A Text Classification Approach for the Automatic Detection of Twitter Posts Containing Self-reported COVID-19 Symptoms

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, Abeed Sarker
Department of Biomedical Informatics, School of Medicine, Emory University,
Atlanta GA 30322, USA
{m.a.al-garadi,yuan-chi.yang,slakama,abeed.sarker}@emory.edu

Abstract

Social media presents a potentially useful resource for conducting automated surveillance of the spread of COVID-19. As part of our efforts to establish a social media based surveillance system, our objective in this paper is to describe the development and evaluation of a natural language processing and machine learning framework using Twitter data for the automatic detection of tweets containing self-reported symptoms by COVID-19-positive patients. We modeled the tweet-level symptom detection as a binary classification task. We used annotated data from a past study, which includes posts from users who had self-identified to have tested positive for COVID-19, and discussed their symptoms over multiple tweets. We trained a BERT-based classifier to automatically detect tweets mentioning COVID-19-related symptoms posted by the users (positive tweets) and those that do not (negative tweets). The F_1-score performance of the Twitter COVID-19 symptom classifier was evaluated using the F_1score metric over a held-out test set. The classifier achieved an F_1score of 0.71 and 0.96 for positive and negative classes, respectively. Following the training and evaluation of the classification approach, we ran it on unlabeled data from December 2019 to early February 2020, and qualitatively analyzed the classified tweets to examine the effectiveness of the classifier.

1 Introduction

On March 11, 2020, the World Health Organization (WHO) declared Coronavirus disease 2019 (COVID-19) as a global pandemic (Cucinotta and Vanelli, 2020). The delay and shortage of accessible testing at the onset of a new pandemic, such as the current one, may make monitoring its spread and preparing an appropriate response major challenge. Multi-faceted epidemiological studies focusing on COVID-19 and various aspects of it (Rothan and Byrareddy, 2020) are being carried out to provide researchers, public health experts, and policy makers with the tools and data to plan an effective public health and policy decisions regarding the global crisis. However, traditional approaches for collecting epidemiological data, such as population health surveys and cohort studies, are time-consuming and slow, which can have critical impact during a pandemic, such as the current one (Eysenbach, 2009). The information available on the Internet and social media have the potential to provide additional means and insights to address this problem (Al-garadi et al., 2016a). A comparative study between data derived from the Internet and those derived from more traditional/formal sources confirmed the practical potential of mining unstructured, text-based, and online forum data for supplementing and validating structured quantitative data collected from clinical studies (Brownstein et al., 2009). In recent years, social media has become a widely used platform where people share their health-related activities and concerns. Now, more than ever, social media offers a progressively important opportunity to examine the usefulness of the data generated in close to real time by users for conducting surveillance of the pandemic and complementing other sources of epidemilogical information.

Social-media-based infectious disease or syndromic surveillance systems rely on the detection of voluntary and spontaneous self-reports of symptoms from the general population on social networks (often only those publicly posted) such as Twitter. Conventional practice, in contrast, depends on an established system of mandatory and voluntary reporting of known infectious diseases by doctors and laboratories to governmental agencies (Velasco et al., 2014). Due to the latency associated with traditional reporting systems, and the rapid growth of social media over the last couple of decades, the use of social media and user-generated data can lead to faster identification of cases of infectious diseases. Direct access to such data can allow epidemiologists interested in surveillance to identify possible public health risks such as rare or new diseases, detect unreported cases, or predict outbreaks for early warnings. The COVID-19 pandemic is, in fact, the first such global outbreak following the invention of internet and the use of social media. Thus, despite the promise of social media, its usefulness is relatively unexplored in practical settings. In this paper, we describe our first steps in developing a social media based COVID-19 based surveillance system.

2 Related works

Several studies have shown that social media contains important data that can be utilized to track infectious diseases and pandemics (Al-garadi et al., 2016a). Unlike conventional surveillance approaches involving traditional data (*e.g.*, emergency department visits), in which data collection is time-consuming and typically expensive, social media information can be collected in close to real time and often at little to no cost (Al-garadi et al., 2016a). For example, Broniatowski et al. (2013) described the development of an influenza surveillance system on the basis of an analysis of Twitter posts. The numbers derived from Twitter correlated with the surveillance data from the Centers for Disease Control and Prevention (CDC), and the Department of Health and Mental Hygiene of New York City. The results validated the usefulness of using Twitter as a surveillance system for monitoring influenza cases.

Social media data have been used for other real time surveillance tasks, including in our past studies focusing on pharmacovigilance (Sarker et al., 2015), toxicovigilance (Sarker et al., 2019) and cyber-crime detection (Al-garadi et al., 2016b). Our ongoing study to utilize social media data for conducting surveillance of COVID-19 is not the first or the only one. The potential of social media has been discussed by multiple studies. For example, Cinelli et al. (2020) studied the diffusion of COVID-19-related information from multiple social networks, specifically focusing on accurate and misinformation. In a separate study, Wang et al. (2020) showed that the phrase "*shortness of breath*" spiked on the Chinese social media platform WeChat weeks before the first few COVID-19 cases were confirmed. Gharavi et al. (2020) collected geo-located tweets from the United States and containing keywords related to COVID-19, such as "cough" or "fever." Their main aim was to observe a temporal lag between the increase in tweets that contain COVID-19-related keywords and officially reported positive cases. Although these studies reveal important information and patterns, their main drawback is the use of simple filtering methods, such as keyword filtering, which do not necessarily suggest that the symptoms being posted are actually from users who have COVID-19. Furthermore, these keywords can also pertain to flu, particularly during flu season. Hence, keyword-based filtering assumes that the official reporting system is *aware* of the ongoing pandemic, and thus, such keyword-based filters may not be effective for prospective tasks such as early warning generation. In our view, an early warning system requires an intelligent system that can distinguish between symptom-mentioning tweets from users to have tested positive from COVID-19 or are likely to be infected with COVID-19 (or, to generalize, a novel infectious disease) and flu-related tweets, and aggregate and report such cases before the official incidence in a specific geo-location is known. Thus, in our approach, we plan to focus on first detecting tweets from potentially COVID-19 positive users, and then identifying relevant symptoms reported.

Data set type	**Positive(1)**	**Negative(0)**	**Total**
Train set	428	3669	4097
Evaluation set	55	474	529
Test set	128	1106	1234
Total	611	5249	5860

Table 1: Annotated data distribution.

3 Material and Methods

3.1 Data

Data Annotation To collect tweets for training and evaluating supervised machine learning algorithms, we first compiled a high-quality, self-reported annotated Twitter dataset (Sarker et al., 2020). We searched for all users who clearly mentioned that they had been tested for COVID-19 and that their test turned out positive. We tracked their profile and annotated their profile tweets that mention self-reported COVID-19 symptoms. This step ensured that we use accurate tweets with symptoms related

Positive Tweets	Negative Tweets
woke up with a headache, extremely sore ribs from the coughing. suspected a chest infection. temperature rose back to 'Number', coughed more. sneezed more. actually felt down. voice sounded weak and flu-like. couldn't make a complete sentence wo gaspin.	the coronavirus bill guarantees sick leave to only 'Number' of workers. big employers like mcdonald's and amazon aren't required to provide paid sick leave. this is what happens when we have a gov't that is controlled by corporations. if you're sick of it then elect more progressives
hey, good people! just want you all to know that my wife and i tested positive for COVID-19 we assume that our entire family has been exposed. we are recovering very well. my wife and son have been fever-free for over three days. this is good. please pray for us. relentless	i can't help but think of dave chappelle's "lip sweat" joke whilst watching boris' address tonight. #COVID-19 url
day 3- symptoms: my cough was a little heavier, normal energy levels.	yea, hospitals are turning people away and tell them "unless your dying please stay home" by that time its to late. i get it, there are too many sick people for hospitals to deal with. its a bad situation for everyone.
nother thing: i noticed a sore throat also in the beginning days of my cough! pain in eyes that came later along with my headaches and loss of taste smell	my family and i have you to thank for recommending preparedness. you may not think so but you are doing god's work # COVID-19 ¡hashtag¿ coronavirus url
been sick for about three days just tested positive for COVID-19 i had a really bad headache on off the past two weeks	amp; just two days ago i started getting more symptoms, take this shit serious
amp; stay in the house fr	nother coronavirus case of unknown origin identified in california.

Table 2: Examples of positive tweets with self-reported COVID-19 symptoms and negative tweets without self-reported COVID-19 symptoms

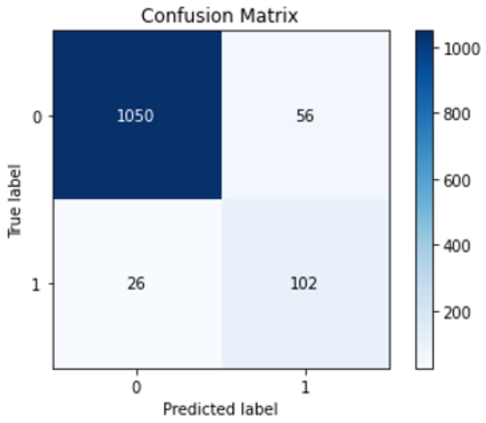

Figure 1: Classifier Confusion Matrix.

Metrics	Precision	Recall	F1-score	Acc
Negative	0.98	0.95	0.96	93%
Positive	0.65	0.80	0.71	

Table 3: Performance of the developed Twitter COVID-19 symptom classifier on the BERT model.

to COVID-19 patients. In the previous work, we quantitatively analyzed tweets that contain self-reported COVID-19 symptoms, which we compared against clinical studies. The results became the basis for a symptom lexicon for the research community (Sarker et al., 2020), and the annotated data associated with symptoms was used in this study for training and evaluating our classification system. Table 1 shows the tweet-level annotated data distribution where 0 means the absence of any potential self-reported COVID-19 symptoms (negative tweets), and 1 means the presence of one or more self-reported COVID-19 symptoms (positive tweets). Examples of positive tweets and negative tweets are shown in Table 2.

Experimental Dataset We used the US-based tweets collected by researcher in (Gharavi et al., 2020), which contained the keywords “fever” or “cough,” to cover early period of pandemic from December 2, 2019 to February 5, 2020.

3.2 Twitter COVID-19 Symptom Classifier

The purpose of the classification model is to classify whether a tweet has potential COVID-19 symptom (positive tweet) or not (negative tweet). To build the Twitter COVID-19 symptom classifier, we needed a supervised classification model that could learn from a relatively small dataset (given that self-reported symptoms comprise only a small portion of all COVID-19 related Tweets) to accurately distinguish tweets with potentially self-reported COVID-19 symptoms from tweets with no COVID-19 symptoms. For this purpose, we used a transformer-based approach, such as bidirectional encoder representations from transformers (BERT), which has significantly improved the state-of-the-art in many NLP tasks (Devlin et al., 2018), and has been shown to work well for supervised classification tasks with relatively small numbers of training data. We used the BERT-large model, which consists of 16 layers (transformer blocks),

Date	Tweets with potential COVID-19 symptoms
2019-12-24	This is literally the worst cough I've ever had. I'm pretty sure I'm dying today
2019-12-27	just when I thought I was getting better I came down with fever, chills, and heavy breathing
2020-01-28	Why did this cough get worse
2020-02-01	Some people have said it makes you sick. I'm not really sure. I got the flu shot every year. Two years ago it knocked me out and my cousin—very tired—then I was fine. This year I got it and still I caught something with fever and cough. Don't know what it was. The flu?
2020-02-05	Trying not to cough from the pain is a sport of shallow breathing now

Table 4: example of early tweets with potential COVID-19 symptoms

1,024 hidden size-16 attention heads with total of 340 M parameters. The tweets were converted into dense vectors, which captured contextual meanings of character sequences. Following vectorization, a neural network (dense layer) with a softmax activation was used to predict the classes of the tweets.

4 Results and Discussion

4.1 Twitter COVID-19 Symptoms Classification

Table 3 shows the results of the developed Twitter COVID-19 symptom classifier. The table and the confusion matrix in Figure 1 show that our BERT-based model can achieve particularly high performance in detecting tweets with non-COVID-19 symptoms, with an F_1-score of 0.96. The classifier shows good performance in detecting tweets with COVID-19 related symptoms with an F_1-score of 0.71. However, the current performance can be improved in the future by adding further training data with a wide spectrum of positive COVID-19-related symptoms.

4.2 Twitter COVID-19 symptom classifier for early warning system

We applied the developed classifier to the previously-mentioned experimental dataset containing tweets from early December 2019 to early February 2020. Our qualitative analyses showed that our classifier detected early tweets that mention potential COVID-19 symptoms. This result suggests that COVID-19 symptoms may have appeared as early as December 2019 (see Table 4). However, the tweets with positive COVID-19 symptoms cannot be interpreted as examples of tweets from confirmed cases because COVID-19 has similar symptoms with seasonal flu. Disproportionality-based methods, which can compare the 2020 distribution of symptoms with distributions of flu-related symptoms from previous years may better indicate the emergence of COVID-19. In the future, researchers can create a baseline of different flu-related symptom distributions or any known symptoms of other infectious diseases that are posted on social media. Such information can help determine whether any new symptoms during a specific time of the year are normal or may be related to a potentially new pandemic. The outcomes of such model can never replace conventional surveillance systems, but they can serve as a complement and an early indicator that work best when integrated with current systems. Furthermore, studying and recording background symptom distributions (*i.e.*, regular number of tweets with infectious disease like symptoms) during different times of the year may be useful in the future to build early warning systems for epidemics of known and unknown infectious diseases.

5 Conclusion

Social media is a potential platform that offers real-time access to millions of geolocated texts (posts) covering information about self-reported health-related posts. Such data can allow surveillance epidemiologists to identify possible public health risks such as rare and new diseases, detect unreported cases, detect common symptoms, or even give early warnings regarding the mental and emotional conditions of people during the pandemic. Our objective was to develop and evaluate NLP and machine learning framework using Twitter data to automatically detect tweets with potential COVID-19-related symptoms. Once this system is fully ready, we plan to deploy it and conduct analyses of temporal data and devise strategies for building an outbreak prediction and early stage warning system. We applied our partially-developed Twitter COVID-19 symptom classifier to tweets from early December and January, and we found that it can potentially detect symptoms reported by users who may be infected with the coronavirus.

Acknowledgments

TBA

References

Mohammed Ali Al-garadi, Muhammad Sadiq Khan, Kasturi Dewi Varathan, Ghulam Mujtaba, and Abdelkodose M Al-Kabsi. 2016a. Using online social networks to track a pandemic: A systematic review. *Journal of biomedical informatics*, 62:1–11.

Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016b. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433 – 443.

David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672.

John S Brownstein, Clark C Freifeld, and Lawrence C Madoff. 2009. Digital disease detection—harnessing the web for public health surveillance. *The New England journal of medicine*, 360(21):2153.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic.

Domenico Cucinotta and Maurizio Vanelli. 2020. Who declares covid-19 a pandemic. *Acta bio-medica: Atenei Parmensis*, 91(1):157–160.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gunther Eysenbach. 2009. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of medical Internet research*, 11(1):e11.

Erfaneh Gharavi, Neda Nazemi, and Faraz Dadgostari. 2020. Early outbreak detection for proactive crisis management using twitter data: Covid-19 a case study in the us. *arXiv preprint arXiv:2005.00475*.

Hussin A. Rothan and Siddappa N. Byrareddy. 2020. The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of Autoimmunity*, 109:102433.

Abeed Sarker, Annika DeRoos, and Jeanmarie Perrone. 2019. Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework. *Journal of the American Medical Informatics Association*, 27(2):315–329.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202 – 212.

Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. 2020. Self-reported covid-19 symptoms on twitter: An analysis and a research resource. *medRxiv*.

Edward Velasco, Tumacha Agheneza, Kerstin Denecke, Goeran Kirchner, and Tim Eckmanns. 2014. Social media and internet-based data in global systems for public health surveillance: a systematic review. *The Milbank Quarterly*, 92(1):7–33.

Wenjun Wang, Yikai Wang, Xin Zhang, Yaping Li, Xiaoli Jia, and Shuangsuo Dang. 2020. Wechat, a chinese social media, may early detect the sars-cov-2 outbreak in 2019. *medRxiv*.