IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received May 10, 2022, accepted May 23, 2022, date of publication June 2, 2022, date of current version June 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3179808

Identifying COVID-19 Personal Health Mentions From Tweets Using Masked Attention Model

LINKAI LUO, YUE WANG, AND DANIEL Y. MO
Department of Supply Chain and Information Management, The Hang Seng University of Hong Kong, Hong Kong, SAR

Corresponding author: Yue Wang (yuewang@hsu.edu.hk)

This work was supported in part by Research Grants Council of Hong Kong under the Research Matching Grant Scheme (RMGS) under Grant 700004.

ABSTRACT Twitter has been an important platform for people to discuss and share health-related information. It provides a massive amount of data for real-time monitoring of infectious diseases (such as COVID-19) and freeing disease-prevention organizations from the tedious labor involved in public health surveillance. Personal health mention (PHM) detection is one of the critical methods to keep up-to-date on an epidemic's condition; it attempts to identify a person's health condition based on online text information. This paper explores PHM identification for COVID-19 through Twitter. We built a COVID-19 PHM data set containing tweets annotated with four types of COVID-19-related health conditions. A masked attention model was devised to classify the tweets as self-mention, other-mention, awareness, and non-health. We obtained promising results on the PHM identification task. The classification results facilitate timely health monitoring and surveillance for digital epidemiology. We also evaluate how the attention mechanism and training method affect the model's predictive performance.

INDEX TERMS COVID-19, deep learning, health monitoring, social media, text mining.

I. INTRODUCTION

Diseases outbreaks like the COVID-19 pandemic have been occurring frequently worldwide in recent years. When faced with unexpected public health threats, it is critical to provide warnings as early as possible to raise the alert and to prevent harm from appearing in countries [1]. Thus, public health surveillance has gained much attention in healthcare research. Public health surveillance consists of activities aimed at continuously and systematically collecting health-related data as well as identifying and interpreting patterns found in the data. However, traditional surveillance methods are costly and time-consuming. Health-related data are usually collected from patients and reported to a public health department for professional analysis [2], but the entire procedure lacks timeliness. When a society faces a rapidly spreading infectious disease, traditional surveillance falls short in monitoring, evaluating, and predicting the trajectory of the disease. This prolongs the effective reaction time for the pandemic and could cause serious consequences.

With the popularity of the internet, large amounts of health-related data can be found on social media, blogs, online forums, and other platforms [3]. The number of social media users has been growing rapidly in the past decade. People discuss and share information and opinions on social media platforms [4]. It has been reported that two-thirds of American adults use social media to post their status, opinions, and other information on a regular basis. This provides the opportunity for public health departments and researchers to monitor the status of public health in real time at minimal cost [5]. Research directed at disease surveillance was initiated to leverage social media data as a means to acquire early warning of epidemics or infectious diseases outbreaks. The results of the analysis aid public health departments in providing timely medical attention and quicker health services to communities. For example, a vast number of tweets with the hashtag "COVID-19" or related keywords have appeared in recent months. Many organizations, such as *The Atlantic*, have launched COVID tracking projects to analyze and monitor the status of COVID based on tweets (https://covidtracking.com/). This may provide greater support for public health departments to intervene in advance of the spread of epidemics. The World Health Organization (WHO) even states that early detection can be found through social media data for more than 60% of epidemics [3]. In previous research and applications, tweets have been

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

used for the early detection of infectious diseases such as Ebola [3], E. coli [5], cholera [40], and seasonal conjunctivitis [41]. Thus, health surveillance based on social media data is highly important for communities and societies globally.

A massive amount of data is generated on social media continuously, with reports that around 3.5 million English tweets related to COVID-19 are posted every day [14], but the majority of them are not informative or are irrelevant for the downstream tasks in public health surveillance. Manual identification of useful tweets is costly and time-consuming. Thus, a critical step in online health surveillance is to develop an automatic way to identify personal health mentions (PHMs) in tweets [6]. The task involves detecting whether a particular text contains a PHM. Specifically, PHM detection attempts to classify each post into one of four categories: non-health, awareness, self-mention, and other-mention. For example, the tweet ''had my COVID-19 nasal swab Saturday. Got a call last night from CDC, test was positive!'' should be categorized as a self-mention, as it mentions that the person who posted the tweet has a disease. The tweet ''COVID-19 can be prevented. Watch for these signs and symptoms'' should be categorized as awareness, as the post provides disease-related information but does not mention a specific patient or reference the person who posted it. ''Need some corona to cure my hangover'' should be classified as non-health, as the word ''corona'' here is the beer brand, not the virus.

PHMs are crucial for health surveillance. They can screen posts to filter out those irrelevant to health and collect those relevant to health in order to keep them for subsequent public health data analysis and downstream public health applications. However, studying PHM detection based on tweets has several challenges. First, tweets are usually short texts in free form; users have no predefined template to express their opinions or health-related information, and tweets are characterized by informal and creative language, including emojis and idiomatic and ambiguous expressions. Consider, for example, ''SRSLY the @LushLtd crashed at the stroke of midnight 😩😩😩.'' For this tweet, it is necessary to contend with linguistic variations and effectively extract the semantic information from the free-form text. Second, though the amount of tweet data is immense, the amount of annotated data is usually limited due to the high cost of manual annotation; this, in turn, limits the application of tools in the task of PHM identification. Because of these challenges, researchers have acknowledged that the methods performed for social media text processing and mining are worse than for normal and standard texts [7].

This paper utilizes a novel deep neural network structure and model-training strategy to address these challenges. We built and annotated a COVID-19 PHM tweets corpus. We encoded tweets using word embeddings. The embeddings are fed into a bidirectional gate recurrent unit (BiGRU) layer, which sorts and extracts semantic information from short texts. The BiGRU layer is followed by a masked attention layer, which fully leverages the tweets' keywords to solve the issue of informal and ambiguous expressions in tweets. The outcome of this layer is then inputted into the SoftMax classifier to identify the corresponding PHM. In addition, we developed a novel epoch-wise moving-average-based training method to improve the efficiency of model training.

This paper makes the following contributions. First, we built the first COVID-19 tweet corpus for PHM identification research. We collected and annotated more than 11,000 tweets with the four types of health mentions. Second, we modeled PHM identification as a text classification task and proposed a masked attention model to classify each tweet into the four categories. The mask was able to handle the different-length issue seen in tweets. The attention mechanism was able to fully utilize the keywords in tweets and thus mitigated the challenge of informal, idiomatic, and ambiguous expressions in tweets. Third, we proposed a novel model-training method based on an epoch-wise moving average of the model parameters. The newly proposed method fully utilizes the information obtained at different training stages and has achieved better results than traditional model training. To summarize, this paper is regarded as the first mover to address the research question of COVID-19 PHM identification from tweets and develop baseline methods. The code and data are released at https://github.com/yw57721/PHM_COVID19_MaskedAtten to promote the research along this direction.

II. RELEVANT LITERATURE

A. PHM IDENTIFICATION BASED ON SOCIAL MEDIA DATA

PHM detection in social media data is a relatively newly defined research topic, although similar topics—such as public health surveillance—have been studied previously. Most existing work has used traditional machine-learning methods and has combined domain knowledge or external resources other than social media texts to identify PHMs. However, the knowledge and resources may be disease-specific, and these developed methods may be difficult to generalize to other diseases. Lamb *et al.* combined linguistic features in detection [8]. Word classes, parts of speech patterns, and stylometry have been incorporated in Twitter texts to detect influenza. Other researchers have investigated various features. For example, Yin *et al.* applied stylistic features to Twitter, such as emoji hashtags, to train a scalable classifier to detect PHMs [9]. Paul and Dredze proposed an ailment topic aspect model (ATAM) to identify ailment-related tweets [10]. This model organizes symptoms and the corresponding treatments of ailments into different topics with different levels of granularities; the combination of keywords and associated topics is then applied to identify the ailment. Gesualdo *et al.* leveraged Twitter data to detect influenza-like illnesses (ILI) [11]. They applied the case definition from the European Center for Disease Prevention and Control, which includes technical jargon related to the disease. However, most of the disease mentions are in layman's terms. Thus, they identified all the layman expressions related to the symptoms or to the disease and the corresponding technical terms or jargon, and they trained a model based on the jargon–layman terms data pairs to detect ILI cases mentioned on the internet. Coppersmith *et al.* applied basic natural-language processing techniques to detect four possible mental health conditions on Twitter [12]; they found that language-model-based methods significantly outperformed traditional methods. Karisani and Agichtein combined four types of features in their neural network structure: lexical features, syntactic features, word-embedding-based features, and context features [6]. They found that incorporating the extra knowledge could improve performance. Iyer *et al.* observed the use of figurative expressions in tweets and combined a figurative-speech detection module with a PHM detection module to augment PHM detection [13].

B. ANALYSIS OF COVID-19 BASED ON MINING SOCIAL MEDIA DATA

Recently, much attention has been drawn to mining social media data, such as tweets, to analyze the status and public response to COVID-19. For example, the Workshop on Noisy User-generated Text in EMNLP 2020 organized one shared task on the identification of informative COVID-19 text from English tweets [14]. The participant teams of the task were provided with a corpus containing 10k of annotated tweets with two labels, "informative" and "uninformative." Informative tweets included the mention of suspected cases, confirmed cases, deaths, number of tests performed, etc. In this shared task, most of the top-ranked teams applied a pre-trained language model such as BERT and its variants for this binary classification task [15]. The techniques of adversary training and ensemble were widely used. The top 10 teams applied model-ensembling to leverage the power of different models [16]. It should be noted that the informative tweets in this shared task covered three classes in the PHM identification task studied in this paper, namely self-mention, other-mention, and awareness. The uninformative tweets corresponded to the non-health class in this paper. Thus, the research question in this paper is finer in granularity. We argue that PHM identification is more useful for downstream applications of health monitoring, as the tweets in the awareness class cannot provide much information on the latest status of the disease.

Researchers have also sought to understand the tweets' content related to COVID-19. Some work has been done to characterize self-reported symptoms, experiences with testing, and other activities related to COVID-19 from social media. Alanazi *et al.* manually identified self-reports of COVID-19 symptoms in tweets. They conducted an offline interview with the posters to rank the appearance of the first three symptoms and then identified the most common ones [17]. However, their work mostly deals with descriptive statistics and cannot identify the symptom automatically. Mackey *et al.* applied a biterm topic model to identify tweets about self-reported experiences and symptoms of COVID-19 [19]. The tweets were then clustered into five main categories, such as the report of symptoms, discussion of recovery, and confirmation of negative COVID-19 diagnoses.

Besides the characterization of symptoms and experiences, some papers have also mined public opinion from tweets. Feldman trained GPT-based models for prompt-based queries on public opinion toward COVID-19 [39]. Hosseini *et al.* combined both manual annotation and topic-modeling tools to identify the frequent topics [18]. They also used the framework to track public responses to the pandemic and its evolution over time.

COVID-19 has had an unprecedent impact on human beings, not only physically but also mentally. Thus, researchers have also investigated tweets for sentiment analysis caused by COVID-19. Nemes and Kiss used RNN to classify tweets into four emotions: weakly positive, weakly negative, strongly positive, and strongly negative [35]. They also used this model to determine which emotional manifestations (such as hashtags) appeared on a specific topic during a given time period. Researchers have also recognized that emojis play a critical role in representing emotional content. A BERT-based model was presented to predict emojis in multilingual tweets [20]. Xue *et al.* used latent Dirichlet allocation (LDA) to detect topics of sentiments, popular unigrams, and bigrams in tweets [36]. They clustered the topics into five categories and found that the feeling of fear is significant in the discussion of COVID-19 cases. Similarly, Jang *et al.* used topic modeling to mine tweets and identify the COVID-19 topics that are most relevant to public health [38]. They also applied aspect-based sentiment analysis to interpret public sentiment on COVID-19-related issues. Kruspe used word2vec, ELMo, and BERT to encode tweets and map the sentiment score into a range of [0, 1] using the sigmoid function [44]. They found that the sentiment started out negative and became positive over time. But the sentiment is still under the average sentiment in most countries during the studied period. To summarize, this stream of research is not well defined. The classes of sentiments / emotions vary across papers, and there is no common data set for researchers to examine.

III. MATH NEURAL PHM IDENTIFICATOIN

A. PROBLEM DEFINITION

We model the detection of personal health mentions from tweets as a text-classification task [21]. Let $T = \left(t_1, t_2, \ldots, t_{|T|}\right) \in \mathbf{T}$ represent a tweet where t_i is the i-th word in the tweet and $\mathbf{T}$ is the tweet space. Let $\mathbf{C} = \{c_1, c_2, \ldots, c_n\}$ be a set of classes that represent different classes or labels. (Note that the terms 'label' and 'class' are used interchangeably in this paper.) There are four possible labels for a PHM task: 1) non-health, 2) awareness, 3) other-mention, and 4) self-mention. Given a training set of tweets $\mathbf{D}$ that consists of labeled tweets $\langle T, c \rangle$ such as $\langle$ "I wish my cough caused by coronavirus could stop for like five minutes", self-mention $\rangle$, where $\langle T, c \rangle \in \mathbf{T} \times \mathbf{C}$, we sought to develop a model $f : \mathbf{T} \rightarrow \mathbf{C}$ to map each tweet to a label.

In the testing or application stage, we can use the learned model to detect which label should be assigned to a new tweet T' for a disease by $f(T')$.

B. NEURAL NETWORK STRUCTURE

Deep-learning methods have been considered an efficient method for extracting the semantic information from texts [34]. Their performance is state-of-the-art in almost all natural-language processing tasks. Thus, we aim to adopt a deep-learning-based method to perform the COVID-19 PHM identification task. Tweets are short texts; thus, complicated deep-learning units may not function well on tweet texts. We use a relatively simple deep-learning unit—gated recurrent unit (GRU)—to extract the contextual information. GRU is a type of recurrent neural network (RNN) unit that extracts information from text. Compared with the popular long short-term memory (LSTM) recurrent neural network, it has a relatively simple structure but without much decreased effectiveness in handling various tasks. In addition, tweets contain informal, idiomatic, and ambiguous expressions; it is harder for a computer to understand these complicated language usages than formal text. We use an attention mechanism to emphasize the keywords in the tweets so the semantic information in the tweets can be better extracted and represented [31].

1) ENCODING OF TWEETS USING PRETRAINED WORD EMBEDDINGS

In this paper, we use a 300-dimension GloVe word embedding developed by Stanford University to encode each word [22]. For the tweet $T = \left(t_1, t_2, \ldots, t_{|T|}\right)$, the corresponding embedding of the tweet is $X = \left(x_1, x_2, \ldots, x_{|T|}\right)$, where x_t is the embedding of the *t-th* word in the text sequence based on GloVe encoding.

2) GRU-BASED TEXT ENCODER

A GRU unit uses a reset gate r_t and an update gate z_t to control how the contextual information is updated [24]. A GRU unit h_t is updated as

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \widetilde{h_t} \tag{1}$$

where $\circ$ is the Hadamard product and $\widetilde{h_t}$ is the candidate state of h_t calculated as

$$\widetilde{h_t} = \tanh(W_h x_t + r_t \circ (U_h h_{t-1}) + b_h) \tag{2}$$

The update gate controls the amount of past information that can be kept and the amount of new information added to the current state t. The reset gate determines the past states' contribution to the candidate state. These are updated as

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{4}$$

where $W_h, W_z, W_r, U_h, U_z, U_r$ represent the weights of each gate and b_h, b_z, b_r are the biases of each gate. To simplify the notation, we use $h_t = GRU(h_{t-1}, x_t)$ to represent the above calculations involved in GRU at state t.

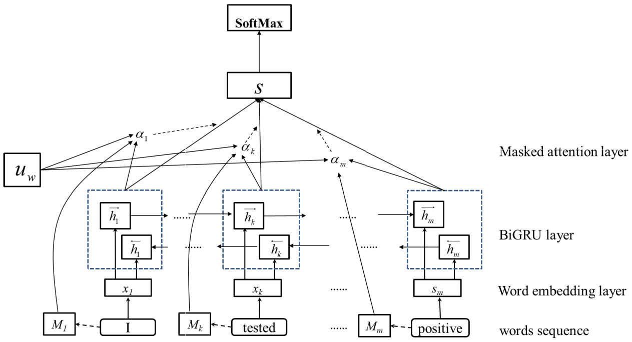


FIGURE 1. The structure of the proposed attention-based model.

3) BIDIRECTIONAL GRU (BIGRU)

Each word in a text is dependent on its previous and future words; thus, an effective approach should capture the relevant information from both past and future directions. To achieve this goal, GRU can be generalized to bidirectional GRU [25]. A BiGRU network consists of two parallel layers propagating both forward and backward. Thus, the past and future information in the text sequence can be encoded in the network. The forward layer is denoted as $\overrightarrow{h_t} = \mathrm{GRU}(\overrightarrow{h_{t-1}}, x_t)$, which reads the tweet $T = \left(t_1, t_2, \ldots, t_{|T|}\right)$ from t_1 to $t_{|T|}$, and the backward layer is denoted as $\overleftarrow{h_t} = \mathrm{GRU}(\overleftarrow{h_{t+1}}, x_t)$, which retrieves the tweet from $t_{|T|}$ to t_1. $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ can be concatenated to form the BiGRU unit at time t—i.e., $h_t = \mathrm{BiGRU}(h_{t-1}, x_t)\left(\overrightarrow{h_t} : \overleftarrow{h_t}\right)$, as shown in Fig. 1. This bidirectional network structure is able to extract the past and future information of each word in the text.

4) MASKED ATTENTION LAYER

Keywords are the critical elements for people to understand the text meaning in a quicker way. In tweet comprehension, keywords play an even more critical role, as tweets are usually in informal and ambiguous language. Thus, we deploy the attention mechanism to quantify the degree of relevance or importance of each word in the meaning of the tweet [26]. The intuitive idea behind the attention mechanism is to reward the keywords by assigning a bigger weight to them. In the traditional attention mechanism, given the output of the BiGRU h_t, the weight assigned to word embedding x_t is calculated as

$$\alpha_{\mathbf{t}} = \frac{\exp(u_t^T \cdot u_w)}{\sum_i \exp(u_i^T \cdot u_w)} \tag{5}$$

where $u_t = \tanh(W_w h_t + b_w)$ and W_w and b_w are parameters estimated in the model-training stage. $u_w$represents a context vector: it is randomly initialized and updated iteratively in the training of the model.

It should be noted that the text lengths of the tweets in the corpus vary. It is essential to handle tweets of different lengths by an appropriate method. The solution is to add padding symbols in the short text to make all the texts have the same length. The padding is meaningless and contains no semantic information; thus, the weight assigned to the padding should be 0 so that the weights for the real informative words are not diluted. In this way, we combine masking with the attention method to update the attention-based weights as

$$\alpha_{\mathbf{t}} = \frac{\exp(u_t^T \cdot u_w + M_t)}{\sum_i \exp(u_i^T \cdot u_w + M_i)} \tag{6}$$

where $M_t = -\infty$ if the t-th word in the text is padding; otherwise, $M_t = 0$. As shown in Fig. 1, the output of the attention layer is calculated as $s = \sum_t \alpha_t h_t = (s_1, s_2, \ldots s_n)$ where n is the dimension of the word embedding. This represents the weighted sum of the BiGRU outputs. To summarize, the attention-based model is used to encode the tweets. The attention layer can give different levels of "attention" to words with different degrees of relevance to the text.

5) SOFTMAX LAYER

The output of the attention layer is fed into a SoftMax function to perform the classification. Specifically, we can calculate the probability that a tweet belongs to class j as

$$\hat{p}_j = SoftMax(Ws)_j = \frac{\exp(W_j s)}{\sum_{k=1}^{C} \exp(W_k s)} \tag{7}$$

where C is the number of classes of the text. In the COVID-19 PHM identification task, $C = 4$, as there are four classes. A new tweet will be assigned to the class with the highest probability.

C. MODEL TRAINING USING EPOCH-WISE MOVING AVERAGE OF THE PARAMETERS

All the parameters in the BiGRU units and the attention layers (i.e., the W, U, b from equations (1) to (7)) are estimated using the annotated tweets. The whole tweet corpus is randomly divided into training, validation, and testing sets (with 80%, 10%, and 10% of the tweets, respectively, in this paper). The model is trained on the training set (i.e., the estimation of the parameters) by minimizing the cross-entropy loss between the true label and the predicted label distributions—i.e., $L = -\sum_{i=1}^{M} \sum_{j=1}^{C} I_j^i \log(\hat{p}_j^i)$, where M represents the training-set size, $\hat{p}_j^i$ is the predicted probability calculated (based on (7)) indicating the i-th tweet having the j-th label, and I_j^i is a binary variable. I_j^i=1 means the i-th tweet is correctly labeled to class j. After each training epoch, the model is tested on the validation set. Early-stopping strategy is used to avoid model overfitting [30]; specifically, the model training stops when the training loss on the validation set stops decreasing. The updated parameters in the current epoch are used as the final training results. If the loss value continues to decrease, the next training epoch is conducted.

However, during the experiment, we noticed that although the overall performance of F1 tended to increase with the training epochs (before the stopping epoch), the classification performance of a particular class might not increase correspondingly. For example, the overall COVID-19 identification performance in the 4th training epoch was better than that of the 3rd epoch; however, the performance for label 2's classification in the 4th epoch worsened. This means that the parameters in the 3rd epoch were better for class 2 COVID-19 PHM identification. Thus, overall performance improvement comes at the expense of certain classes. This may not be good for health surveillance purposes.

These findings motivated us to consider previous epochs' parameter values in the current epoch. Specifically, we used the epoch-average parameters to update the current epoch's parameter. Let $\left\{W_i^t, U_i^t, b_i^t : i \in \{h, z, r, w\}\right\}$ be the set of parameters learned at epoch t, where t starts from 1. Then, we reset the parameter of the current epoch as the three-period moving average of the latest three epochs,

$$\begin{aligned} W_i^t &\leftarrow \frac{W_i^t + W_i^{t-1} + W_i^{t-2}}{3} \\ U_i^t &\leftarrow \frac{U_i^t + U_i^{t-1} + U_i^{t-2}}{3} \\ b_i^t &\leftarrow \frac{b_i^t + b_i^{t-1} + b_i^{t-2}}{3} \end{aligned} \tag{8}$$

for $t > 2$, $i \in \{h, z, r, w\}$; for $t \leq 2$, we simply take the average of all the previous epochs' parameters as the current epoch's parameter value.

IV. EXPERIMENT SETTINGS

A. DATA PREPARATION

We built a COVID-19 tweet corpus for PHM identification containing 11,231 tweets posted from February to May 2020. The tweets were collected using hashtags such as "COVID", "SARS," "coronavirus," "corona," "pandemic," and "quarantine." We further processed the tweets to remove the mentions, hashtags, and links and only keep the relevant textual content. Two annotators with relevant background knowledge of medical and public health independently annotated the 11,231 tweets. We used Fleiss's kappa to evaluate the inter-annotator agreement between the annotators. The Fleiss's kappa value was 0.76, suggesting a substantial agreement of the annotation [32]. For the tweets annotated with different labels, the two annotators and one consolidator with expertise in public health conducted further discussions to resolve the annotation disagreement. The four labels are as follows:

1 (self-mention): The tweet mentions a disease or health condition for the person who posted it.

2 (other-mention): The tweet mentions a disease or health condition for a person other than the person who posted it.

3 (awareness): The tweet contains the name of the disease but is not related to any specific people being sick.

4 (non-health): The tweet may contain the name of the disease but is not related to health.

Table 1 shows examples of these four types of tweets.

The four labels are in descending order of importance from a health-monitoring perspective. The first two reflect the status of the disease and the other two do not. Thus, if a tweet can be assigned multiple labels, we will keep the more important label only. For example, "We got our coronavirus

TABLE 1. Examples of PHMs in tweets.

Category of PHM	Example tweet
Self-mention	had my COVID-19 nasal swab Saturday. Got a call last night from CDC, test was positive!
Other-mention	I found out today that my Grandad has COVID-19. Literally heartbroken.
Awareness	to stay away from coronavirus? Wear your mask and wash your hands!
Non-health	Need some corona to cure my hangover.

TABLE 2. Total numbers and percentages of each label for COVID-19.

Data Size	1 (self-mention)	2 (other-mention)	3 (awareness)	4 (non-health)
11,231	2.8%	9.8%	72.7%	14.7%

test results. I am positive. A is also positive. The basement quarantine continues'' should have both self-mention and other-mention labels, but we will only assign the self-mention label to the more important tweet.

The numbers for the data size and the distribution of the data among the four classes are shown in Table 2. The data distribution is highly imbalanced. The data sizes for classes 1 and 2 are small, and class 3 has a large majority of the data. Although the annotated tweets are not on a massive scale due to time and cost constraints, these statistics in Table 2 roughly indicate the proportion of different types of COVID-19-related tweets on Twitter. It is reasonable that significantly fewer patients will post tweets and most of the tweets reflect people's awareness of the disease.

B. MODEL TRAINING

The entire data set was randomly divided into training, validation, and testing sets, with 80%, 10%, and 10% of the tweets, respectively. We used stratified splitting to make sure that the proportion of these four labels was roughly the same in the training, validation, and testing sets. Adam optimization was used, with the learning rate of 0.005. The embedding dropout rate was set at 0.3, and the hidden dropout rate was set at 0.5. The hyperparameters were primarily set by trial based on the performance. We lowercased all the words in the tweets and used twikenizer to tokenize the tweets.

C. PERFORMANCE MEASURE

We use accuracy, precision, recall, and F1 as performance measures, as they are the major metrics for classification tasks [27]. COVID-19 PHMs are modeled as a multiple classification problem. Precision and recall for each class can be defined as: $\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$, $\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$, where TP_i, FP_i, and FN_i are the true positive, false positive, and false negative classification results for class I, respectively. The F1 ratio is defined as $\text{F1}_i = \frac{2*\text{precision}_i*\text{recall}_i}{\text{precision}_i+\text{recall}_i}$. To evaluate the overall performance across multiple classes, weighted precision, recall, and F1 score are used. The overall accuracy is defined as the ratio between the total number of true positive and the size of the testing set.

V. RESULTS AND DISCUSSION

A. OVERALL PERFORMANCE

To evaluate the effectiveness of the proposed approach, we use four popular methods as the baseline for performance comparison. These methods are:

- fastText [23]: fastText was developed by Facebook in 2017 for text classification purposes. It has become the de facto approach for text classification, due to its simplicity and effectiveness. The embeddings of each word in the text and the corresponding n-gram features are fed into a hierarchical SoftMax to get the corresponding predicted labels. For example, the 4-grams of the word ``cough'' are ``<cou'', ``coug'', ``ough'', and ``ugh>'', where ``<'' and ``>'' indicate the start and end of a word. A 300-dimension GloVe is used for the embeddings.
- Convolutional neural network (CNN) [28]: CNN takes the embeddings of the text (in the format of an embedding matrix) as the input. A set of convolutional kernels is applied to the input to extract the characteristics of the text. After certain operations, such as max-pooling and dropout, the features are fed into a SoftMax function to classify the text. CNN has been acknowledged to effectively extract the hierarchical and ordering information in the text [42]. In our experiment, the word embeddings dimension was 64; the kernel sizes were set at 2, 3, and 4; the dropout rate was 0.5, and the learning rate of Adam was 0.005. All hyperparameters were optimized by trial.
- Bidirectional long short-term memory network (BiLSTM) [29]: LSTM is a type of RNN unit that is capable of capturing the long-range contextual dependency in a text. Each LSTM unit contains a forget gate, an input gate, and an output gate to control the information flow and update in the network. BiLSTM is a parallel LSTM structure that fully leverages the forward and backward contextual information [37]. Similar to CNN, in our experiment, 64-dimension embeddings and a dropout rate of 0.5 were used, and the learning rate of Adam optimizer was 0.005.
- Bidirectional Encoder Representations from Transformers (BERT) [33]: BERT is a language model trained on words from BooksCorpus and Wikipedia with more than three billion words. BERT and its variants have shown an extremely good capability of extracting the sematic features from texts and have achieved state-of-the-art performance in many NLP tasks [43]. This paper uses $\text{BERT}_{\text{BASE}}$ to encode the tweets. The representation of the tweets was fed to a SoftMax function to perform the classification task. The batch size was 32 and the learning rate was 0.00005. The $\text{BERT}_{\text{BASE}}$ model was fine-tuned on the training set and tested on the testing set.

TABLE 3. The overall performance of the models for COVID-19 PHM identification.

Model	Accuracy	Precision	Recall	F1 score
fastText	0.7811	0.7637	0.7811	0.7667
CNN	0.7785	0.7626	0.7785	0.7626
BiLSTM	0.7785	0.7625	0.7785	0.7618
BERT	0.7690	**0.7943**	0.7690	0.7560
Masked-Attn	**0.8043**	0.7936	**0.8043**	**0.7942**

The accuracy, precision, recall, and F1 scores of all the methods used for COVID-19 PHM identification are shown in Table 3. The best performances are highlighted in **bold**. The attention model with masking operation achieves the best performance identification in terms of accuracy, recall, and the F1 score. For precision, the Masked-Attn model is outperformed only by BERT. Thus, the Masked-Attn model achieves the best overall performance due to its efficiency at extracting information from the text, even the short texts used in tweets.

It should be noted that the powerful BERT model does not achieve a satisfactory overall result, although it has the highest precision rate. The reason is that the BERT model is very heavy. It has been pre-trained using the general BooksCorpus and Wikipedia. There may be a great semantic difference between the PHM corpus and the corpus to train BERT. The PHM corpus built in this paper is on a small-to-medium scale. Fine-tuning BERT using the partial PHM data may not adapt the BERT to the PHM domain well. However, if we have a massive amount of annotated PHM data, BERT will have great potential to outperform existing methods.

The text length of tweets is relatively short compared with the long text. Research in natural language processing has observed that the same technique usually performs worse for short texts in some tasks (such as name entity recognition) due to the lack of sufficient contextual information in short texts [8]. Thus, the power of other deep-learning-based methods to extract semantic information from texts cannot be fully exploited. Some methods' performances are even worse than the simple fastText.

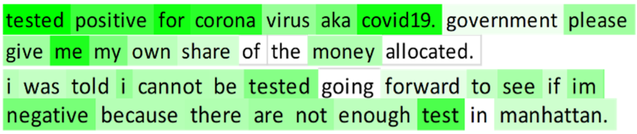
B. THE EFFECT OF ATTENTION AND MASKING OPERATION ON PERFORMANCE

The attention mechanism emphasizes the role of keywords in understanding the text. To show its effectiveness, we conducted the experiment using the following two network structures.

- BiGRU: The details of BiGRU can be found in section III.B.3 of this paper. The output of the BiGRU layer was fed directly to the SoftMax function without passing the attention layer. In the experiment, 128-dimension embeddings and a dropout rate of 0.3 were used, and the learning rate of the Adam optimizer was 0.0005. The training batch size was 32 and the number of training epochs was 10.

TABLE 4. The overall performances of the models for COVID-19 PHM identification.

Model	Accuracy	Precision	Recall	F1 score
BiGRU	0.7731	0.7550	0.7731	0.7472
BiGRU-Attn	0.7482	0.7566	0.7482	0.7517
Masked-Attn	0.8043	0.7936	0.8043	0.7942


tested positive for corona virus aka covid19. government please give me my own share of the money allocated.

i was told i cannot be tested going forward to see if im negative because there are not enough test in manhattan.

FIGURE 2. Visualization of attention values of different tokens.

- BiGRU-Attn: To show the effect of the masking operation on the performance, we conducted an experiment using the vanilla attention architecture, i.e., a BiGRU network followed by an attention layer. The model architecture was the same as Fig.1, except all the masks were removed. The hyperparameters setting was the same as BiGRU.

The experiment results in Table 4 also show that the attention mechanism can significantly improve the performance, as Masked-Attn outperformed BiGRU by a significant margin (p-value = 0.031 for the one-tailed two-sample p-test on the overall accuracy). We also noticed that masking did improve the performance significantly, as shown in Table 2 (p-value = 0.001 for the two-sample p-test on the overall accuracy between BiGRU-Attn and Masked-Attn). Thus, to calculate the attention among tokens in the text, it is better to mask out the padding tokens so they will not decrease the contribution of real tokens in the classification task.

According to the attention value calculated in the model, it was found that COVID-19-related keywords tended to have bigger attention weight. These keywords improved the performance of PHM identification. The visualization of tokens' attention in typical tweets is shown in Fig. 2. Darker shades indicate bigger values. For the first tweets, the tokens of "tested," "positive," "for," "corona," "covid19," and "me" have relatively big attention values. They are either directly related to COVID-19 or pronouns ("me"). The word "for" also has big attention. This is because its adjacent words ("tested," "positive," and "corona") all have big attention values. Thus, the attention of "for" is affected. For the second tweets, similar findings can be observed. The words of "i" (the second one), "tested," "negative," and "test" have relatively big attention value. Based on the observation, we know the COVID-19-related keywords contribute more in the masking attention model and thus lead to better performance.

C. THE EFFECT OF AVERAGING-BASED MODEL TRAINING ON PERFORMANCE

We leveraged the model parameters trained in different training epochs to update the final model parameters. To show the effectiveness of this training strategy, we conducted another

TABLE 5. performance comparisons between models trained with and without epoch-wise averaging.

Training method	Label	Precision	Recall	F1 score
Model training without epoch-wise average	Self-mention	0.5263	0.3571	0.4255
	Other-mention	0.7083	0.5354	0.6099
	Awareness	0.8274	0.9373	0.8789
	Non-health	0.6818	0.3846	0.4918
Model training with epoch-wise average	Self-mention	0.5625	0.6429	0.6000
	Other-mention	0.6900	0.5433	0.6079
	Awareness	0.8439	0.9176	0.8792
	Non-health	0.6574	0.4551	0.5319

experiment using the same network structure as shown in Fig. 1 but with an early-stopping-model training method, wherein the parameters are tuned without considering previous epochs' values, i.e., without epoch-wise averaging. The results of PHM identification for each class are shown in Table 5. It can be observed that the precision, recall, and F1 scores have mostly increased across the different labels. These increases were relatively large for classes with fewer samples, such as class 1 (from 0.4255 to 0.6000) and class 4 (from 0.4918 to 0.5319). For classes 2 and 3, the F1 values are nearly unchanged (by 0.0003 and 0.006, respectively). In addition, the overall performance combing the outcomes of the four classification results also improved with the use of the proposed training method. The overall F1 score increased from 0.7835 (without epoch-wise average) to 0.7942 (without epoch-wise average). It can be observed that the overall improvement was mainly from the self-mention class. The precision, recall, and F1 increased by 6.9%, 80.0%, and 41.0%, respectively. In the application of public health surveillance, it is more important to identify the posts in self-mention and other-mention classes, as they reflect the true public disease status. Using epoch-wise averaging of model parameters during the training stage, the performance of self-mention PHM identification is improved significantly, without the significant decrease of the performance of other-mention PHM detection (the F1 score slightly decreased from 0.6099 to 0.6000). Thus, epoch-wise training is more suitable for downstream tasks in public health control.

As shown in Table 5, the method tends to favor classes with more training data—label 3 (awareness)—as the F1 score of class 3 is the highest. Thus, with the training method without epoch-wise average, model parameters are largely determined by the result of the awareness label; the performance of other labels' classifications are not fully considered in the training. The epoch-wise, average-based training could compensate for such class imbalance. The advantage of the epoch-wise average can be attributed to its generalization capability. During model training, the parameters obtained in each epoch are distributed on the periphery of the parameter space. The center of space usually has a higher generalization capability. Epoch-wise averaging makes the parameters closer to the center of the periphery, and thus exhibits better generalization and PHM identification performance.

TABLE 6. overall performance of the models for COVID-19 PHM identification.

Model	Accuracy	Precision	Recall	F1 score
Testing set I	0.8043	0.7936	0.8043	0.7942
Testing set II	0.7900	0.7786	0.7900	0.7811
Testing set III	0.7750	0.7673	0.7750	0.7675

D. THE GENERALIZABILITY OF THE PROPOSED METHOD

Twitter provides a promising data source for more effectively identifying PHMs for public health monitoring. In addition, tweets are continuously produced and updated. They can quickly capture the latest trend in the COVID-19 public health condition in a region. Thus, new textual and semantic information related to COVID-19 may be generated with the emergence of new tweets. For example, the discussion on the omicron variant of COVID-19 emerged in 2022. In this section, we would like to explore the generalization capability of the proposed method. Specifically, we will investigate whether the performance of the model learned from old data can detect PHMs accurately for new tweets.

To conduct the analysis, we collected 200 tweets posted from January to June 2021 (denoted as testing set II; Testing set I was the original test set used in Table 3), and 200 tweets posted from September 2021 to March 2022 (denoted as testing set III). Besides the hashtags used to build the COVID-19 corpus containing 11,231 tweets, we also used more hashtags such as "vaccine," "lockdown," "socialdistancing," and "omicron." The newly collected tweets were cleaned to remove non-textual content, such as mentions, hashtags, and links. They were annotated by two annotators independently in the same manner as before. The annotation disagreements were solved by a group comprising the two annotators and two more senior experts in public health and social media mining. The percentages of each label in the newly annotated tweets were 5.5% (self-mention), 18% (other-mention), 67% (awareness), and 9.5% (non-health) for testing set II and 6.5% (self-mention), 19.5% (other-mention), 64% (awareness), and 10% (non-health) for testing set III. The class imbalance issue still existed but was not as serious as the original tweets corpus shown in Table 2.

The masked attention model trained on the original COVID-19 tweets corpus was used to identify the PHMs in the newly collected two data sets. The performance comparison on various testing sets can be found in Table 6. Testing set I contains the tweets from the original corpus, with postings dated from February to May 2020. It can be noticed that the accuracy, precision, recall, and F1 score in testing sets II and III were worse than the results for testing set I, but not significantly (p-value = 0.647 on the test of difference on the accuracies between testing sets I and II; p-value = 0.357 on the test of differences in the accuracies of testing sets I and III). Thus, the experiment's results show that the model trained on the original tweets corpus has a very satisfactory generalization capability.

VI. CONCLUSION

This paper has aimed to automate COVID-19 personal health mention detection processes based on deep-learning and natural-language processing techniques. We constructed a COVID-19 tweets corpus containing 11,231 annotated tweets. Each tweet was annotated according to non-health, awareness, self-mention, and other-mention categories. The COVID-19 PHM identification was modeled as a text classification task. An attention-based model was trained to classify each tweet according to the four classes. Promising results have been achieved in terms of the overall F1 score. Additional experiments have also been conducted to study the effect of training data size on performance. It was found that the methods tended to favor the classes with larger numbers of training samples, with classes holding more data resulting in greater reliability and a higher classification performance. Through extensive experiment, we have also shown that the proposed method has a good generalization capability. Thus, the model developed using the old data set can be applied to the new tweets data set with satisfactory performance.

However, there are several limitations to this paper. The methods leverage only Twitter data, which are short texts. It is expected that incorporating domain knowledge in public health and medicine will improve the performance of short-text classification. In the fields of public health and medicine, domain knowledge and resources are especially useful, as many professional terms and instances of jargon appear in the data set. In our future work, we will combine information from a knowledge base for the COVID-19 PHM identification task. In addition, new investigations will be conducted to compensate for classes with low sample sizes. Data resampling and even text-style transfer techniques will be attempted to mitigate the data imbalance issue. Furthermore, the regions from which the posters are tweeting are not considered in the paper. Thus, the developed method is a general method and can be applied to any region in the world. However, people in different countries may have their own tweeting style and language expression syntactics. Another future study would be to incorporate the regional information into the model so that the model is more adaptive to the regions.

REFERENCES

[1] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact," *IEEE Access*, vol. 8, pp. 90225–90265, 2020.

[2] S. Declich and A. O. Carter, "Public health surveillance: Historical origins, methods and evaluation," *Bull. World Health Org.*, vol. 72, no. 2, pp. 285–304, 1994.

[3] A. Joshi, R. Sparks, S. Karimi, S.-L.-J. Yan, A. A. Chughtai, C. Paris, and C. R. MacIntyre, "Automated monitoring of tweets for early detection of the 2014 Ebola epidemic," *PLoS ONE*, vol. 15, no. 3, Mar. 2020, Art. no. e0230322.

[4] A. A. Al-Shargabi and A. Selmi, "Social network analysis and visualization of Arabic tweets during the COVID-19 pandemic," *IEEE Access*, vol. 9, pp. 90616–90630, 2021.

[5] E. Diaz-Aviles and A. Stewart, "Tracking Twitter for epidemic intelligence: Case study: EHEC/HUS outbreak in Germany, 2011," in *Proc. Web Sci. Conf.*, 2012, pp. 82–85.

[6] P. Karisani and E. Agichtein, "Did you really just have a heart attack? Towards robust detection of personal health mentions in social media," in *Proc. World Wide Web Conf. World Wide Web*, Lyon, France, Apr. 2018, pp. 137–146.

[7] G. Rizzo, B. Pereira, A. Varga, M. van Erp, and A. E. C. Basave, "Lessons learnt from the named entity recognition and linking (NEEL) challenge series," *Semantic Web*, vol. 8, no. 5, pp. 667–700, Apr. 2017.

[8] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on Twitter," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Atlanta, GA, USA, 2013, pp. 789–795.

[9] Z. Yin, D. Fabbri, S. T. Rosenbloom, and B. Malin, "A scalable framework to detect personal health mentions on Twitter," *J. Med. Internet Res.*, vol. 17, no. 6, p. e138, Jun. 2015.

[10] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health," in *Proc. 5th Int. Conf. Weblogs Social Media*, Barcelona, Spain, 2011, pp. 265–272.

[11] F. Gesualdo, G. Stilo, E. Agricola, M. V. Gonfiantini, E. Pandolfi, P. Velardi, and A. E. Tozzi, "Influenza-like illness surveillance on Twitter through automated learning of Naïve language," *PLoS ONE*, vol. 8, no. 12, Dec. 2013, Art. no. e82489.

[12] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. Workshop Comput. Linguistics Clin. Psychol., From Linguistic Signal Clin. Reality*, Baltimore, MA, USA, Jun. 2014, pp. 51–60.

[13] A. Iyer, A. Joshi, S. Karimi, R. Sparks, and C. Paris, "Figurative usage detection of symptom words to improve personal health mention detection," 2019, *arXiv:1906.05466*.

[14] D. Q. Nguyen, T. Vu, A. Rahimi, M. H. Dao, L. T. Nguyen, and L. Doan, "WNUT-2020 task 2: Identification of informative COVID-19 English tweets," in *Proc. 6th Workshop Noisy User-Generated Text (W-NUT)*, 2020, pp. 314–318.

[15] P. Kumar and A. Singh, "NutCracker at WNUT-2020 task 2: Robustly identifying informative COVID-19 tweets using ensembling and adversarial training," in *Proc. 6th Workshop Noisy User-Generated Text (W-NUT)*, 2020, pp. 404–408.

[16] A. G. Møller, R. van der Goot, and B. Plank, "NLP north at WNUT-2020 task 2: Pre-training versus ensembling for detection of informative COVID-19 English tweets," in *Proc. 6th Workshop Noisy User-Generated Text (W-NUT)*, 2020, pp. 331–336.

[17] E. Alanazi, A. Alashaikh, S. Alqurashi, and A. Alanazi, "Identifying and ranking common COVID-19 symptoms from tweets in Arabic: Content analysis," *J. Med. Internet Res.*, vol. 22, no. 11, Nov. 2020, Art. no. e21329.

[18] P. Hosseini, P. Hosseini, and D. Broniatowski, "Content analysis of Persian/Farsi tweets during COVID-19 pandemic in Iran using NLP," in *Proc. 1st Workshop NLP COVID*, 2020, pp. 1–7.

[19] T. Mackey, V. Purushothaman, J. Li, N. Shah, M. Nali, C. Bardier, B. Liang, M. Cai, and R. Cuomo, "Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: Retrospective big data infoveillance study," *JMIR Public Health Surveill.*, vol. 6, no. 2, Jun. 2020, Art. no. e19509.

[20] S. Stoikos and M. Izbicki, "Multilingual emoticon prediction of tweets about COVID-19," in *Proc. 3rd Workshop Comput. Modeling People's Opinions, Pers., Emotion's Social Media*, 2020, pp. 109–118.

[21] K. Liu and L. Chen, "Medical social media text classification integrating consumer health terminology," *IEEE Access*, vol. 7, pp. 78185–78193, 2019.

[22] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

[23] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," 2016, *arXiv:1607.04606*.

[24] P. Jia, H. Liu, S. Wang, and P. Wang, "Research on a mine gas concentration forecasting model based on a GRU network," *IEEE Access*, vol. 8, pp. 38023–38031, 2020.

[25] Q. Wang, C. Xu, Y. Zhou, T. Ruan, D. Gao, and P. He, "An attention-based BI-GRU-CapsNet model for hypernymy detection between compound entities," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Madrid, Spain, Dec. 2018, pp. 1031–1035.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[27] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on residual convolutional recurrent neural network with semi-supervised loss function," *IEEE Access*, vol. 9, pp. 7564–7575, 2021.

[28] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.

[29] M. Zeng and N. Xiao, "Effective combination of DenseNet and BiLSTM for keyword spotting," *IEEE Access*, vol. 7, pp. 10767–10775, 2019.

[30] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Comput.*, vol. 7, no. 2, pp. 219–269, Mar. 1995.

[31] Y. Wang, W. Zhao, and W. X. Wan, "Needs-based product configurator design for mass customization using hierarchical attention network," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 1, pp. 195–204, Jan. 2020.

[32] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[33] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[34] Y. Wang, X. Li, and F. Tsung, "Configuration-based smart customization service: A multitask learning approach," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 4, pp. 2038–2047, Oct. 2020.

[35] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19," *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–15, Jan. 2021.

[36] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, and T. Zhu, "Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach," *J. Med. Internet Res.*, vol. 22, no. 11, Nov. 2020, Art. no. e20550.

[37] Y. Wang, X. Li, and D. Mo, "Knowledge-empowered multitask learning to address the semantic gap between customer needs and design specifications," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 8397–8405, Dec. 2021.

[38] H. Jang, E. Rempel, G. Carenini, and N. Janjua, "Exploratory analysis of COVID-19 related tweets in North America to inform public health institutes," 2020, *arXiv:2007.02452*.

[39] P. Feldman, S. Tiwari, C. S. L. Cheah, J. R. Foulds, and S. Pan, "Analyzing COVID-19 tweets with transformer-based language models," 2021, *arXiv:2104.10259*.

[40] R. Chunara, J. R. Andrews, and J. S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak," *Amer. J. Tropical Med. Hygiene*, vol. 86, no. 1, pp. 39–45, Jan. 2012.

[41] M. Deiner, T. Lietman, S. McLeod, J. Chodosh, and T. Porco, "Surveillance tools emerging from search engines and social media data for determining eye disease patterns," *J. Amer. Med. Assoc. Ophthalmol.*, vol. 134, no. 9, pp. 1024–1030, 2016.

[42] Y. Wang, L. Luo, and H. Liu, "Bridging the semantic gap between customer needs and design specifications using user-generated content," *IEEE Trans. Eng. Manag.*, early access, Sep. 21, 2020, doi: 10.1109/TEM.2020.3021698.

[43] L. Luo and Y. Wang, "EmotionX-HSU: Adopting pre-trained BERT for emotion classification," 2019, *arXiv:1907.09669*.

[44] A. Kruspe, M. Haeberle, I. Kuhn, and X. X. Zhu, "Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic," in *Proc. ACL Workshop Natural Lang. Process. COVID*, 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.14/