COVID-19 Topic Modeling and Visualization

Grace Tao
Department of Computer Science and
Software Engineering
Swinburne University, Australia
xtao@swin.edu.au

Yuan Miao
College of Engineering and Sciences
Victoria University,
Melbourne, Australia
yuan.miao@vu.edu.au

Sebastian Ng
Department of Computer Science and
Software Engineering
Swinburne University, Australia
sng@swin.edu.au

2020 24th International Conference Information Visualisation (IV) | 978-1-7281-9134-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/IV51561.2020.00129

***Abstract*—The World Health Organization announced the novel coronavirus (COVID-19) outbreak as a global pandemic on March 11, 2020. To date, this virus has infected over 13 million people worldwide, causing a significant impact on human societies. More than 570,000 people have lost their lives. This study tries to understand the concern and experiences of the general public in this pandemic via social media data analysis. Twitter is a very popular social media platform where people share their ideas and express their views in a timely manner. It is also one of the most comprehensive sources of public conversation. In this work, we collected COVID-19 related Tweets during this pandemic, analyzed the volume, location, frequent words, main topics and people's emotion. We further visualized the evolution of the main topics along with the development of this pandemic. The analysis and visualization can help us to construct a picture of people's concerns and to understand their behaviors. It would also be helpful to the public health systems to devise proper policies to better guide and support the pubic.**

Keywords—topic modeling, information visualization, Latent Dirichlet Allocation (LDA), Tweets, COVID-19.

I. Introduction

The World Health Organization (WHO) announced the novel coronavirus (COVID-19) outbreak as a global pandemic on March 11, 2020. This virus is having a significant impact on our human societies. Till now, global confirmed cases of this coronavirus are still continuing to rise rapidly. There have been over 13 million positive cases and more than 570,000 people have lost their lives [1]. It will be useful to understand how people concern and react to the pandemic. Twitter is a social media platform where people share their ideas and express their views in a timely manner. It is also one of the most comprehensive sources of public conversation. There are over 500 million tweets shared on the Internet every day, making it impossible for us to manually analyse the tweets to form a big picture. Designing algorithms to classify Twitter data into topics or categories would be very helpful for us to understand how people react and communicate about this pandemic.

There have been several pioneer research reporting their analysis of Twitter topics during this COVID-19 pandemic. For example, Abd-Alrazaq and colleagues studied the top concerns of Tweeters during the COVID-19 Pandemic, based on the coronavirus related tweets between February 2, 2020 and March 15, 2020 [2]; Ordun and colleagues mined the topics that discuss case spread, healthcare workers, and personal protective equipment (PPE). They also analyzed and visualized the tweeter's retweeting behaviour [3]. Their tweets were collected from March 24 to April 9, 2020.

Such analysis works were conducted at the early stage of this pandemic when many people had not been affected or taken it seriously. However, the COVID-19 situation has developed rapidly. Now the number of people who have lost their lives has already well exceeded the number of confirmed cases at that time. The findings at the early stage of this pandemic may not reflect the current position. It would be very valuable to analyze and draw a big picture from then to now to allow us better understand the impact of this pandemic to the public. Thus we studied the Twitter discussion topics on the latest tweets from March 20 to July 8, 2020, covering the full critical period to date. Further, we analyzed and visualized the dynamics of the topics corresponding to the development of this pandemic.

The paper is organized as follows: First, we introduced in Section II our study method, including data collection, preprocessing and topic modeling; Section III described the findings and the visualized results. In Section IV we discussed and analyzed our results using social psychological theories. The final section concluded the paper by highlighting key findings, gained understanding and interpretation.

II. Methods

A. Data Collection

We used the Coronavirus (COVID-19) Tweets Dataset [4] hosted at IEEE DataPort. This dataset only contains the IDs of geo-tagged English tweets and it does not include any retweets. The dataset is captured by an ongoing project which monitors the real-time Twitter feed for the following keywords: "corona", "coronavirus", "covid", "pandemic", "lockdown", "quarantine", "hand sanitizer", "ppe", "n95", different possible variants of "sarscov2", "nCov", "covid-19", "ncov2019", "2019ncov", "flatten(ing) the curve", "social distancing", "work(ing) from home" and the respective hashtag of all these keywords. We used the data from March 20 to July 8 in this dataset. Due to some accidental technical faults, one day of tweets, i.e. that of March 29, are missing in the dataset. We believe the missing data on this day should have negligible impact on our study. Altogether there are 129,711 tweet IDs within 110 days.

We developed an R program using the *rtweet* package [5] to retrieve tweets based on their IDs in the dataset. In total, 126,476 tweets have been retrieved. After removing all duplicated tweets, there are 126,264 unique tweets.

B. Preprocessing

In this research, our main interests are those discussion topics in people's social media conversation. Therefore, we mainly focused on the tweet text. As the coronavirus pandemic develops, people's concerns also change accordingly. We combined all tweet text from one day into a consolidated text document. The 110 text documents are corresponding to the 110 days of data collection, forming the corpus of this study.

© IEEE 2021. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

DOI 10.1109/IV51561.2020.00129

The pre-processing process is to reduce "noise" in our dataset. We converted the text to lower case first, then performed the following procedures: 1) Removing urls, username, emoji, number, punctuation and white space. 2) Removing the words that do not convey the core meaning of the text. We removed all stopwords that occur often in texts but do not add extra meaning, such as "is", "the" and "but". We also removed "coronavirus", "covid-19", "covid19", "2019-ncov", "2019ncov" because all tweets for this analysis are coronavirus-related. We identified a list of high frequency words that do not convey extra meaning, including: "covid", "corona", "day", "new", "time", "will", "get", "today", "now", "one", "can", "back", "just", "may", "first", "make", "week", "take". We therefore also removed them. 3) Stemming words in the corpus to ensure words that have same meanings or different verb forms of the same word are not duplicated. The cleaned up corpus is the dataset for the topic modeling and other analysis.

C. *Topic Modeling*

Topic modelling is an unsupervised machine learning technique to discover the hidden topics that reflect the meaning of the documents. It has been proven to be a powerful tool for information filtering, text categorization, automatic recommendation, etc. [17]

We used the Latent Dirichlet Allocation (LDA) [13] topic model to identify hidden topics from Tweets. LDA is a generative probabilistic model that leverages Bayesian probability to discover latent (i.e. hidden) topics. It assumes a document to be a bag-of-words with no regard to the order of the words or syntax structures. LDA models all documents as probabilistic combinations of topics. One document may include multiple topics based on certain probabilities, with the total probabilities for all topics of a document being 100%. The topics use words following another probability distribution, and the total probabilities for all words of a topic is 100%.

We developed an R program using the LDA function from the *topicmodels* package [6] to find hidden topics. The LDA function needs the document-term matrix as the input. The document-term matrix describes the word count for each term in each document. As LDA considers documents as "bags of words", the document-term matrix can be a representation of the corpus.

LDA requires the number of topics to be specified beforehand. There are statistical methods (such as the coherence value score used in [3]) that can be adopted as guidance for determining the most appropriate topic numbers. However, these numbers can be too large for human beings to comprehend. In this research, our focus is to investigate people's feelings, concerns and experiences under this pandemic outbreak, thus we decided to use a small topic number to provide a comprehensible overview of the COVID-19 related tweets. We ran our topic modeling algorithm with topic numbers from 3 to 8. After studying the top words for all topics identified by the program and examining sample tweets from each day, we concluded to categorise the tweet discussions as 4 topics.

D. *Emotion Analysis*

To analyze the emotion expressed in the collected tweets, we developed a program using the *syuzhet* R package [14] and compared the tweets with the NRC Emotion lexicon [15,16], which classifies eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust). For each tweet, we calculated the number of occurrences of each emotion. We then summarized the emotions by day and visualized the result to show the evolution of such emotions during the pandemic period. Our result reveals the criticality of government policies in responding to this pandemic.

III. Result and Visualization

In this section we will first report our findings with descriptive statistics, followed by the main topics and emotion analysis of the dataset.

A. *Number of Tweets*

Our dataset contains 126,264 tweets, with an average of 1148 tweets per day. Fig.1 shows the number of tweets changes along with dates. The tweet numbers started to increase from mid-April and reached peak in early May. As we know, the third meeting of the Emergency Committee convened by the WHO Director-General regarding COVID-19 was on 30 April 2020. The Committee unanimously agreed that the outbreak still constitutes a public health emergency of international concern. Around this time, many countries entered into a state of emergency. The COVID-19 related tweet numbers increased significantly. Later, the public got used to the situation and expected the pandemic to be over in June. The tweet numbers also gradually dropped. Until mid-June, the coronavirus situation in many countries around the world, particularly the US, showed significant deterioration instead of any improvement, contradicting to what most people envisaged in early days of this pandemic. The COVID-19 related tweet numbers increased again and remained at a high level since mid-June.

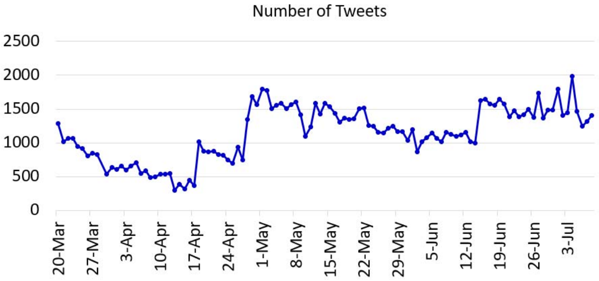


Fig. 1. Number of tweets

B. *Tweets Location*

There are 125,690 tweets with country name identifiable. Out of which, 43.7% of the tweets were from the United States, followed by the United Kingdom and Canada. The top five countries and the total number of tweets are presented in Table I and Fig.2 respectively.

TABLE I. NUMBER OF TWEETS FOR THE TOP FIVE COUNTRIES

Country	No. Tweets	Percentage
United States	54875	43.7%
United Kingdom	18716	14.9%
Canada	9783	7.8%
India	9348	7.4%
Nigeria	4020	3.2%

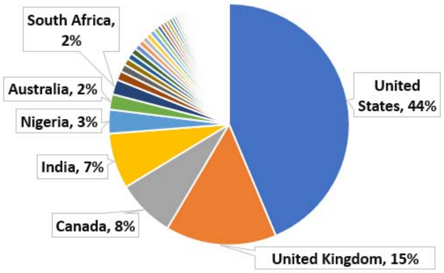


Fig. 2. Nnumber of Tweets by countries

Based on the location coordinates, we plotted the tweets on the map (Fig. 3).

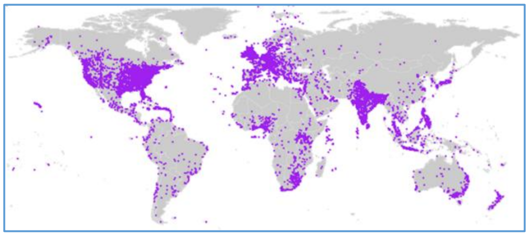


Fig. 3. Tweets on map

C. Most Frequent Words

The top 30 most frequent words of the tweets are summarized in Table II. The top 300 words are visualized as word cloud in Fig. 4. Words are stemmed during the preprocessing. The stemmed words have been changed back to the normal form for easy understanding (e.g. 'happi' was reverted to 'happy') in the table. The first three most frequent words are 'lockdown', 'pandemic' and 'quarantine'. 'Home', 'work' and 'love' are also popular in discussions.

TABLE II. MOST FREQUENT WORDS

	Word	Count		Word	Count
1	lockdown	14684	16	fight	5155
2	pandemic	11476	17	distance	4991
3	quarantine	7795	18	safe	4692
4	home	7740	19	due	4486
5	virus	6813	20	live	4362
6	work	6751	21	stayhome	4350
7	mask	6280	22	see	4160
8	stay	6070	23	socialdistance	4131
9	people	5854	24	life	4103
10	love	5482	25	help	4102
11	like	5466	26	Face	4093
12	social	5338	27	Stigma	3980
13	case	5294	28	Happy	3941
14	health	5231	29	Test	3850
15	thank	5187	30	World	3729

Up to the time of writing this paper, the coronavirus pandemic has already resulted in over 13 million confirmed cases, with more than 570,000 people lost their lives. Social distancing, travel restrictions, work and learn from home, etc. become the new way of life. The pandemic has largely altered the pace, fabric and nature of our lives. It is also causing a global economic recession. There are also lots of uncertainties such as when the vaccine will become available and whether the pandemic can be put under control. The discussion about individual responsibilities such as wearing mask, social distancing and hygiene are not among the most popular words, showing that most governments and public health systems were not very successful in guiding the general public to protect themselves from this extremely infectious disease, contributing to the failure of containing the virus.

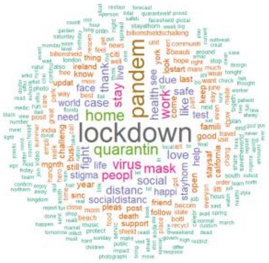


Fig. 4. Word cloud of the top 300 most frequent words

Fear of the situation and uncertainty of changes can leave people feeling stressed, anxious and powerless. Before this study, we expected to see words like "panic", "depressed" or similar words to show people's negative emotions. However, even among the top 50 words, there is none showing fear or stressed about the situation. This is a good indicator which means the majority of people are managing to live along with the pandemic situation. On the other hand, this may also indicate that lots of people have underestimated the severity of the pandemic. By not paying sufficient attention to the pandemic, the public are therefore not seriously and strictly following advice from public health experts.

D. Main Topics

Through our topic modeling program, we discovered 4 topics. The top 20 words for each topic are listed in Table III.

Fig. 5 illustrated the words and the probabilities to the topic they belong to. The words were stemmed by the program thus some words are not in their normal form, e.g. '*happi*' represents all forms of the word '*happy*'.

TABLE III. Top Words for Each Topic

Topic 1	Topic 2	Topic 3	Topic 4
lockdown	pandem	virus	challeng
home	health	stay	face
work	lockdown	quarantin	hero
love	social	stayhom	becom
like	stigma	case	empti
quarantin	fight	see	billionshieldschalleng
socialdistanc	distanc	home	shield
thank	june	everi	bottl
happi	case	india	health
pandem	mask	staysaf	billionshield
safe	test	stayathom	recycl
look	peopl	life	soda
need	see	last	fight
famili	ireland	safe	help
still	busi	fight	savetheworld
come	open	isol	test
keep	last	dead	faceshield
mask	month	forecast	mask
pleas	due	death	peopl
thing	reopen	lock	pandem

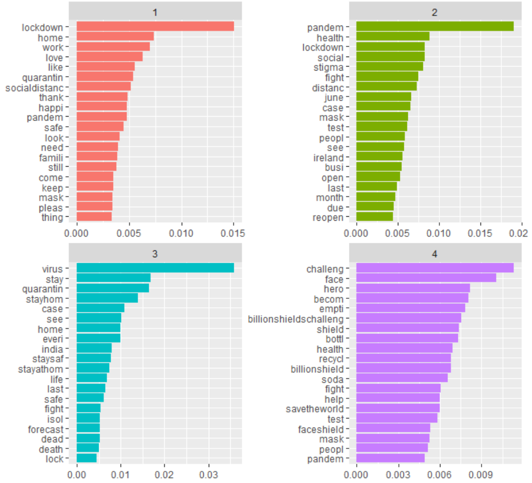


Fig. 5. Words and probabilities in each topic

By analyzing the words in each topic and referring to the original sample tweets, we summarized the meaning of the topics as follows.

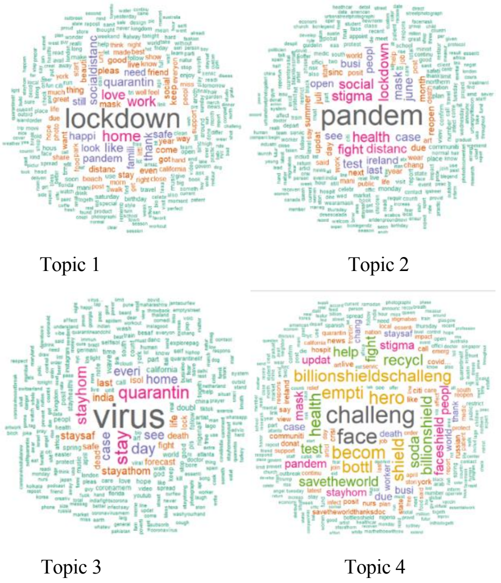
Topic 1: Work and life in the pandemic . In this topic, people discuss their work and life experiences under the lockdown.

Topic 2: Social issues. In this topic, people talk about the social issues (eg. stigma) and political issues.

Topic 3: Understanding the virus. In this topic, people discuss about the COVID-19 cases, death numbers, and the methods of staying home and quarantine to stop transmission.

Topic 4: Prevention methods. In this topic, people discuss methods to prevent getting the virus. Examples include, sharing the method which uses soda bottles to make face shields.

We used word cloud to visualize the top 300 words for each topic, which are presented in Fig. 6.


Topic 1

Topic 2

Topic 3

Topic 4

Fig. 6. Word cloud for the four topics

In each day, the tweets may cover all the 4 topics but with different percentages. We presented the topic coverage in Fig.7, with the x axis showing the date, y axis is the percentage of each topic over the overall discussions on that day.

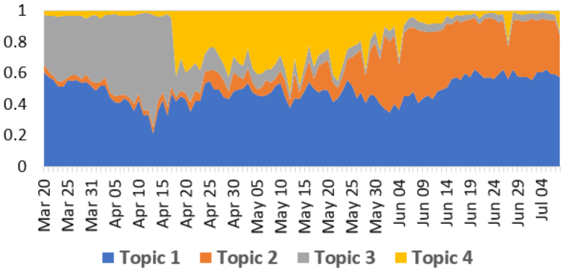


Fig. 7. Topic distribution

Topic 1 always takes about half of the area throughout the whole study period. Work and life in the pandemic is always the most popular topic for discussion. There are more discussions on Topic 3 during March and April, but then gradually becoming less. This can be understood as people tried to understand the virus, and make sense of the situation at the beginning of the pandemic outbreak. Once people attained some knowledge, they tended to focus more on how to prevent themselves from being infected. From mid-April onwards, the discussion on prevention methods (Topic 4) increased a lot. In the first few months of this year, the pandemic caused severe impact on people's lives and the economy. People faced a great uncertainty, not knowing when the pandemic would be over. The long-time frustrating circumstances caused social issues (e.g. social stigma). From June, Topic 2 discussions increased. The relationships of the four main topics are illustrated in Fig.8.

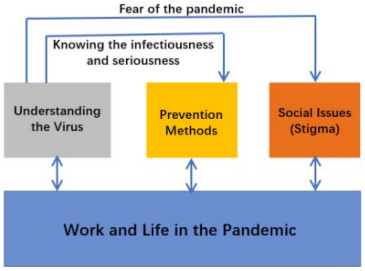


Fig. 8. Links between topics

E. Tweets Emotion

We analysed the emotion of the tweets and extracted the occurrence of each emotion. The emotions ranked from the highest frequency to the least are *trust, anticipation, joy, fear, sadness, anger, surprise and disgust*. The percentage of these emotions are illustrated in Fig.9. The 300 most frequent words by emotion are presented in Fig.10.

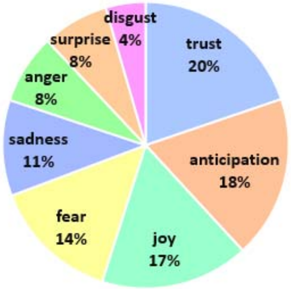


Fig. 9. Emotion percentage among all tweets

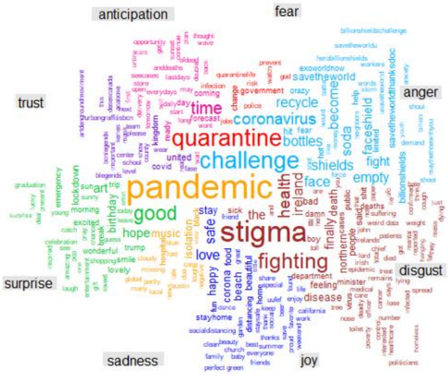


Fig. 10. Top words representing each emotion

The emotion counts are summarized daily in Fig.11 together with the total number of tweets. We can see that the counts of all the eight emotions follow almost the same pattern as the total number of tweets. There are no obvious changes in the eight emotions. We further analyzed the percentage of emotions by days. Again, the percentage of emotions have remained very stable. The results are presented in Fig. 12. It clearly showed that the pandemic did not really have a significant impact on people's emotions. It confirmed our earlier findings that people did not really take the pandemic seriously. Therefore, the public health guidance and strong government campaign are indeed very critical in introducing efficient prevention mechanisms to their citizens.

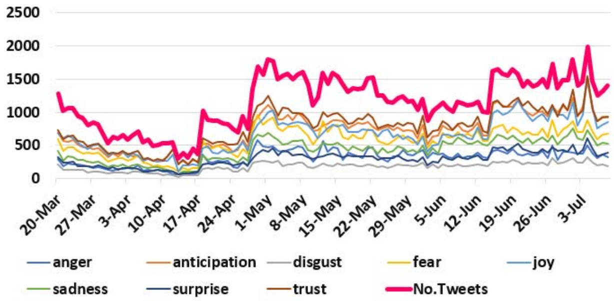


Fig. 11. Emotion count by day

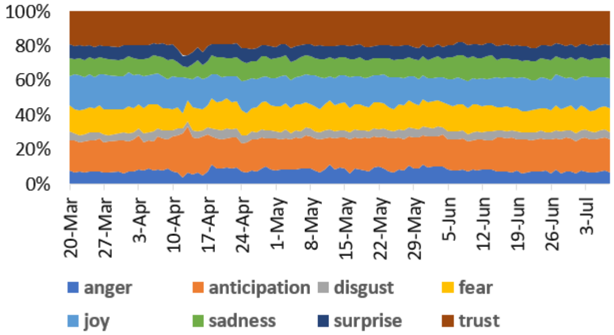


Fig. 12. Emotion percentage by day

IV. Discussion

This study analyzed COVID-19 related tweets to discover the key topics and their evolution along the pandemic. In this section, we use social psychological theories to understand and interpret people's behavior.

A. Health Belief Model

Since the early 1950s, the Health Belief Model (HBM) has been one of the most widely used conceptual frameworks in health behavior research [7]. Over the past decades, the HBM has been expanded and used to support interventions to change health behavior. Fig.13 illustrates the relationships among the main components of the health belief model. According to this model, in order for a behavior to change, people must feel personally vulnerable to a health threat, view the possible consequences as severe, and see that taking action is likely to either prevent or reduce the risk at an acceptable cost with few barriers. In addition, a person must feel competent (have self-efficacy) to execute and maintain the new behaviour.

In March and April, there was increased awareness of the susceptibility and severity of the coronavirus, which showed the increase of perceived threat. Based on the HBM, if the perceived threat of a disease is high, then people are more likely to engage in behaviors associated with prevention techniques, such as wearing mask, keeping social distance or setting up protection shields. Hence, from the middle of April, people started to change their behaviors and engaged in making protection tools by themselves (Topic 4).

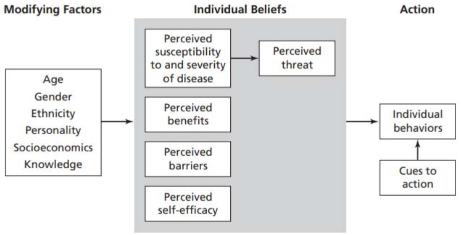


Fig. 13. Health belief model components and linkages [7]

It is advised that the government and media can provide the public with sufficient and accurate information in a timely manner to help people fully understand the susceptibility and severity of COVID-19 and the benefit of following the health advices, therefore people can actively and consciously follow.

B. Social Stigma

Stigma was a Greek word that in its origins referred to a type of signs that were cut or burned into the skin of criminals, slaves, or traitors in order to visibly identify them as blemished or morally polluted persons. These individuals were to be avoided particularly in public places [8]. Nowadays, stigma is understood to mean a social construction whereby a distinguishing mark of social disgrace is attached to others in order to identify and to devalue them. It will result in stereotypes, prejudice and discrimination [9]. Social stigma in an pandemic outbreak may mean people are being labelled, stereotyped, discriminated against, treated separately, and/or experiencing loss of status because of a perceived link with a disease [11].

Stangor and Crandall [10] proposed that stigma develops out of an initial, universally held motivation to avoid danger, followed by (often exaggerated) perception of characteristics that promote threat, and accompanied by a social sharing of these perceptions with others. In the current coronavirus outbreak, stigma is associated with a lack of knowledge about how COVID-19 spreads, a need to blame someone, fears about disease and death, and gossip that spreads rumors and myths [12].

Since March, the pandemic has led to record high numbers of illness and fatalities. It has changed the society and heavily affected the economy. This is a novel virus and there are still many unknowns. The crisis and uncertainties create fear among the public. This fear creates the need to blame “others” and scapegoating. Thus, people of certain ethnic backgrounds as well as anyone perceived to have been in contact with the virus became the “others”. From June, the discussion about social issues (Topic 2), such as stigma, increased considerably in tweets.

V. CONCLUSION

This paper reported our analysis of COVID-19 related tweets from the end of March to early July this year, in terms of the number, location, frequent words, main topics and emotion. From the analysis and visualization, we found the key topics and emotions that people are actively discussing in their tweets. We also analyzed their evolution along with this pandemic, as well as any potential reasons and impacts. Our findings provide insights and opportunities for public health systems to devise strategies which will help the public control transmission of the virus and reduce social issues.

REFERENCES

[1] WHO COVID-19 Dashboard, https://covid19.who.int/.

[2] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and S. Shah, "Top concerns of Tweeters during the COVID-19 pandemic: infoveillance study", J Med Internet Res 2020, 22(4):e19016. https://www.jmir.org/2020/4/e19016.

[3] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of Covid-19 Tweets using topic modeling, UMAP, and DiGraphs", arXiv:2005.03082v1.

[4] R. Lamsal, "Coronavirus (COVID-19) geo-tagged Tweets dataset", IEEE Dataport, 2020. [Online]. Available: http://dx.doi.org/10.21227/fpsb-jz61. Accessed: Jul. 12, 2020.

[5] M.W. Kearney, “rtweet: Collecting and analyzing Twitter data”, *Journal of Open Source Software*, 4(42), 1829. doi: 10.21105/joss.01829, R package version 0.7.0, https://joss.theoj.org/papers/10.21105/joss.01829. 2019

[6] B. Grün, K. Hornik, “topicmodels: An R package for fitting topic models”, *Journal of Statistical Software*, 40(13), pp:1–30, 2011. doi: 10.18637/jss.v040.i13.

[7] V.L. Champion, C.S. Skinner, “The health belief model” in Health behavior and health education: theory, research, and practice, 4th ed., K. Glanz, B.K. Rimer and K. Viswanath Eds. San Francisco, USA: John Wiley & Sons, 2008, pp:45–62.

[8] E. Goffman, Stigma: Notes on the Management of Spoiled Identity, New York: Simon and Schuster, 2009.

[9] J. Arboleda-Florez, “What causes stigma?”, World Psychiatry, vol. 1(1), pp:25–26, 2002.

[10] C. Stangor and C.S. Crandall, “Threat and the social construction of stigma” in The Social Psychology of Stigma, T.F. Heatherton, R.E. Kleck, M.R. Hebl, and J.G. Hull Eds. Guilford Press, 2000, pp: 62–87.

[11] “A guide to preventing and addressing social stigma associated with COVID-19”, COVID-19: Risk Communication and Community Engagement, 24 February 2020, World Health Organization, https://www.who.int/publications/m/item/a-guide-to-preventing-and-addressing-social-stigma-associated-with-covid-19

[12] ”Reducing stigma”, Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/reducing-stigma.html

[13] D.M.Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, Vol.3, pp:993-1022, 2003.

[14] M. Jockers, "syuzhet: Extracts sentiment and sentiment-derived plot arcs from text", 2017. https://CRAN.R-project.org/package=syuzhet

[15] NRC Emotion lexicon, http://www.purl.org/net/NRCemotionlexicon

[16] S.M. Mohammad, S. Kiritchenko and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of Tweets", 2013, arXiv:1308.6242.

[17] B.V. Barde and A.M. Bainwad, "An overview of topic modeling methods and tools," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp:745-750, doi: 10.1109/ICCONS.2017.8250563.