Understanding Public Sentiments, Opinions and Topics about COVID-19 using Twitter

Jolin Shaynn-Ly Kwan
Information Systems Technology and Design Pillar
Singapore University of Technology and Design
Email: jolin_kwan@mymail.sutd.edu.sg

Kwan Hui Lim
Information Systems Technology and Design Pillar
Singapore University of Technology and Design
Email: kwanhui_lim@sutd.edu.sg

Abstract—The COVID-19 pandemic has caused widespread devastation throughout the world. In addition to the health and economical impacts, there is an enormous emotional toll associated with the constant stress of daily life with the numerous restrictions in place to combat the pandemic. To better understand the impact of COVID-19, we proposed a framework that utilizes public tweets to derive the sentiments, emotions and discussion topics of the general public in various regions and across multiple timeframes. Using this framework, we study and discuss various research questions relating to COVID-19, namely: (i) how sentiments/emotions change during the pandemic? (ii) how sentiments/emotions change in relation to global events? and (iii) what are the common topics discussed during the pandemic?

I. Introduction

The COVID-19 outbreak is a global pandemic that has infected millions and claimed the lives of hundred thousands. Social media, e.g., Twitter, has proved to be valuable in times of such global pandemics, as it reveals the real-time sentiment and discussions of Twitter users [1]. Additionally, with the implementation of quarantine measures and the prohibition of social gatherings across the world, the usage of social media has soared to extraordinary levels.

Towards this effort, we aim to answer three research questions through analysis of global tweets about COVID-19 in the duration after the virus was named as a pandemic. The research questions (RQ) are:

- RQ1: What are the sentiments and emotions experienced during the COVID-19 pandemic?
- RQ2: How do these sentiments and emotions change in relation to global events?
- RQ3: What are the common topics discussed during the COVID-19 pandemic?

Related Work. Social media data have been a popular source of information for studying pandemics and diseases [2]. For example, Signorini et al. studied the use of Twitter to track rapidly-evolving public interest with respect to the 2009 H1N1 pandemic in the United States [3]. The study found that interest in hand hygiene and face masks seemed to be timed with public health messages from the Centers for Disease Control and Prevention (CDC). Chew studied public sentiment with regards to the 2009 H1N1 pandemic from manual content analysis of tweets [4]. Keywords corresponding to certain sentiments were collated and the tweets were categorised into the different sentiments by string detection of the different keywords.

Other researchers worked on different aspects of studying the COVID-19 pandemic. Dubey [5] studied the sentiments of COVID-19 related tweets between 11-31 March 2020, while Rajput et al. [6] used word frequency analysis to identify trends in the words used. Similarly, Lwin et al. [7] studied the prevalence of the four emotions of fear, anger, sadness, and joy in relation to the COVID-19 pandemic. Abd-Alrazaq et al. [8] focused on understanding topics related to COVID-19 from 2 February to 15 March 2020 and [9] used mobility patterns within and out of Wuhan, China to understand COVID-19 cases.

Contribution. While there has been significant research with regards to pandemics, the use of Twitter as a research tool for this is still relatively new and focus on one aspect of the research. In contrast, our work adopts a holistic framework that combines sentiment analysis, emotion detection, topic modelling, spatial/temporal analysis to understand the range of issues surrounding COVID-19. For example, we combine sentiment analysis with topic modelling to provide a deeper understanding about how individuals feel about specific topics, which can aid public health officials in developing strategies and policies.

II. Research Framework

In this section, we discuss the framework used to utilize public tweets for understanding sentiments, emotions and discussion topics related to COVID-19.

A. Data Collection

Our dataset comprises 910,000 public tweets from users across the world, curated from a combination of three larger COVID-19 related datasets. We extracted 70,000 tweets per week that were more likely to contain geolocation information (more details later), for the period of 12 March to 10 June 2020. Out of the 910,000 tweets, there are a total of 896,031 geolocated tweets in English. We extracted this subset of tweets from three larger datasets, namely: (i) Shane Smith's datasets on Kaggle [10], where daily datasets of tweets containing hashtags surrounding the topic of COVID-19 from 12
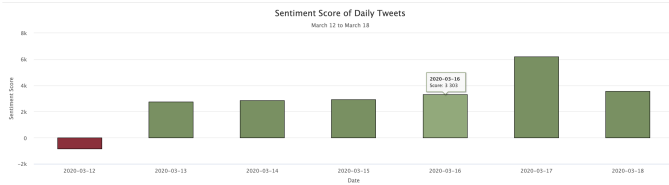
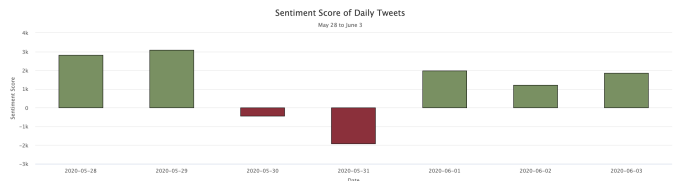Fig. 1. Time Series Plot for Sentiment Score of Daily Tweets

Fig. 2. Sentiment Score of Daily Tweets for Week of 28 May to 3 June

March to 30 April 2020; (ii) The Github repository of [11] contained tweet IDs of collected tweets from 1 May to 22 May 2020; and (iii) our own data collection from 23 May 2020 onwards.

B. Data Preprocessing

Previous studies have estimated that only 0.85% of tweets are GPS geotagged [12], [13]. Likewise for the original larger datasets, the proportion of tweets that are geotagged is extremely small. To address this issue, we utilized other attributes of the tweets to infer the location of these tweets, such as: (i) the self-declared location in the user profiles; and (ii) the "entities" section of the tweets which contain metadata and additional contextual information about them. This resulted in the significantly higher number of 896,031 geolocated tweets that will be used for spatial analysis. Thereafter, we perform the standard text preprocessing steps such as conversion to lowercase, tokenizing, lemmatizing and stemming.

C. Sentiment/Emotion Analysis

The process of sentiment/emotion analysis involves taking the pool of words that have been identified from the preprocessing steps and comparing them to lexicons to determine their sentiments or polarity [14]. There are two main lexicons used in this study: the NRC Word-Emotion Association Lexicon (also known as EmoLex) and the AFINN lexicon. The NRC lexicon is a list of more than 14,000 English unigrams and their associations with eight fundamental emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive) [15]. The assignment and labelling of a word to its sentiment were manually done by crowdsourcing through Amazon's Mechanical Turk service. The eight fundamental emotions in the lexicon correspond to the eight primary emotions introduced by Plutchik [16].

The AFINN lexicon, on the other hand, assigns English words on a scale of -5 to 5 (excluding 0), with negative scores representing negative sentiment and positive scores representing positive sentiment [17]. The sentiment score for each tweet is calculated as a sum of the sentiment scores for words in the tweet. A net positive overall sentiment score indicates a positive statement, a net negative overall sentiment score indicates a negative statement and a net overall sentiment score of zero indicates a neutral statement.

Therefore, two aspects of sentiment analysis were carried out for this research. The NRC lexicon was used for the detection of the different emotions that Twitter users have towards the pandemic, while the AFINN lexicon was used for detecting the extent of negativity/positivity of the tweets regarding the pandemic.

D. Topic Modelling

Latent dirichlet allocation (LDA), a topic modelling algorithm, was carried out to identify key topics talked about in tweets surrounding the COVID-19 pandemic. LDA is a generative probabilistic model of a corpus [18]. The algorithm works by representing documents as random mixtures over latent topics, where each topic is characterized by a collection of words ("bag-of-words") [19].

The tweets were grouped into weeks (7 days) with a total of 13 weeks and the tweet pool of each individual week is treated as an individual document. To identify the optimal number of topics for LDA on the 13 different documents, the measure of perplexity was used. Perplexity is a statistical measure of how well a probability model predicts a given sample. In the context of LDA, the theoretical word distribution represented by the given number of topics is compared to the actual topic mixtures or distribution of words in the documents [20].

III. FINDINGS AND DISCUSSION

A. RQ1: Is it all doom and gloom?

We observe that the emotion "trust" surprisingly makes up a consistently higher proportion of tweets as compared to other emotions. Additionally, the emotion "disgust" makes up for a consistently lower proportion of tweets as compared to others. These results are surprising because a pandemic is usually associated with devastation and it would be natural to assume a greater prevalence of negative emotions. These negative emotions have been reported to stem from areas such as the fear and anxiety people feel in the face of a novel virus, the feelings of isolation and loneliness from social distancing [21], as well as the stress and worry from economic uncertainties [22]. However, a study by Barford et al. found that negative emotions rarely overwhelm positive ones in everyday circumstances [23], and the findings from our research displays the same phenomenon even in the midst of the COVID-19 pandemic. A study surveyed 854 Australian residents about their emotions specifically during the COVID-19 pandemic in March 2020 and the results showed that only 3% of the participants felt purely negative emotions, but almost all the participants actually reported feeling positive emotions at some point in time. These positive emotions were attributed to factors such as the novelty and flexibility of different working arrangements and also feelings of approval
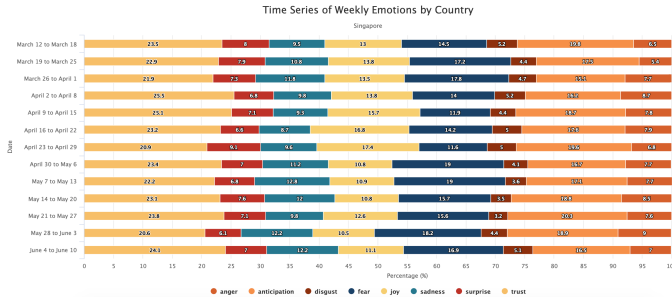
Fig. 3. Time Series of Weekly Emotions in Singapore

Fig. 4. Time Series of Weekly Emotions in the United Kingdom

towards social distancing for the sake of everyone's collective health [24].

B. RQ2: Global Events and Sentiments

The first global event that will be covered is the declaration of COVID-19 as a pandemic by WHO Director-General on 11 March 2020. This came after the number of cases COVID-19 outside China increased by 13 folds and the number of affected countries tripled in just two weeks [25]. This study started the analysis of tweets on 12 March, the day after the declaration and it can be seen from the visualization of daily sentiment score (see Figure 1) that people's sentiments were significantly more negative on 12 March as compared to other days within the week as well as in the next few weeks. This is likely due to the fact that anxiety and fear levels were heightened after people realised the gravity of the COVID-19 situation and how many aspects of their lives will be impacted – be it in daily routines or in finances.

The second global event is the protests sparked by the death of George Floyd on 25 May 2020 where a police officer killed him by kneeling on his neck during an arrest. The Black Lives Matter movement kick-started many protests not just in the USA, but all over the world as well. The weekend of May 30 and May 31 saw protests all over the world in countries like Britain, Germany, France and Spain, where people stood "in a global act of solidarity" [26]. To add on to the anger and devastation that people had towards this issue, the President of the USA, Donald Trump tweeted controversial and insensitive comments in response to the protests. His tweets on 29 May and 30 May threatened violence against the protesters outside the White House with one tweet saying that they could have been attacked with "vicious dogs and ominous weapons" by the US Secret Service [27]. The culmination of the protests and the comments by Trump fuelled people's rage towards this situation and this can be seen in the significantly more negative sentiment scores for 30 May and 31 May (see Figure 2).

Figure 3 shows the time series of the proportions of emotions exhibited by Twitter users in Singapore with regards to COVID-19. The week after the circuit breaker[1] started for Singapore on 7 April (April 9 to April 15), it can be seen that percentages of the emotions "fear" (11.9%) and "sadness" (9.3%) dropped to one of the lowest percentages for

[1]Measures to reduce movement, activity and interactions in the country.
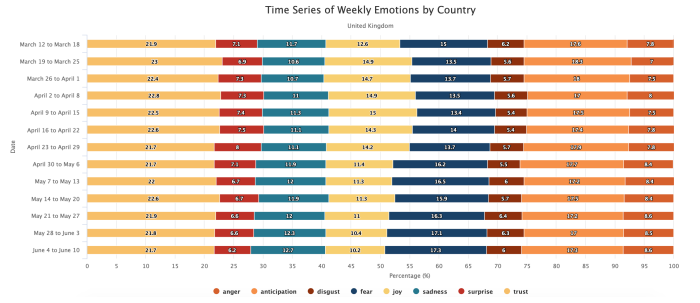
their respective emotions throughout the whole duration of the dataset; while the percentages of the emotions "trust" (25.1%) and "joy" (15.7%) rose to one of the highest percentages for their respective emotions. From these results, it can be seen that contrary to popular belief that being in a "lockdown" dampens moods with the prohibition of social gatherings as well as harsh consequences associated with breaching the associated rules, Singaporeans generally still do feel a sense of relief and satisfaction at the implementation of the circuit breaker. According to a recent analysis of Singaporeans' sentiment online, positive sentiments could be attributed to the Singapore government's initiatives such as the Solidarity Budget that offers help to Singaporeans' livelihoods while coping with the economic effects of the pandemic, and also the sense of relief at stricter control measures in light of growing cases of infection [28].

Figure 4 shows the time series of the proportions of emotions exhibited by Twitter users in the United Kingdom with regards to COVID-19. Looking at the percentages of the emotions across the weeks, it can be seen that the percentages for the emotions "fear" (16.2%) and "sadness" (11.9%) saw significant increases while emotions like "joy" saw a significant decrease (11.4%) starting from the week of April 30 to May 6. During the first coronavirus press conference since being infected with COVID-19 on 30 April, the Prime Minister of the United Kingdom (UK) , Boris Johnson, announced that the UK was "past the peak" of the virus outbreak and that a "road map" would be laid out to phase out the lockdown in the UK [29]. The "road map" was eventually revealed on 10 May where Johnson announced a phased end to existing lockdown restrictions. However, many criticisms were made against the way Johnson handled the reopening – that there was a lack of clear plans and guidance for safety [30] at a critical time in managing the pandemic [31]. In addition, the change of message for the UK's coronavirus safety campaign from "stay home" to "stay alert" has also been criticized as "confused" and "nonsensical" by the public and rejected by nations such as Scotland and Wales [32]. The feelings of uncertainty and anxiety felt by people in the UK during this period is reflected in how the overall levels of negative emotions increased significantly from the week of April 30 onwards, while the overall levels of positive emotions decreased.

C. RQ3: Discussion Topics

The analysis from longitudinal topic modelling for 3 months of tweets related to COVID-19 has proven to be effective in identifying key trends of public dialogue and could potentially be used to guide targeted and timely interventions. For example, the overall trend of topics started off containing words with negative connotations such as "violence", "assault", "marginalize", with the topics suggesting the woes and frustrations associated with the COVID-19 lockdown/quarantine. However, as time went by, the topics started shifting towards the more positive end of the spectrum, with topics that involved tweets about keeping productive during the lockdown and tweets about approaching a "new normal" with telecommuting and e-commerce. This signals a potential "glimmer of hope" amidst the bleak outlook because it shows that society is accelerating the use of technology, progressing forward in becoming "smart" cities, and adapting to the new future.

IV. CONCLUSION

In this paper, we present a framework that allows the general public or public health officials, to understand the public sentiment towards the pandemic, the key topics discussed. For example, with an idea of what the key topics are at a specific time in point, as well as across time, the general public can be educated about events that might not have been reported that widely in their specific countries. Using this framework, we studied three research questions relating to COVID-19, namely: (i) how sentiments/emotions change during the pandemic? (ii) how sentiments/emotions change in relation to global events? and (iii) what are the common topics discussed during the pandemic?. We believe that this research's methodology can be easily tweaked and applied to different contexts as well to examine other research topics such as that of a natural disaster or political events.

REFERENCES

[1] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, and T. Zhu, "Twitter discussions and concerns about covid-19 pandemic: Twitter data analysis using a machine learning approach," *arXiv preprint arXiv:2005.12830*, 2020.

[2] J. Liu, T. Singhal, L. T. Blessing, K. L. Wood, and K. H. Lim, "EPIC30M: An Epidemics Corpus of Over 30 Million Relevant Tweets," in *Proceedings of the 2020 IEEE International Conference on Big Data (BigData'20)*, Dec 2020.

[3] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[4] C. Chew and G. Eysenbach, "Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak," *PloS one*, vol. 5, no. 11, p. e14118, 2010.

[5] A. D. Dubey, "Twitter sentiment analysis during covid19 outbreak," *Available at SSRN 3572023*, 2020.

[6] N. K. Rajput, B. A. Grover, and V. K. Rathi, "Word frequency and sentiment analysis of twitter messages during coronavirus pandemic," *arXiv preprint arXiv:2004.03925*, 2020.

[7] M. O. Lwin, J. Lu, A. Sheldenkar, P. J. Schulz, W. Shin, R. Gupta, and Y. Yang, "Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends," *JMIR public health and surveillance*, vol. 6, no. 2, p. e19447, 2020.

[8] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the covid-19 pandemic: infoveillance study," *Journal of medical Internet research*, vol. 22, no. 4, p. e19016, 2020.

[9] D. Bisanzio, M. U. Kraemer, I. I. Bogoch, T. Brewer, J. S. Brownstein, and R. Reithinger, "Use of twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of covid-19 at global scale," *Geospatial health*, vol. 15, no. 1, 2020.

[10] S. Smith, "Coronavirus (covid19) tweets - early april, version 4," https://www.kaggle.com/smid80/coronavirus-covid19-tweets-early-april/version/4, 2020.

[11] Rrishik, "Covid-19-tweets," https://github.com/Rrishik/COVID-19-tweets, 2020.

[12] L. Sloan and J. Morgan, "Who tweets with their location? understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter," *PloS one*, vol. 10, no. 11, p. e0142209, 2015.

[13] L. Chi, K. H. Lim, N. Alam, and C. J. Butler, "Geolocation Prediction in Twitter Using Location Indicative Words and Textual Features," in *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT'16)*, Dec 2016, pp. 227–234.

[14] K. H. Lim, K. E. Lee, D. Kendal, L. Rashidi, E. Naghizade, Y. Feng, and J. Wang, "Understanding Sentiments and Activities in Green Spaces using a Social Data-driven Approach," in *Smart Cities: Issues and Challenges*. Elsevier, Jun 2019, pp. 77–107.

[15] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[16] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[17] University of Cincinnati, "Text mining: Sentiment analysis," https://uc-r.github.io/sentiment_analysis, 2018.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[19] C. Montenegro, C. Ligutom III, J. V. Orio, and D. A. M. Ramacho, "Using latent dirichlet allocation for topic modeling and document clustering of dumaguete city twitter dataset," in *Proceedings of the 2018 International Conference on Computing and Data Engineering*, 2018, pp. 1–5.

[20] University of Cincinnati, "Topic modeling. computing for the social sciences," 2020.

[21] Centers for Disease Control and Prevention, "Coping with stress: Pandemics can be stressful," 2020.

[22] L. Bareket-Bojmel, G. Shahar, and M. Margalit, "Covid-19-related economic anxiety is as high as health anxiety: Findings from the usa, the uk, and israel," *International Journal of Cognitive Therapy*, p. 1, 2020.

[23] K. A. Barford, P. Koval, P. Kuppens, and L. D. Smillie, "When good feelings turn mixed: Affective dynamics and big five trait predictors of mixed emotions in daily life," *European Journal of Personality*, 2020.

[24] J. Anglim, K. Barford, P. O'Connor, and L. Smillie, "Not all doom and gloom: even in a pandemic, mixed emotions are more common than negative ones," *The Conversation*, 2020.

[25] World Health Organization, "Who director-general's opening remarks at the media briefing on covid-19 (11 march 2020)," 2020.

[26] N. Colarassi, "Photos show how the world is demanding justice for george floyd at global black lives matter protests," *The Insider*, 2020.

[27] O. Milman, M. Pengelly, R. Luscombe, and D. Smith, "George floyd: as protests rage across us, trump comments inflame tensions," *The Guardian*, 2020.

[28] N. Syed, "Circuit breaker: How are singaporeans affected?" *Human Resource Director*, 2020.

[29] N. Triggle, "Coronavirus: Boris johnson says uk is past the peak of outbreak," *BBC*, 2020.

[30] A. Bienkov and T. Colson, "Boris johnson has revealed his 'road map' for ending the uk coronavirus lockdown," *Business Insider*, 2020.

[31] A. Forsyth, "Coronavirus: Government urges 'caution' on lockdown easing," *BBC*, 2020.

[32] L. Handley, "Uk pm boris johnson's new 'stay alert' coronavirus warning criticized as 'confused' and 'nonsensical'," *CNBC*, 2020.