Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220552

Discovery of COVID-19 Symptomatic Experience Reported by Twitter Users

Keyuan JIANG[a,1], Minghao ZHU[b] and Gordon R. BERNARD[c]
[a] *Purdue University Northwest, Hammond, Indiana, U.S.A.*
[b] *College of Electronic & Information Engineering, Tongji University, Shanghai, China*
[c] *Vanderbilt University, Nashville, Tennessee, U.S.A.*

Abstract. Since the beginning of the COVID-19 pandemic, patients shared their personal experiences of the viral infection on social media. Gathering their symptomatic experiences reported on Twitter may help better understand the infectious disease and supplement our knowledge of the disease gathered by healthcare workers. In this study, we identified personal experience tweets related to COVID-19 infection using a pre-trained and fine-tuned language model, and annotated the machine-identified tweets in order to extract the information of infection status, symptom concepts, and the days the symptomatic experience occurred. Our result shows that the top 10 most common symptoms mentioned in the collected Twitter data are in line with those published by WHO and CDC. The symptoms along with the day information appear to provide additional insight on how the infection progresses in infected individuals.

Keywords. Novel coronavirus, COVID-19 symptoms, personal health experience, Transformer-based language model, Twitter data

1. Introduction

The outbreak of the novel coronavirus, which occurred in January 2020, has led to the pandemic of COVID-19, declared by the World Health Organization (WHO) on March 11, 2020. Since the outbreak, this highly contagious, life-threatening virus has infected more than 433 million people and caused over 5.9 million deaths around the world, according to WHO [1]. Even with the successful development and wide availability of various COVID-19 vaccines, the world is still struggling to contain the pandemic.

At the onset of the breakout, it became challenging for healthcare workers to accurately diagnose the disease due to the limited understanding and knowledge of the new disease and the similarity of its symptoms to those of flu and common cold. Healthcare givers were learning about the disease while caring and treating the patients, to gain the knowledge about the disease and to discover effective treatments. COVID-19 symptoms were reported by caregivers in the early days of the pandemic [2-4].

Meanwhile, COVID-19 symptoms were shared on social media by those who experienced the viral infection. Twitter posts have received noticeable attention since the beginning of the pandemic and efforts were made to leverage the Twitter data to understand the new outbreak. Chen and colleagues [5] started collecting COVID-19

[1] Corresponding Author, Keyuan Jiang, Department of Computer Information Technology and Graphics, Purdue University Northwest, Hammond, Indiana 46323, U.S.A; E-mail: kjiang@pnw.edu.

related tweets on 28 January 2020. The collection has continued since and is ongoing. Tweet IDs of more than 2 billion tweets collected are publicly shared on GitHub[2]. To make the Twitter data useful for various natural language processing tasks, Müller and colleagues (2020) generated a transformer-based language model, named COVID-Twitter BERT (CT-BERT), based on Google's BERT (Bidirectional Encoder Representations from Transformers) language model [6]. The language model was learned with a corpus of 22.5 million unannotated COVID-19 related tweets. Sarker et al. [7] attempted to investigate self-reported COVID-19 symptoms from Twitter posts, by manually annotating nearly 500,000 tweets. Their study showed that 203 positive-tested Twitter users reported 1,002 symptoms with 668 unique expression.

COVID-19 experiences posted by Twitter users may provide richer information and more details. For example, "*So I tested positive for Covid19 yesterday...[omitted]. Day 1. My Symptoms:. . Throbbing headache with pressure right behind the eyes. Fever ranging from 102-104*" describes (1) whether the infection was confirmed (*tested positive*), (2) the symptoms (*headache, eye pressure,* and *fever*) and (3) when the symptoms were experienced (*Day 1*). Collecting this type of the information can help enhance our knowledge and understanding of the deadly infectious disease. In this study, we attempt to gather such information from the Twitter data, with machine learning methods and manual annotation to discover the mentions of symptoms along with the day information.

2. Method

Our data processing pipeline starts with collecting COVID-19 related Twitter data which are known for noisiness and informal writings. The collected tweets were preprocessed to remove duplicates, retweets (RTs), and non-English tweets. Tweets pertaining to personal experience were identified from the preprocessed tweets using a method based on the fine-tuning of the pre-trained RoBERTa (Robustly Optimized BERT Pretraining Approach developed by FaceBook AI) language model [8]. The personal experience tweets were later processed by MetaMap Lite [9] to identify any potential symptom terms. Afterwards, tweets with any identified symptoms were annotated manually to extract the infection status, day information and symptoms which were missed by MetaMap Lite.

2.1. Identifying Personal Experience Tweets

To predict personal experience tweets (PETs), we utilized our pre-trained and fine-tuned language model based on RoBERTa (Robustly Optimized BERT Pretraining Approach [10]) for identifying PETs related to medication effects [8]. The motivation for this transfer learning is twofold: (1) both problems are somehow similar in the same domain – all about the personal experience in health, and (2) the annotated PETs for COVID-19 symptoms were not available and would take a significant amount of effort to do so.

The RoBERTa-based language model, with a structure of 12 layers, 768 hidden neurons, 12 self-attention heads and 110M parameters, was initially pre-trained with more than 160GB uncompressed texts [10]. The model was further fine-tuned with 12K annotated tweets related to personal experience of medication effects. With the medication effect tweets, the model achieved 0.877 for accuracy, 0.734 for precision,

[2] https://github.com/echen102/COVID-19-TweetIDs

0.775 for recall, and 0.754 for F1 score. The same model was transferred, without relearning/re-training with the COVID-19 data, to identify personal experience tweets related to disease (COIVD-19) symptoms.

2.2. Data

A corpus of 12 million tweets was collected in May 2020 by querying Twitter.com with a home-made crawler which was implemented in compliance with the crawling policy of Twitter.com (as documented in its robots.txt file). The tweets posted between March 11, 2020 and April 23, 2020 were gathered. The keywords used in querying the tweets are: *covid19*, *COVID-19*, *coronavirus*, *Wuhan pneumonia*, and *nCoV*. This corpus of raw tweets was preprocessed, and non-personal experience tweets were filtered out using a pre-trained transformer-based method developed by our team [8]. Afterwards, the personal experience tweets were processed by MetaMap Lite through its RESTFul API. Concepts extracted by MetaMap Lite were considered symptoms if they belong to the semantic type of *sosy* (sign and symptom). Tweets containing no symptom concepts were discarded, and this process yielded about 11K tweets.

Symptomatic experiences occurred on the first 14 days were of interest in this study, as COVID-19 symptoms develop between 1 and 14 days after exposure [13]. Tweets containing day information of experience were extracted using a set of phrases. For instance, for day 1, the following phrases were used: *day 1*, *day1*, *day one*, *1st day*, and *first day*, to cover various ways expressing the first day from the reference point. However, it is noted that this also resulted in tweets containing any days starting with 1, such as day 18. Finally, a corpus of 699 tweets was identified for manual annotation.

2.3. Data Annotation

After gathering the personal experience tweets, we annotated them to extract information of the infection status, symptoms and the days of symptomatic experience. There were two steps of annotation. The first step was to help set the reference points, start points of infection. The labels listed in Table 1 were assigned to tweets during the first step.

Table 1. Labels for the confirmation/status of COVID-19 infection. This information was used to set the reference point.

Label	Description	Tweet Count
Isolation	Isolation, quarantine	93
Confirmed	Tested Positive	15
Infected	Highly certain without testing	372
Suspected	Not sure, without testing	173
Not Personal	Not a personal experience tweet	2
Unrelated	Not related to COVID-19	33
Lockdown	During a lockdown	8
WFH	Work from home	1
Negative	Tested negative	2

The reference point was defined as the day zero (0) of infection. We decided to combine tweets with labels of *isolation*, *confirmed*, *infected* and *suspected* so that their reference points (start points) are the same. The rational was that (1) there was a lack of testing at the time (March and April of 2020), partially due to an increasing demand for testing and lack of available testing kits, (2) those in isolation include individuals who

showed symptoms or had been exposed to the infected individuals, and (3) all of the tweets contain mentions of symptoms.

The second annotation step was to extract symptom expressions, which can be a single word (e.g., *breathlessness*) or made up of multiple words (e.g., *temp at 7 EST is 100.6*). Listed in Table 2 are the top 10 most common symptoms in our data set after the second step of annotation. As shown in the table, there are multiple ways to express each symptom concept. In particular, each of the three symptom concepts listed (Fever, Breath Difficulty, and Lost Smell/Taste) shows more than 60 different expressions.

Table 2. Top 10 most common COVID-19 symptoms. Unique Expression Count: the number of unique expressions for a symptom concept. Count of Mention: the number of mentions of a symptom concept. A single tweet may contain more one mention of the same symptom concept.

Symptom	Unique Expression Count	Count of Mention
Fever	64	314
Cough	23	315
Headache	20	153
Fatigue	10	150
Ache	14	102
Breath Difficulty	64	126
Lost Smell/Taste	67	74
Sore Throat	14	69
Tight Chest	35	56
Chest Pain/Discomfort	25	49

It is also noted that Twitter users use many layman's terms or consumer health vocabulary (CHV) terms to express the symptom concepts. For example, *temp* and *temperature* were commonly used to describe the concept of fever.

3. Result and Discussions

Table 3 below is the result of aligning the mentions of the top 10 most common symptoms with the day information as reported in the Twitter posts. The figure in each cell is the number of mentions of the corresponding symptom.

Table 3. Mentions of top 10 most common COVID-19 symptoms by day. Chest pain symptoms include other chest discomforts other than tight chest.

	Day													
Symptoms	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Fever	21	22	30	29	29	13	29	15	17	21	15	11	3	12
Cough	19	26	33	31	36	29	34	17	18	26	15	10	3	16
Headache	10	19	23	13	20	12	21	11	6	16	8	2	6	4
Fatigue	3	10	19	22	19	6	15	16	11	11	6	1	6	1
Ache	11	21	17	10	8	9	6	4	3	7	2	3	0	2
Breath Difficulty	7	14	16	16	16	8	15	6	9	11	11	3	0	3
Lost Smell/Taste	2	2	4	5	15	5	6	3	4	4	1	3	2	2
Sore Throat	15	11	12	5	7	9	5	4	4	0	2	0	1	0
Tight Chest	3	10	6	6	6	1	7	1	0	4	0	0	0	1
Chest Pain	1	3	4	0	6	3	3	4	2	3	0	0	1	0

It is worth noting that the most common symptoms identified from the study Twitter data are well in line with those published by WHO [11] and CDC of the United States [12]. Besides, the symptomatic experiences reported on Twitter provide additional useful information such as the day of a particular symptom and its severity.

The number of mentioning a particular symptom on a particular day may indicate the likelihood that the symptom would be experienced on that day. This information may help us understand how the infection progresses on individuals. It is known that COVID-19 infection starts at the upper respiratory track, and it is interesting to note that the number of mentioning sore throat is the highest on day 1 during the 14 day period, and the number of mentions of other symptoms (fever, cough, headache, fatigue, ache, breath difficulty, and tight chest) increases from day 1. More people reported loss of smell/taste on day 5 than any other days.

4. Conclusion

In this study, we investigated utilizing a pre-trained and fine-turned language model to identify personal experience tweets related to COVID-19 infection. These personal experience tweets were manually annotated for symptoms. Our result shows that the top 10 common symptoms are in line with those reported by both WHO and CDC, and experiences of infected individuals shared online provide additional and more detailed information of the viral infection. This demonstrates the utility of our approach which helps gather additional information and enhances our knowledge of the COVID-19.

References

[1] World Health Organization: Weekly epidemiological update on COVID-19 - 8 March 2022. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---8-march-2022.
[2] Chen J, Qi T, Liu L, Ling Y, Qian Z, Li T, Li F, Xu Q, Zhang Y, Xu S, Song Z. Clinical progression of patients with COVID-19 in Shanghai, China. Journal of infection. 2020 May 1;80(5):e1-6.
[3] Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DS, Du B. Clinical characteristics of coronavirus disease 2019 in China. New England journal of medicine. 2020 Apr 30;382(18):1708-20.
[4] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The lancet. 2020 Feb 15;395(10223):497-506.
[5] Chen E, Lerman K, Ferrara E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR public health and surveillance. 2020 May 29;6(2):e19273.
[6] Müller M, Salathé M, Kummervold PE. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503. 2020 May 15.
[7] Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi MA, Yang YC. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. Journal of the American Medical Informatics Association. 2020 Aug;27(8):1310-5.
[8] Zhu M, Song Y, Jin G, Jiang K. Identifying personal experience tweets of medication effects using pre-trained Roberta language model and its updating. InProceedings of the 11th international workshop on health text mining and information analysis 2020 Nov (pp. 127-137).
[9] Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association. 2017 Jul 1;24(4):841-4.
[10] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.
[11] World Health Organization: Common symptoms of COVID-19. https://www.who.int/mongolia/multi-media/item/common-symptoms-of-covid-19
[12] Centers for Disease Control and Prevention. Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html
[13] World Health Organization: Diagnostic testing for SARS-CoV-2. https://apps.who.int/iris/bitstream/handle/10665/334254/WHO-2019-nCoV-laboratory-2020.6-eng.pdf.