudes were the highest (79%) at 4/5. It then decreased to its lowest point on 4/7, which is only around 29% of the total population. However, the percentage of people holding positive attitudes reaches its peak (67%) on the same day. Though there are some fluctuations, the proportion of negative attitudes was the highest most of the time.

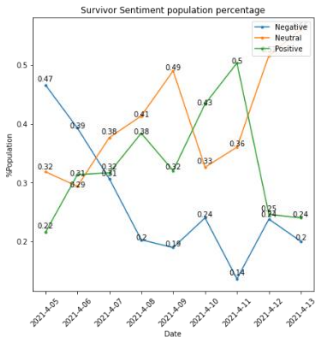




(Figure 14. Tweets number with different sentiments (Victim)) (Figure 15. Tweets percentage with different sentiments (victim))

Similarly, the number of people holding negative attitudes was the highest (47%) on the first day in the survivor group, but it then kept decreasing to 0.2 at the end of the period. That of neutral attitudes showed the opposite trend: the proportion of people holding neutral attitudes kept increasing until it reaches 56% at the end. The trend of positive attitude changing is very similar to that of neutral, excepting the sudden decrease on 4/13.





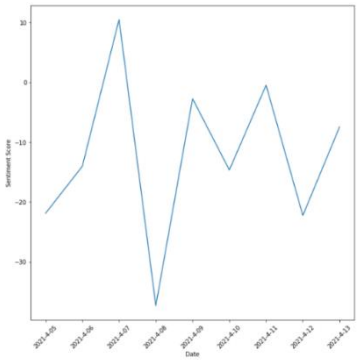
(Figure 15. Tweets number with different sentiments (survivor)) (Figure 16. Tweets percentage with different sentiments (survivor))

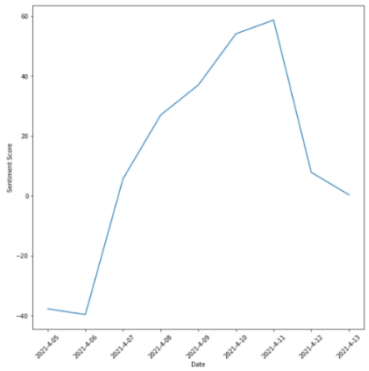
4.1.2.2 Temporal Sentiment Score

Based on the sentiment population analysis, daily sentiment scores have been calculated by simply adding the polarity score to observe the trend of attitude change. If the sum of the sentiment score is positive, it simply implies that more people hold a positive attitude.

Figure 11 and 12 illustrate the sentiment-changing trend over 9 days. The sum of sentiment daily scores in the victim dataset is more fluctuated than that in the survivor group. People started withholding a very negative attitude (-20) and become more optimistic in the following two days. However, the score suddenly dropped to the lowest (-40) and later still end with a negative attitude (-10).

Similarly, users were very pessimistic at the beginning in the survivor group. But people were getting more optimistic, and the score reached its peak (60) on 4/11.





(Figure 17. Daily sentiment change (victim)) (Figure 18. Daily sentiment change (survivor))

Overall, the emotion-changing trends between victim and survivor groups have both similarities and differences. Both of them started with having the greatest number of people holding negative attitudes and ended with holding more neutral opinions. The difference between the two groups lies in the trend evolution process. The change is more intense in the victim group, while that of

the survivor group remains relatively stable. In addition, most users hold a negative attitude in the victim datasets, while users are generally more optimistic in the survivor group.

Though the attitude changing trend is clearer, it is hard to give a reason about why there is a such trend or what influenced the attitude of users. In this case, it is necessary to look into the tweet topics under each group.

4.2 LDA Analyze

Table 2 presents 4 topics in the victim dataset by applying LDA. There are some correlations between topics. Due to the pandemic, businesses could not operate normally, and many people lost their jobs. With the increasing unemployment rate, keywords such as family, and credit card have been mentioned very often in topic 4. This may further cause the increasing crime rate and racism, which is related to topic 1. Another downside of lockdown is the lack of communication, people have to rely more heavily on social media to communicate and get information. This is why the keywords of topic 3 have interview and skype. As the keyword included, victims, especially young victims have also been mentioned in topic 2.

Topic	**Keywords**
Crimes	Victim, crime, people, die, hate, race, nypd, urges, death, blood
Young victims	Victim, young, Quebec, Montreal, die, hospital, case, teenager, evidence, disease
Communication	Victim, police, interview, unable, request, skype, blood, clot, sister, jab
Family & economy	Victim, people, fall, parent, family, report, vaccination, patient, soar, card, credit

(Table 2. Topic modeling of the victim dataset)

Table 3 presents keywords and corresponding topics in the survivor dataset. Similar to the victim group, users also discussed family issues (topic 4). Other three topics are mainly about the therapy of covid 19. They may need plasma for medical cure, and users may also share some experiences about self-cure after getting affected (topic 1). Some users may think the spread and mutation are still a mystery (topic 3), and therefore another topic is about the science, which also mentioned the vaccine (topic 2).

Topic	Keywords
Therapy	Survivor, long, polio, need, people, old, plasma, leave, blood, therapy
Science	Survivor, experience, lose, get, veteran, vaccinated, email, science, contribute, life
Solution	Survivor, cancer, solution, lifeordeath, jab, dead, medicare, skeptical, mystery
Family	Work, survivor, medical, pay, mother, sick, treatment, single, year, research, fear, family

(Table 3. Topic modeling of the survivor dataset)

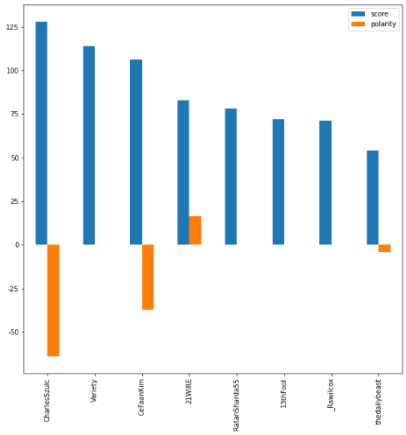
Overall, users in both datasets mentioned victims, including infestor and business owners. However, users in the victim group mainly focus on the crime and economic destruction due to the pandemic; while users in the survivor group discussed more on the therapy and scientific solutions.

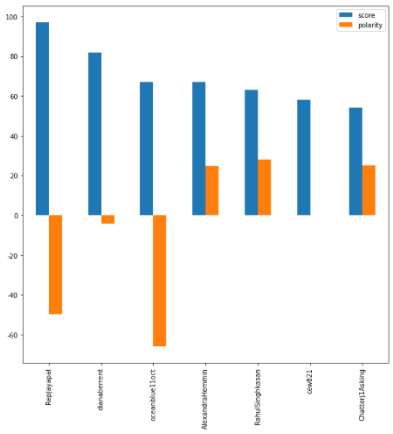
4.3 Influencer Analyze

As the description part has mentioned, the influencer analysis only focuses on the top 10 users that have been mentioned most frequently.

Figure 11 and 12 illustrates the number of mentions and the sum of tweet sentiment score related to these influencers. In the victim group, *CharlesSzulc* has the highest mentioned times (120), but the lowest sentiment score (-65). *CeFaanKim* has the second most negative sentiment score, which is around -37. Among these 10 influencers, only *21WIRE* has a positive score, and the scores of others are all around 0, which implies that users hold a more neutral attitude.

Different from the victim group, three influencers in the survivor group, *AlexandraHemmin*, *RahulSinghkasan*, and *Chatterj1Asking*, all got a positive sentiment score. This implies tweets mentioned these influencers are more likely to hold a positive attitude.





(Figure 19. influencer sentiment score (victim)) (Figure 20. influencer sentiment score (survivor))

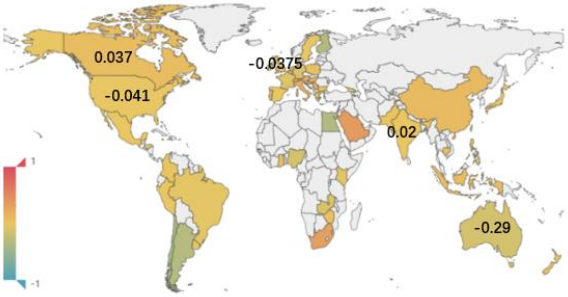
According to data, the attitude of users who mentioned influencers in the victim group is more pessimistic, while the tweet sentiment score that mentioned influencers in the survivor group is more positive. However, it is not clear about whether these influencers changed or manipulated the attitudes of users. In this case, a further analysis suggestion discussed in the following section.

4.4Geographic Analyze

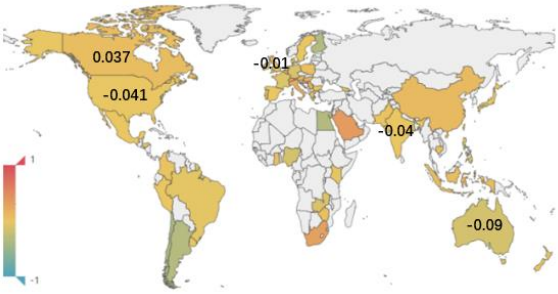
As the description part has mentioned, the geographic analysis mainly focused on these 5 countries, which are the USA, UK, Canada, Australia, and India. The mean sentiment score of each country has been calculated in order to examine the sentiment differences across regions.

Figure 21 presents the mean sentiment score in the victim group in different countries. Obviously, Canada is the most optimistic country, followed by the India. Scores of USA, UK, and Australia are both negative, which implies users are more pessimistic in these two countries.

Figure 22 presents the mean score in the survivor group, which is very similar to the score of the victim. However, there do exist some differences in several countries. The score of Australia and UK in the survivor group is more positive, while that in India decreased significantly from 0.02 to -0.04.



(Figure 21. Tweet global sentiment map (victim))



(Figure 22. Tweet global sentiment map (survivor))

Overall, users from Canada are the most optimistic, while users from USA are most pessimist. Users in India generally held a neutral attitude, and the sentiment score increase in the survivor dataset compared with score in the victim dataset in terms of the rest of 2 countries.

5. Further Analysis Recommendation

Based on the influencer analysis, two recommendations will be provided for further analysis. The first is influencer text analysis. The influencer analysis in this report only focused on the result (sentiment score of user tweets), instead of how these influencers influenced the attitudes of users. In this case, it is recommended to conduct further sentiment analysis on the original text posted by these influencers. And by tracking the attitudes before and after mentioning these influencers, the influence of these opinion leaders on users will be clearer.

The second is topic time-series analysis. The current analysis only discussed the general topics of two datasets, but not covered the changes of topics. By learning the changes of topics along with time, reasons of attitude change would also be clearer.

6. Conclusion

Overall, this tweet analysis mainly focused on three parts, which are tweet sentiment, influencer, and geographical analysis.

In the sentiment analysis, there are two key findings. First, the general attitude of users in the victim group is more pessimistic than users in the survivor group. Second, the variance of daily sentiment scores in the victim dataset is higher than that of another group. It implies that users in the survivor group are more emotionally stable.

In the topic modeling part, there are some similarities and differences between the two groups. Both groups mentioned the victims of covid19. However, users in the victim datasets also discussed the negative impacts on the economy due to the pandemic; while users in the survivor group mentioned more about therapy and medical treatment.

Based on the user location, geographical sentiment map has also been generated, and it has found that users from Canada are most optimistic while users from USA are most pessimistic.

Finally, In the influencer analysis, the sentiment score of tweets that mentioned the top 10 influencers is generally more negative in the victim dataset rather than that of another group. There are two reasonable explanations. The first is users originally agreed with the opinion of these influencers, and this explanation is consistent with the earlier finding that users in the victim datasets are more pessimistic. While another explanation could be that the attitudes of these users were influenced and shaped by these influencers.

Therefore, sentiment analysis on influencer text is recommended to extend. By analyzing the text sentiment score from these influencers and comparing the attitude before and after users mention these influencers, the influences of these opinion leaders on users will be clearer. In addition, the temporal changes of topic analysis could also provide some reasonable explanations of the attitude changing of users.