Expert Systems With Applications 198 (2022) 116882

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US

Yiming Zhang [a,b], Ke Chen [a,b], Ying Weng [a,*], Zhuo Chen [c,d], Juntao Zhang [a,b], Richard Hubbard [b]

[a] *School of Computer Science, Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo, China*
[b] *School of Medicine, Faculty of Medicine and Health Sciences, University of Nottingham, Nottingham, United Kingdom*
[c] *Department of Health Policy and Management, University of Georgia, Athens, USA*
[d] *School of Economics, Faculty of Humanities and Social Sciences, University of Nottingham Ningbo China, Ningbo, China*

ARTICLE INFO

Keywords:
COVID-19 surveillance
Early warning system
Text classification
BERT
Epidemic intelligence

ABSTRACT

The World Health Organization (WHO) declared on 11th March 2020 the spread of the coronavirus disease 2019 (COVID-19) a pandemic. The traditional infectious disease surveillance had failed to alert public health authorities to intervene in time and mitigate and control the COVID-19 before it became a pandemic. Compared with traditional public health surveillance, harnessing the rich data from social media, including Twitter, has been considered a useful tool and can overcome the limitations of the traditional surveillance system. This paper proposes an intelligent COVID-19 early warning system using Twitter data with novel machine learning methods. We use the natural language processing (NLP) pre-training technique, i.e., fine-tuning BERT as a Twitter classification method. Moreover, we implement a COVID-19 forecasting model through a Twitter-based linear regression model to detect early signs of the COVID-19 outbreak. Furthermore, we develop an expert system, an early warning web application based on the proposed methods. The experimental results suggest that it is feasible to use Twitter data to provide COVID-19 surveillance and prediction in the US to support health departments' decision-making.

1. Introduction

The spread of the coronavirus disease 2019 (COVID-19) was declared by the World Health Organization (WHO) as a pandemic on 11th March 2020 (World Health Organization, 2020). The traditional surveillance system used by the public health authorities was unable to implement timely interventions to prevent COVID-19 from spreading into a pandemic. CDC notifiable disease reports typically are published weekly, with a 1-week delay, thus providing an estimate of the current situation before the release of an outbreak report is helpful (Şerban, Thapen, Maginnis, Hankin, & Foot, 2019). Social media platforms such as Weibo and Twitter may provide an alternative tool to assist the public health authorities to reduce the delay in surveillance and reporting. Twitter has more than 500 million users worldwide, and users may post their status and thoughts on Twitter, including their health conditions and other health-related conditions, in real-time. These tweets are valuable resources for researchers conducting public health surveillance.

Previous studies have used Twitter for epidemic surveillance (Culotta, 2010; Masri et al., 2019; Odlum & Yoon, 2015; Modu et al., 2017; Yousefinaghani, Dara, Poljak, Bernardo, & Sharif, 2019). These studies focused on epidemics such as Zika, Ebola, and Avian Influenza. There are, however, no COVID-19 Twitter surveillance research work at the time of our research. Moreover, many research works on epidemic surveillance using machine learning algorithms. Chen, Tozammel Hossain, Butler, Ramakrishnan, and Prakash (2016) proposed a weakly supervised temporal topic model on syndromic surveillance. Missier et al. (2016) used clustering (Twitter topic modelling) and classification for Dengue outbreak detection. Yousefinaghani et al. (2019) applied a semi-supervised approach for Avian Influenza surveillance. The authors first manually labeled 4200 sample tweets and trained Naïve Bayes (NB) classifier, then they applied a semi-supervised algorithm to label all Twitter data and then built a Twitter-based data analysis framework. Additionally, Gomide, Veloso, Meira, Almeida, Benevenuto, Ferraz, and Teixeira (2011) performed sentiment analysis to filter irrelevant Twitter data, and the study found that the activity on Twitter reflected the Dengue incidence in Brazil. In another study by Culotta (2010), the author analyzed influenza outbreaks by tracking flu-related Twitter with

* Corresponding author.
E-mail address: ying.weng@nottingham.edu.cn (Y. Weng).

https://doi.org/10.1016/j.eswa.2022.116882
Received 3 May 2021; Received in revised form 14 September 2021; Accepted 10 March 2022
Available online 14 March 2022

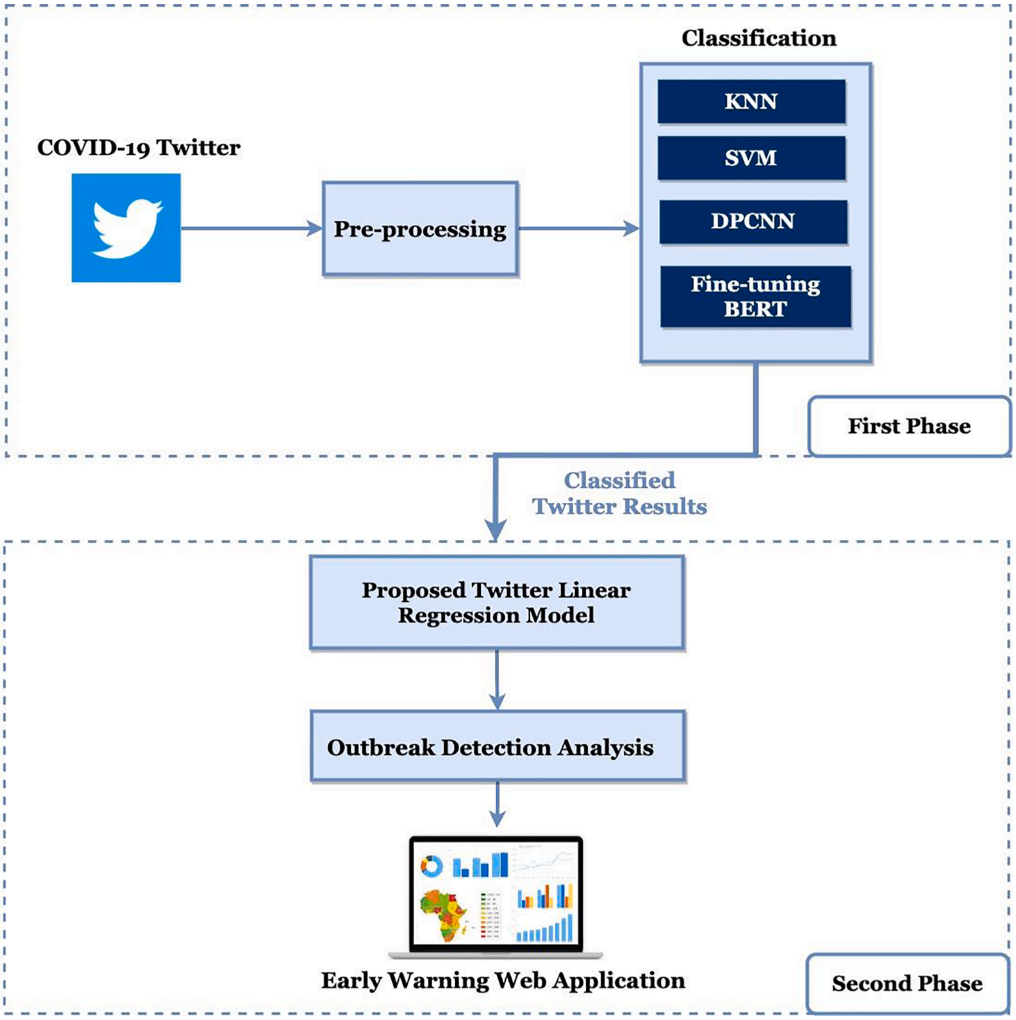

Classification
KNN
SVM
DPCNN
Fine-tuning BERT
COVID-19 Twitter
Pre-processing
First Phase
Classified Twitter Results
Proposed Twitter Linear Regression Model
Outbreak Detection Analysis
Early Warning Web Application
Second Phase

Fig. 1. The proposed COVID-19 early warning system framework.

supervised learning and then using logistic regression to predict influenza rate. Similar to the Culotta (2010) study, Santos and Matos (2014) applied Twitter data and web queries to predict the incidence of influenza.

Most of the previous works focus on using the traditional machine learning methods, such as Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbour (KNN). Compared with these traditional machine learning algorithms, our paper applies fine-tuning Bidirectional Encoder Representation from Transformers (BERT) to classify health-related Twitter messages (Devlin, Chang, Lee, & Toutanova, 2019). To the best of our knowledge, this is the first work on COIVD-19 Twitter surveillance using fine-tuning BERT.

In this paper, an intelligent early warning system using Twitter on COVID-19 in the US is proposed. We collect the COVID-19 related tweets and manually annotate them to train the supervised learning text classification algorithms. To address this Twitter classification task, we propose fine-tuning BERT. We compare classical machine learning algorithms, including K-Nearest Neighbour (KNN) and Support Vector Machine (SVM). We also implement a CNN-based model Deep pyramid convolutional neural network (DPCNN) for comparison. The experimental results demonstrate that fine-tuning BERT outperforms these methods. We then analyze the classified tweet results using a Twitter-based linear regression model to perform epidemic forecasting and epidemic early warning detection. Finally, we deploy a web application that can display the evolution of COVID-19 reported in tweets and provided early warning messages from 2020 January to 2020 April in the US. The web application is available at https://nbreservoir.com/warning/visualization.html.

The main contributions of this work are summarized as follows:

- We construct a geo-located COVID-19 Twitter dataset that can be used for COVID-19 surveillance.
- We show that fine-tuning BERT outperforms other text classification methods in COVID-19 Twitter classification with a 0.98 F1 score.
- We propose a Twitter-based linear regression model for COVID-19 forecasting and outbreak detection, and it achieves good performance.
- We develop an expert system, an early warning system web application based on the previous classification and regression results. It predicts COVID-19 outbreaks in advance and provides support for health departments' decision-making.

The rest of the paper is organized as follows. Section 2 introduces the literature review of the epidemic surveillance. Section 3 presents the early warning system framework, including Twitter text classification, Twitter-based linear regression model and web application. Section 4 includes the extensive experiment details. In Section 5, results are presented and discussed. The paper is summarized in Section 6.

2. Related works

2.1. Web and social media-based epidemic surveillance

Many previous works have analyzed and detected incidences of infectious disease using the web and social media data. As people may use the web to search for information about specific diseases, the web is a critical source of health information, and it provides a new way for disease surveillance. For example, Google proposed a Google flu trend that can detect the influenza epidemic using Google search queries data and applied a linear regression to predict the weekly influenza status in the US (Ginsberg et al., 2009). Polgreen, Chen, Pennock, and Nelson (2008) used search queries data from Yahoo! search engine to predict influenza activity. The authors fitted a linear regression model with lags of 1–10 weeks search queries data as explanatory variables to study influenza surveillance. It indicated a strong correlation between web search data and influenza activity.

Social media platforms such as Twitter may also include health-related information, and many researchers have made efforts to explore the potential to use Twitter data for public health surveillance (Jordan et al., 2019). Odlum and Yoon (2015) collected tweets that mentioned Ebola and applied time series analysis of the data. The authors reported that tweets began to rise a few days earlier than the official announcement of Ebola in Nigeria. Masri et al. (2019) first filtered the Twitter data containing keywords “Zika” and “mosquito” in Florida and the entire US. Then they used a time series model autoregression to analyze the filtered Twitter data, and the experiments demonstrated the ability to utilize Zika Twitter data for disease surveillance. Di Martino et al. (2017) proposed a similar study. The authors identified the epidemic outbreak by using Early Aberration Reporting System (EARS) anomaly detection algorithms. In another study, multiple regression methods, e.g., stacked linear regression, Support Vector Machine regression, AdaBoost with Decision Trees Regression, were applied on multiple data sources, including Google search logs, Twitter data, and hospital visit records for influenza surveillance (Santillana et al., 2015).

2.2. Intelligent epidemic surveillance application

Besides performing theory epidemic surveillance research, some studies also built an application based on the research. Freifeld, Mandl, Reis, and Brownstein (2008) built HealthMap, a free and open-sourced website, to generate disease outbreak information. Marcus and Bernstein (2011) implemented a system called Twitnfo, which can highlight peaks of high Twitter activity. Lee, Agrawal, and Choudhary (2013) proposed real-time disease surveillance using Twitter data, which provides flu and cancer surveillance. Another study provided a malaria outbreak early warning system and deployed it on a mobile platform (Modu et al., 2017). In a study by Ji, Chun, and Geller (2012), the authors gathered tweets containing specified health-related keywords and then developed an Epidemics Outbreak and Spread Detection System (EOSDS) and provided different map visualizations for epidemic detection. Şerban et al. (2019) proposed a real-time syndromic surveillance system that can detect disease outbreaks using Twitter data.

3. Methods

3.1. Framework overview

The primary objective of the study is to build an intelligent early warning system that can provide COVID-19 short-time prediction and outbreak detection. In this paper, we propose a two-phase framework for the COVID-19 early warning system. The overview of the system architecture is shown in Fig. 1. In the first phase of the framework, we first collect tweets and processed the data. Then we apply Twitter text classification algorithms to classify the Twitter content. We conduct an outbreak detection analysis in the second phase using our proposed Twitter-based linear regression model. We then build a web application that can visualize the early warning messages.

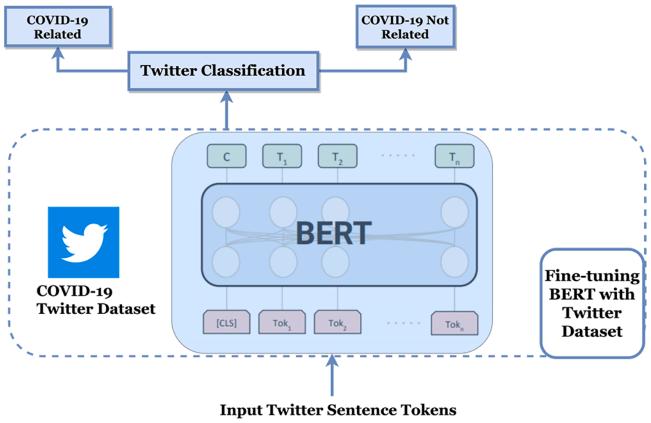


Fig. 2. Architecture of fine-tuning BERT for COVID-19 Twitter classification.

3.2. Twitter text classification method

In this paper, we employ text classification algorithms to classify the tweets. The purpose of the text classification algorithm is to identify tweets that are related to the COVID-19 outbreak. We classify each tweet as ‘COVID-19 related’ or ‘COVID-19 unrelated’ according to its content.

3.2.1. Fine-tuning BERT

Bidirectional Encoder Representation from Transformers (BERT) is a state-of-the-art language representation model pre-trained on unsupervised Wikipedia and Bookcorpus datasets. The pre-trained BERT model can be adapted to various natural language processing tasks, such as question answering and text classification, with an additional output layer (Devlin et al., 2019). The BERT model consists of multiple layers of bidirectional transformer encoder, as Fig. 2. shows. [CLS] is a special token added at the head of every input indicating the classification. For the classification, the corresponding state to [CLS] in the final hidden layer is used to generate a class prediction. With the bidirectional transformers, the model can sense the relationship in the text both from left-to-right and right-to-left, which in some cases can extract more information than the unidirectional transformer encoder model. Meanwhile, the BERT model is pre-trained using two unsupervised tasks: masked language model and next sentence prediction, which ensures the understanding of natural language. In this paper, we use fine-tuning BERT for Twitter classification. We use the same tokenizer and tokens index mapping as the BERT-Base is per-trained and input the Twitter sentence tokens. Then we add a linear layer at the output hidden state of the [CLS] token to perform a binary text classification task (COVID-19 related or COVID-19 not related).

3.3. The proposed linear regression method

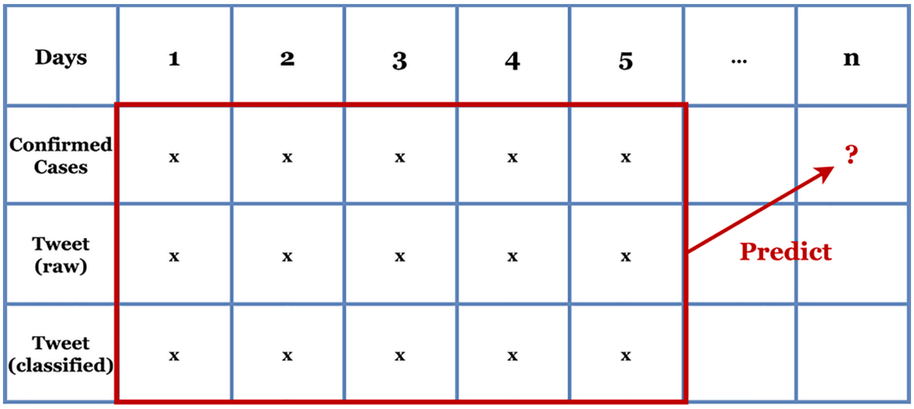
3.3.1. Linear regression

Linear regression is a statistical data analysis approach for modelling the linear relationship between a dependent variable and one or more variables (Freedman, 2009). The linear model follows the general equation below:

$$y = WX \tag{1}$$

where y denotes for the prediction, $W = (w_0, w_1, \cdots, w_n)$ denotes for the coefficients and $X = (x_0, x_1, \cdots, x_n)^t$ denotes for the data points.

The model is fitted with the coefficients using the ordinary least squares method to minimize the squared error between the dataset’s



Days	1	2	3	4	5	...	n
Confirmed Cases	x	x	x	x	x		?
Tweet (raw)	x	x	x	x	x		
Tweet (classified)	x	x	x	x	x		

Fig. 3. Illustration of the proposed linear regression model: use day 1 to day 5 data to predict the confirmed cases of day n.

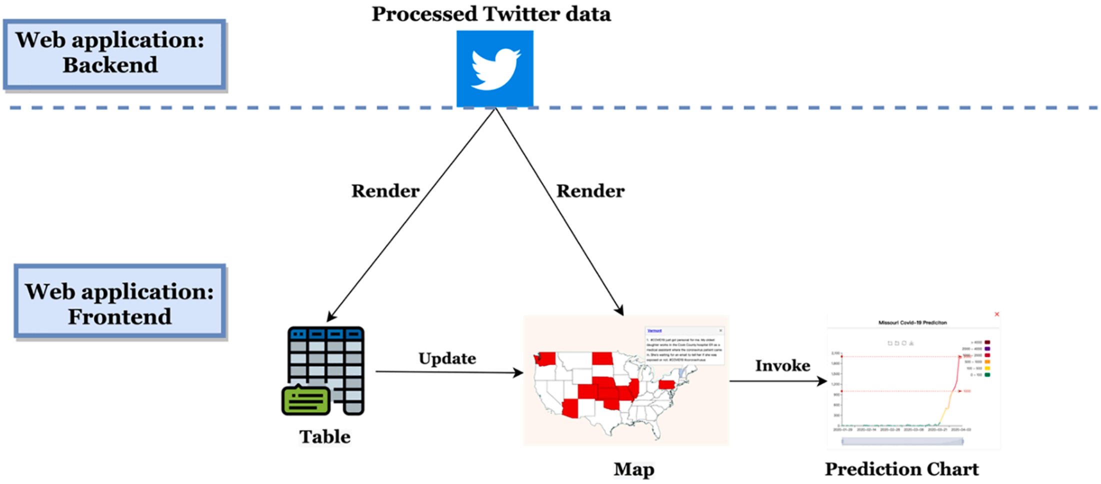


Fig. 4. Early warning system web application architecture.

observed targets and the corresponding predictions. In other words, linear regression aims at finding the best linear hyperplane for the prediction.

3.3.2. *The proposed Twitter-based linear regression model*

In this paper, we propose a Twitter-based multiple linear regression as a COVID-19 forecasting model. Early warning detection is realized by predicting future confirmed cases using the proposed linear regression model. In this work, the dependent variable *y* is the COVID-19 confirmed cases. Three explanatory variables are COVID-19 confirmed cases per state, tweet counts per state per day, and the positive classified tweet counts per state per day, all in the past time available. We apply this model to forecast future COVID-19 confirmed cases, which is illustrated in Fig. 3. The general equation is as follows:

$$y = \sum_{k=1}^{m} \alpha_k Conf_k + \sum_{k=1}^{m} \beta_k Tweet_k + \sum_{k=1}^{m} \gamma_k ClsTweet_k + \delta \quad (2)$$

where $Conf_k$ is the confirmed case count at day k, α_k denotes the effect estimate on $Conf_k$, $Tweet_k$ indicates the tweet count at day$k$. β_k denotes the effect estimate on $Tweet_k$, $ClsTweet_k$ is the classified tweet count at day$k$, γ_k represents the effect estimate on $ClsTweet_k$, m indicates a selected day prior to the prediction day, and δ denotes the bias of the regression model.

With data in the past few days as input, temporal information is included in the linear model. It helps the model make more accurate forecasting since confirmed cases in the epidemic have a time serial relationship.

3.4. *The proposed expert system: Intelligent early warning web application*

The intelligent early warning system comprises a backend for ‘COVID-19 related’ Twitter text classification, Twitter-based regression analysis, and a frontend to visualize the early warning results. The text classification is discussed in Section 3.2, and the regression analysis has been introduced in Section 3.3. Tweets containing keyword *coronavirus* is collected from Twitter API, and then the tweets are processed. BERT, which is applied as the text classification in this system, is used to classify the processed tweets. Next, the proposed regression model is performed on the classified results, and then the analyzed results are forwarded to the frontend of the system. The overall architecture of the web application is shown in Fig. 4.

In terms of the frontend of the system, it consists of two parts, a tabular component and a mapping component. In the mapping component, if the user clicks the map, a prediction chart will pop up. The frontend is implemented using HTML, CSS, JavaScript. Besides, we use Echarts for the table component and Gaode Loca for the map component (Li et al., 2018). Echarts provides various examples of interactive and intuitionistic charts for data visualization and is open source. It is implemented by JavaScript and can run smoothly on most current browsers such as Google Chrome, Safari on PCs and some mobile devices. Gaode Loca is based on the Gaode JS API map, which can call and maintain the existing JS API mapping products to realize the map visualization data system. Besides, visualizations such as scattered points, trajectories, areas, and heat maps can also be created.

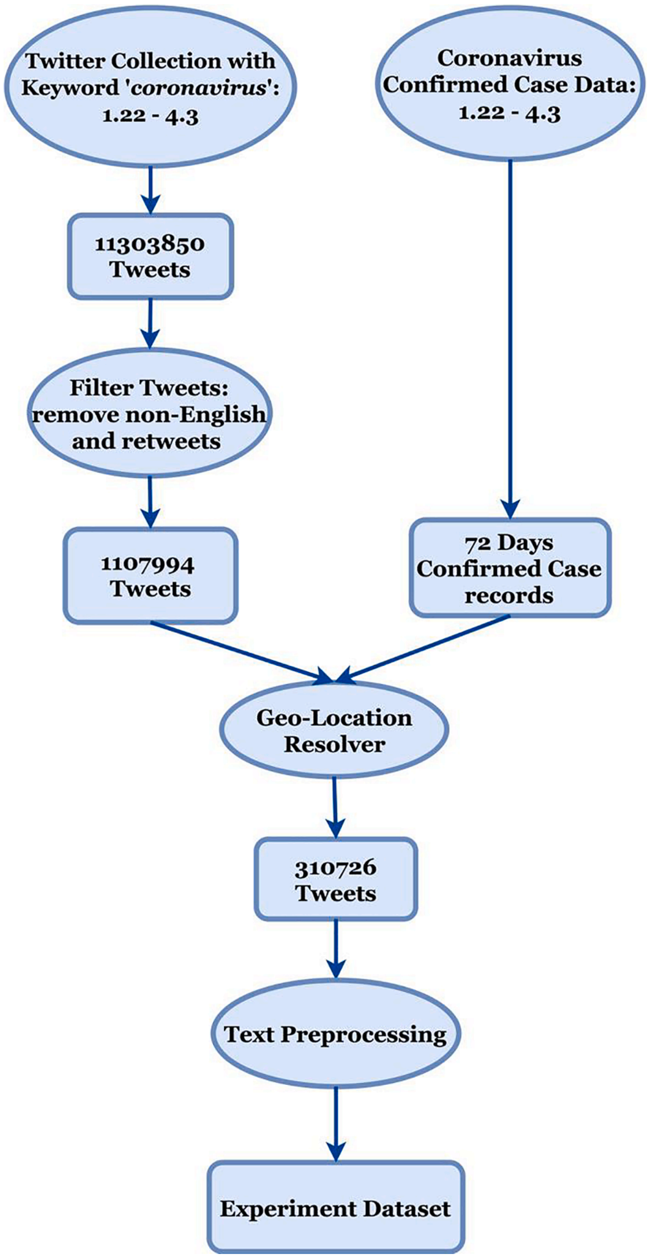

Twitter Collection with Keyword 'coronavirus': 1.22 - 4.3
Coronavirus Confirmed Case Data: 1.22 - 4.3
11303850 Tweets
Filter Tweets: remove non-English and retweets
1107994 Tweets
72 Days Confirmed Case records
Geo-Location Resolver
310726 Tweets
Text Preprocessing
Experiment Dataset

Fig. 5. Dataset collection and pre-processing flowchart.

4. Experiments

4.1. Dataset collection and pre-processing

The whole dataset collection and the pre-processing flowchart is shown in Fig. 5. The original Twitter dataset we used is a subset of the public dataset released by Lopez, Vasu, and Gallemore (2020), collected via a Twitter API. Additionally, we utilize the keyword *coronavirus* to search tweets from 22nd January 2020 to 3rd April 2020. By using this Twitter collecting method, we collect 11,303,850 tweets. These collected tweet data contained different languages, and we filter non-English tweets as we only consider English-related tweets. Besides, retweets are not useful in our study, and they are also filtered. After these two processing steps, the dataset contained 1,107,994 tweets. These tweets constitute our final experiment dataset and are used for Twitter text classification.

Next, we apply a geo-location filter step to resolve each tweet's geo-location into a US state-level and convert them into a uniform format. To do that, we first select tweets that either has a tweet location or Twitter profile location. Next, we select tweets with metadata that have geo-location across the US. Then, we geo-code these tweets at the state level. For example, we gather tweets with geo-location "Denver, CO," and converted them to "Colorado." In terms of the COVID-19 confirmed case data, we use the dataset provided by Johns Hopkins University (Dong, Du, & Gardner, 2020). Furthermore, we also apply the same geo-location resolver step to convert all daily COVID-19 confirmed case data with a US state level.

Table 1
Annotation Guidelines with Examples.

Annotation Guidelines	Example
Guideline 1: If the tweet indicates the user or someone in the user's community/city/state has been tested positive for COVID-19, this tweet should be considered 'COVID-19 related'.	"Not a hoax as #Trump said in SC rally. Illinois officials say patient has tested positive for #coronavirus https://t.co/lbgsYxqRo8"
Guideline 2: If the tweet indicates the user or someone in the user's community/city/state might know some suspected patients with COVID-19, this tweet should be considered 'COVID-19 related'.	"A letter sent to the @PlymouthSch community warns that a student who just got back from Italy last month was hospitalized with flu-like symptoms. We're tracking this potential case of #coronavirus on @boston25 at 5 and 6:30 https://t.co/yZAswzS2dd"
Guideline 3: A description of symptoms of COVID-19, such as cough, tiredness, pains were mentioned in the Twitter text. Then the tweet is considered to be 'COVID-19 related'.	"I'm sure I've got that #Coronavirus. Been in absolute tatters since about 4 pm yesterday. Hardly slept all night, coughing and sweating, and my head is totally throbbing. Not my most productive day so far. Literally getting up now to see if I can eat anything. Thanks, China!"

Table 2
The Attribute Description of the Dataset.

Attribute	Description
created_at	the creation time of the tweet
id_str	the unique identifier (id) of the tweet
state	the tweet's geo-location at US state level
full_text	the text content of the tweet
label	the result of text classification

Text pre-processing steps are then utilized to clean the tweet text dataset, and they include:.

- Convert all text to lower case.
- Remove punctuations and stop words.
- Apply lemmatization to convert all words to their basic form.
- Apply tokenization.

To conduct supervised learning text classification, we manually annotate 7064 geo-located tweets with three annotation guidelines shown in Table 1. The annotation guideline is confirmed by public health experts. Annotators are then instructed to label all tweets by following these annotation guidelines. The tweet is labelled as '1′ if it is 'COVID-19 related' and labelled as '0′ if it is 'COVID-19 unrelated'. Two authors annotate the whole dataset several times.

The statistics of the dataset used in the experiment are following:.

- 11,303,850 original tweets contain keywords *coronavirus* (between 22nd January 2020 to 3rd April 2020).
- 1,107,944 filtered tweets after the data pre-processing pipeline with geo-location at US state level (between 22nd January 2020 to 3rd April 2020).
- 7064 tweets with manually annotated for classifying 'COVID-19 related' tweet content.

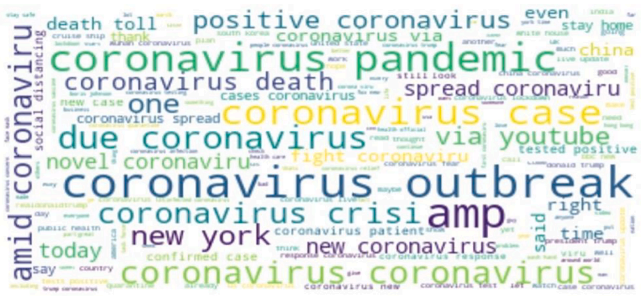


Fig. 6. Word cloud of the processed Twitter dataset.

- 72 days report on coronavirus confirmed case (between 22nd January 2020 to 3rd April 2020).

Besides, the main attributes of the dataset are shown in Table 2, which includes created_at, id_str, state, full_text, and label. We plot the word cloud of the processed Twitter dataset, which is shown in Fig. 6. It indicates strong concern about the COVID-19 outbreak. For example, 'coronavirus pandemic', 'coronavirus outbreak' are prominently shown in the word cloud.

4.2. *Twitter text classification*

4.2.1. *Experiment setup*

The experiments are performed on a workstation with an Intel CPU and a Tesla K80 GPU. All the experiments are implemented in Python. 70% of the annotated COVID-19 Twitter dataset is the training dataset, and 30% of the annotation dataset is the validation dataset. KNN and SVM are trained using the Scikit-learn toolkit (Fabian et al., 2011). For KNN, the hyperparameter K is set to 10. For SVM, we use linear kernel (LinearSVC). DPCNN is reproduced using PyTorch, and we choose stochastic gradient descent (SGD) as the optimization method, the learning rate is set as 0.0001, and the momentum is 0.949. BERT is implemented using TensorFlow, the batch size is 32, the learning rate is 1e-5, and the epoch is set at 16.

4.2.2. *Text classification model comparison*

To compare the evaluation performance of the Twitter text classification algorithms, KNN and SVM are chosen to be baseline as Modu et al. (2017) applied KNN and SVM for epidemic surveillance (Malaria outbreak warning system). And we also reproduce one CNN-based text classification DPCNN for comparison.

KNN is an algorithm that builds the model only consisting of storing the training data, and the algorithm finds the closest data points in the training set to predict the new data (Keller & Gray, 1985).

SVM is a kernel-based supervised machine learning technique. It assumes the data is linearly separable and aims to find a linear hyperplane (decision boundary) to separate the data. An SVM model represents the examples as points in space, mapped to divide the examples of the separate categories by a linear hyperplane (Boser et al., 1992).

DPCNN is a low-complexity word-level convolutional neural network used for text classification (Johnson & Zhang, 2017). The model requires sequentially embedded text as input. According to our limited dataset, we choose to use pre-trained unsupervised word vectors obtained with the Global Vectors for Word Representation (GloVe) algorithm (Pennington, Socher, & Manning, 2014).

4.3. *The proposed linear regression and early warning detection*

We use the Scikit-learn library for linear regression model implementation. Meanwhile, we randomly separate the dataset into training set and validation set with a ratio of 7:3. The validation set is used to validate the linear regression model's performance obtained with the training set to avoid overfitting. After implementing the Twitter-based linear regression model to forecast the future confirmed cases, we then utilize 30/100 confirmed cases to be the early outbreak threshold, which can inform the level of urgency in the state's warning. Zhang, Tao, Wang, Ong, Tang, Zou, Bai, Ding, Shen, Zhuang, and Fairley (2020) found that 30 confirmed cases are a critical threshold for a transition point from a slow to a fast-growing phase for COVID-19, and it is an essential early characteristic of the COVID-19 outbreak. This paper utilized the 30 confirmed cases as the outbreak threshold for partial US states. In another study by Gharavi, Nazemi, and Dadgostari (2020), the outbreak threshold for a US state is defined as 100 confirmed cases. Also, to address the substantial variation in the number of total populations in different states, we apply K-means clustering to cluster US states into two clusters by population. Then we assign the states in the cluster with a higher population with 100 confirmed cases as the outbreak threshold and the states in the cluster with a lower population with 30 confirmed cases as outbreak threshold (Hartigan & Wong, 1979). Hence, if the regression model predicts the confirmed case in a state that exceeds its threshold (30 or 100), then we assume this state will have an outbreak in the future. Moreover, we will generate an early warning message in the backend of the system. Besides, the frontend user interface (UI) will visualize this early warning message by plotting this state area in red.

Table 3
Binary Classification Confusion Matrix.

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	True positive (TP)	False positive (FP)
Predicted Negative Class	False negative (FN)	True negative (TN)

4.4. *Evaluation metrics*

4.4.1. *Twitter text classification*

The evaluation metrics utilized include Accuracy, Precision, Recall, and F1 score, which are commonly used in a binary classification problem (Mohammad & Md Nasir, 2015). A confusion matrix of the binary classification is shown in Table 3. If the result is a predicted positive class for an actual positive class, this is a true positive (TP). Otherwise, if the result is a predicted negative class for an actual positive class, it is a false negative (FN). If the result is a predicted positive class for an actual negative class, this is a false positive (FP). Otherwise, if the result is a predicted negative class for an actual negative class, it is a true negative (TN).

Accuracy (Eq. (3)) is the statistic measure of correct predictions of two classes over the total number of instances evaluated. Precision (Eq. (4)) is used to measure how many instances are correctly predicted from the total predicted instances in a positive class, and Recall (Eq. (5)) represents how many actual correct instances are identified correctly. F1 score (Eq. (6)) is the harmonic mean for Precision and Recall values.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

4.4.2. *Linear regression*

R^2 correlation coefficient is used to evaluate the performance of the linear regression model by measuring how well the independent variable explains the dependent variable in the linear regression model. It is defined as the following equation:

Table 4
Twitter classification results (The best performance is marked in bold font).

Algorithm	Precision	Recall	F1 score	Accuracy
KNN (Modu et al., 2017)	0.83	0.61	0.66	0.95
SVM (Modu et al., 2017)	0.82	0.67	0.72	0.95
DPCNN	0.20	0.75	0.32	0.94
Fine-tuning BERT	**0.96**	**0.99**	**0.98**	**0.99**

Table 5
The number of total parameters, model size, and run time in the four models. Run time is the inference time that measured over 2120 samples.

Model	Total parameters	Model size (memory)	Run time (ms)
KNN (k = 10)	N/A	1.43 MB	364.4
SVM (linear kernel)	N/A	0.75 MB	40.8
DPCNN	0.98 M	3.2 MB	29494.0
BERT (Base, Uncased)	110 M	1.22 GB	81957.7

$$R^2 = 1 - \frac{u}{v} \tag{7}$$

$$u = \sum \left(y_{true} - y_{predict}\right)^2 \tag{8}$$

$$v = \sum \left(y_{true} - y_{true\_mean}\right)^2 \tag{9}$$

where u denotes for the unexplained variation, v denotes for the total variation, $\mathrm{y_{true}}$ is the observed data, $\mathrm{y_{predict}}$ is the predicted value, and $\mathrm{y_{true\_mean}}$ is the mean of the observed data.

High R^2 score means the unexplained variation u is low, indicating that the linear model has good performance in explaining the dependent variable by a set of independent variables. However, $R^2 = 1$ is almost impossible with natural data due to random errors. Generally, $R^2 > 0.7$ is considered a strong prediction.

5. Results and discussion

5.1. Evaluation of the Twitter text classification algorithms

For the tweet text classification, Precision, Recall, F1 score and Accuracy are used as evaluation metrics discussed in Section 4.4.1. The size of the manually annotated dataset is 7064 tweets with 434 ‘COVID-19 related’ tweets and 6630 ‘COVID-19 unrelated’ tweets. The feature model used for KNN and SVM is Term Frequency-Inverse Document Frequency (TF-IDF). The DPCNN applies GloVe as the feature model. The tweet classification result is shown in Table 4. It indicates that our proposed fine-tuning BERT is the best model in terms of four evaluation metrics (Precision, Recall, F1 score, Accuracy). Fine-tuning BERT shows an Accuracy of 0.99, a Precision of 0.96, a Recall of 0.99, and an F1 score of 0.98, which outperforms previous algorithms. Compared to traditional machine learning methods (KNN, SVM) and CNN-based methods (DPCNN), our approach is clearly more adapted to the COVID-19 Twitter classification task.

5.2. Classification model complexity

Model complexity is an essential aspect for assessing the performance of machine learning models, especially deep learning models. Moreover, it is usually assessed by the number of parameters in the model, model size, and run time. A model’s size and number of parameters indicate how much space it takes up, while its run time indicates how fast it can make an inference. Table 5 shows the number of parameters, model size, and run time of the four models in our experiment. Run time is measured with the average of over 2120 samples and tested with the Telsa K80 GPU. KNN and SVM are traditional machine learning methods, and their model complexity is relatively low. KNN is simple, and the complexity of KNN depends on the hyperparameter K. In the experiment, SVM uses a linear kernel, and the complexity of SVM is tied to the number of support vectors. The latency and throughput grow linearly with the number of support vectors. More specifically, KNN has a 1.43 MB model size and a 364.4 ms run time. While SVM is smaller and faster, it has a 0.75 MB memory size and took 40.8 ms to infer 2120 samples. DPCNN and BERT are deep learning models. In the experiment, we reproduce the DPCNN that has 15 weight layers, and we apply BERT-Base (uncased) for fine-tuning. The BERT-Base has 12 transformer layers, and the hidden embedding size is 768. Compared with traditional machine learning models, DPCNN and BERT have a larger size, a larger number of parameters, and a longer run time. It can be seen that DPCNN has 0.98 M parameters, 3.2 MB model size, and 29494.0 ms run time. BERT (base, uncased) has 110 M parameters, the model size is 1.22 GB, and the run time is 81957.7 ms. As BERT is pre-trained on a large corpus (Wikipedia and BookCorpus) and its transformer architecture, it is the most complicated model among the four models.

Table 6
State level prediction model with Twitter data (seven days prior) results.

Prediction model (predict seven days prior) + Twitter data	Pseudo R^2
California Linear Regression Model	0.842
Oregon Linear Regression Model	0.946
Massachusetts Linear Regression Model	0.889

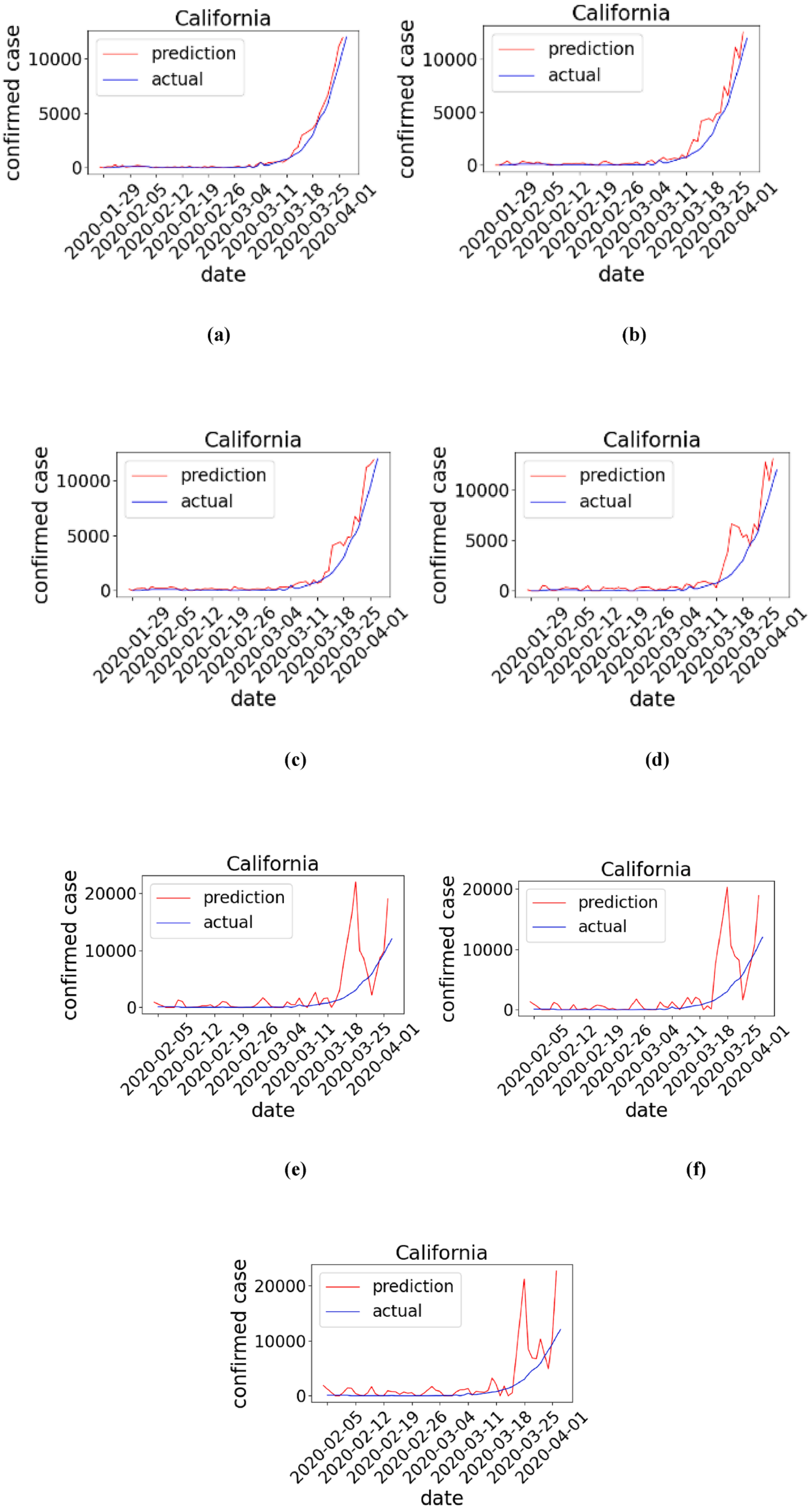
Table 7
US level prediction model with Twitter data results.

Prediction Model	Pseudo R^2
US Linear Regression Model (predict one day prior)	0.977
US Linear Regression Model (predict two days prior)	0.979
US Linear Regression Model (predict three days prior)	0.977
US Linear Regression Model (predict four days prior)	0.979
US Linear Regression Model (predict five days prior)	0.789
US Linear Regression Model (predict six days prior)	0.909
US Linear Regression Model (predict seven days prior)	0.621

5.3. Twitter-based linear regression results

As for the linear regression model, we choose contiguous seven-day data with three variables (confirmed cases per state, tweet counts per state per day, and the positive classified tweet counts per state per day) to predict future confirmed cases for every state in the US. The data on confirmed cases are from Johns Hopkins University. The tweet data, which contain tweet counts data, and positive classified tweet data, are from the experiment results of fine-tuning BERT. We train several linear regression models using only its state data and its validation set pseudo $\mathbf{R}^2$ score results are shown in Table 6. However, the limitation of the state-level linear regression model is overfitting, as the training dataset is small. To achieve a better generalization performance and avoid overfitting, we train a Twitter-based US-level linear regression model that uses all states’ data. The result of the US-level pseudo $\mathbf{R}^2$ scores of the validation set are shown in Table 7. From Table 7, the pseudo $\mathbf{R}^2$ scores for prediction one to four days ahead and six days ahead are all above 0.9 in the validation set, which validates the effectiveness of our regression model.

Though the US-level regression model pseudo $\mathbf{R}^2$ score is similar, the result for two days prior prediction has a better performance than the rest. It implies the accuracy will be lower when using a linear regression model to predict a more extended range. However, according to the results, prediction for six days prior with 0.909 pseudo $\mathbf{R}^2$ in the validation is generally acceptable. After balancing the prediction period and forecasting performance, prediction for six days prior linear regression



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 7. Visualization of prediction for California. (a) one day prior (b) two days prior (c) three days prior (d) four days prior (e) five days prior (f) six days prior (g) seven days prior. Red denotes prediction, and blue denotes the true values.

Table 8
Twitter classification results with over-sampling.

Algorithm	Precision	Recall	F1 score	Accuracy
KNN (Modu et al., 2017)	0.73	0.91	0.79	0.93
SVM (Modu et al., 2017)	0.92	0.91	0.92	0.98
DPCNN	0.30	0.70	0.42	0.95
Fine-tuning BERT	0.96	0.99	0.98	0.99

Table 9
State-level prediction model without Twitter data (seven days prior) results.

Prediction model (predict seven days prior)	Pseudo R^2
California Linear Regression Model	0.976
Oregon Linear Regression Model	0.694
Massachusetts Linear Regression Model	0.694

model is used as the prediction model to detect early warnings of the outbreak. The visualization of prediction using the US level linear regression model for California ranges from one day to seven days prior are shown in Fig. 7.

5.4. Ablation study

As the class distribution of the annotated dataset is imbalanced ('COVID-19 related' class accounting for 6.14% and 'COVID-19 unrelated' class accounting for 93.86%), it may lead the classification algorithm to have a bias towards the majority class. To determine how the imbalanced Twitter data impact the classification result, we conduct experiments using the over-sampling approach. The general idea of over-sampling is to increase the size of the minority class and get a balanced class dataset. The tweet classification results with over-sampling are shown in Table 8.

By applying the over-sampling approach and compared the results with Table 4, we show that the over-sampling can slightly improve the performance of the classification algorithms. It improves the F1 score of KNN from 0.66 to 0.79 and improves the SVM F1 score from 0.72 to 0.92. Moreover, it improves DPCNN F1 score from 0.32 to 0.42. However, the over-sampling has the disadvantage that it may cause the overfitting issue. Among these four classification algorithms, BERT performs best and achieving an F1 score of 0.98, Precision of 0.96, Recall of 0.99, and Accuracy of 0.99. We can conclude that the imbalanced dataset has a more significant impact on the traditional machine learning algorithms and less on the pre-trained NLP model BERT. And fine-tuning BERT can achieve significant performance on the imbalance dataset.

Moreover, in terms of the regression model, we also evaluate the impact of the Twitter data for the model. We train these state-level linear regression models without Twitter data (tweet counts per state per day and the positive classified tweet counts per state per day). The results of the state-level prediction model without Twitter data are shown in Table 9. Compared with the results in Table 6, it shows that some states, such as Oregon and Massachusetts, have lower Pseudo $\mathbf{R}^2$ scores, indicating that Twitter data are helpful to the prediction model.

5.5. Early warning detection results

In the experiment, K-means clustering is used to cluster US states according to their population into two clusters. We implement the K-means clustering using the Python Scikit-learn library (Pedregosa Odlum & Yoon, 2011). Then we designate the state with the higher population cluster to apply 100 confirmed cases as the outbreak threshold and designated the state with the lower population cluster to apply 30 confirmed cases as the outbreak threshold. The higher population cluster states include four states, California, Texas, Florida, and New York. Moreover, the rest of the states are in the cluster with a lower population.

Table 10
Early warning detection results: the early warning detection date is the date that the system sends an early warning message (six days before the predicted outbreak date), the predicted outbreak date is the date that the linear regression model forecasting the COVID-19 confirmed cases exceeds the outbreak threshold.

State	Early warning detection date	Predicted outbreak date
California	2020.01.30	2020.02.05
Colorado	2020.02.24	2020.03.01
Washington	2020.01.30	2020.02.05
New York	2020.02.04	2020.02.10
Florida	2020.02.04	2020.02.10

Using the 30 and 100 confirmed cases as the early outbreak threshold, if the regression model predicts the number of confirmed cases in a state exceeds the outbreak threshold, we project that the state may encounter an outbreak. The system will send the early warning message when the prediction result exceeds the outbreak threshold. The frontend of the system will visualize this early warning message to the users. As indicated in Table 7, the linear regression model that predicts a six-day prior has a good performance, and we apply it to detect early warning of the COVID-19 outbreak.

The predicted outbreak date is when the model predicts the state will exceed the outbreak threshold, and the early warning detection date is when the system will send early warning signals. For example, the prediction model predicts that California exceeds its outbreak threshold on 5 February 2020. As it predicted six days prior, the system would generate the early warning signal six days before 5 February 2020, i.e., on 30 January 2020. The partial results of the early warning detection dates using a six-day prior linear regression model are presented in Table 10. The results provide evidence of our early warning system's effectiveness in providing early warning signs of the COVID-19 outbreak in the US.

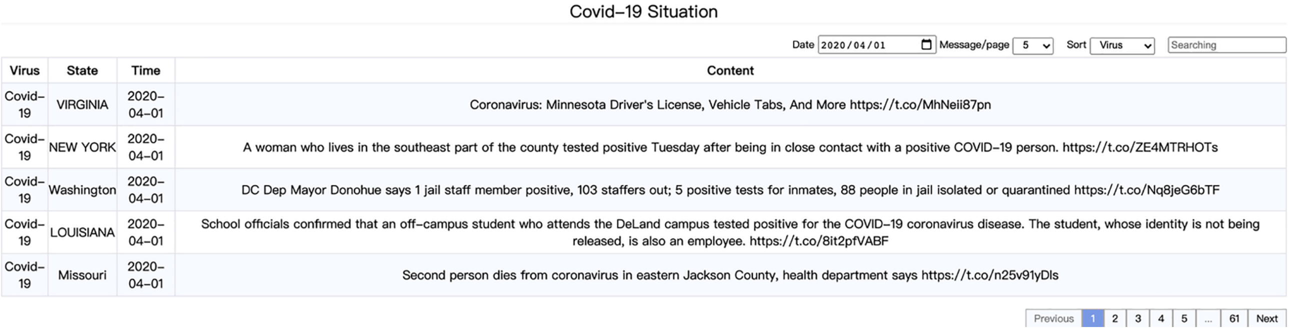
5.6. Intelligent early warning system

A simple demonstration of our early warning system is available online (https://nbreservoir.com/warning/visualization.html). The frontend UI of the system consists of a table component and a map component. The table stores the classified Twitter text by applying fine-tuning BERT (Fig. 8.). Users can interact with the web application by selecting the data of a specific date to view. Also, users can set the rows of tables (five rows, ten rows, fifteen rows) displayed on the webpage, and the table information can be sorted by state name, time and Twitter content. Also, users can search the information through the keywords in the "Searching", then the information in the table will update instantly according to the user's searching input data.

The mapping component of the web application visualizes the US at the state level, and each state shows corresponding tweet classification information (Fig. 9.). If the state in the US map is rendered in red, it indicates that it has an 'early warning' level, and the system predicates it will have an outbreak in this state in the future. When the user clicks the state on the map, it will pop a small window with corresponding tweet information. Users can change the page by clicking Pre or Next. Moreover, clicking the state name will show a prediction chart of COVID-19 development of the current state to users (Fig. 10.). The y-axis of the prediction chart represents the prediction of the COVID-19 confirmed cases in the current state, and the x-axis of the prediction chart represents the date. If the predicted number of confirmed cases exceeds the outbreak threshold, the chart's line will be highlighted in red. Users can also click the download button over the chart to download the current chart.

6. Conclusions

In this paper, we construct a Twitter dataset with geo-location that



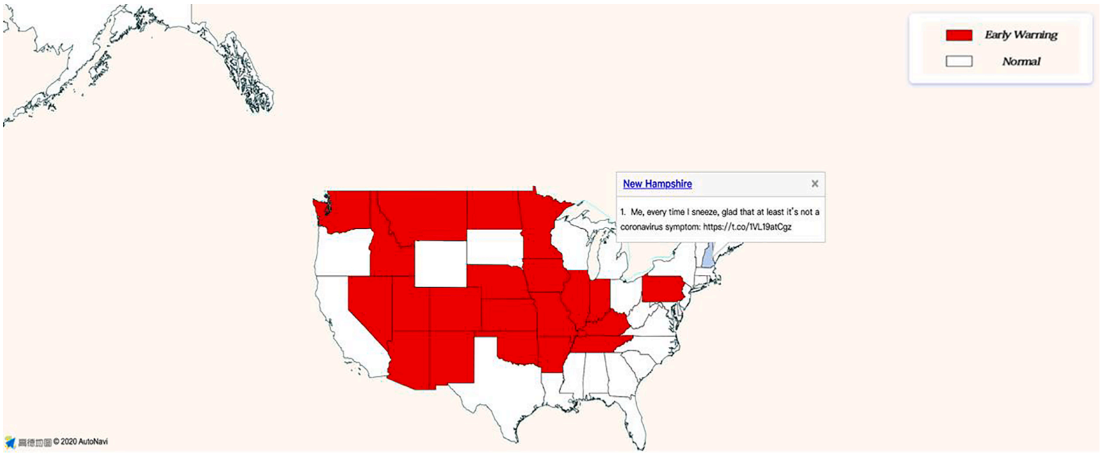
Covid–19 Situation

Date 2020/04/01 Message/page 5 Sort Virus Searching

Virus	State	Time	Content
Covid–19	VIRGINIA	2020–04–01	Coronavirus: Minnesota Driver's License, Vehicle Tabs, And More https://t.co/MhNeii87pn
Covid–19	NEW YORK	2020–04–01	A woman who lives in the southeast part of the county tested positive Tuesday after being in close contact with a positive COVID–19 person. https://t.co/ZE4MTRHOTs
Covid–19	Washington	2020–04–01	DC Dep Mayor Donohue says 1 jail staff member positive, 103 staffers out; 5 positive tests for inmates, 88 people in jail isolated or quarantined https://t.co/Nq8jeG6bTF
Covid–19	LOUISIANA	2020–04–01	School officials confirmed that an off–campus student who attends the DeLand campus tested positive for the COVID–19 coronavirus disease. The student, whose identity is not being released, is also an employee. https://t.co/8it2pfVABF
Covid–19	Missouri	2020–04–01	Second person dies from coronavirus in eastern Jackson County, health department says https://t.co/n25v91yDls

Previous 1 2 3 4 5 ... 61 Next

Fig. 8. Table component UI in the early warning system.



New Hampshire

1. Me, every time I sneeze, glad that at least it's not a coronavirus symptom: https://t.co/1VL19atCgz

Fig. 9. Map component UI in the early warning system.

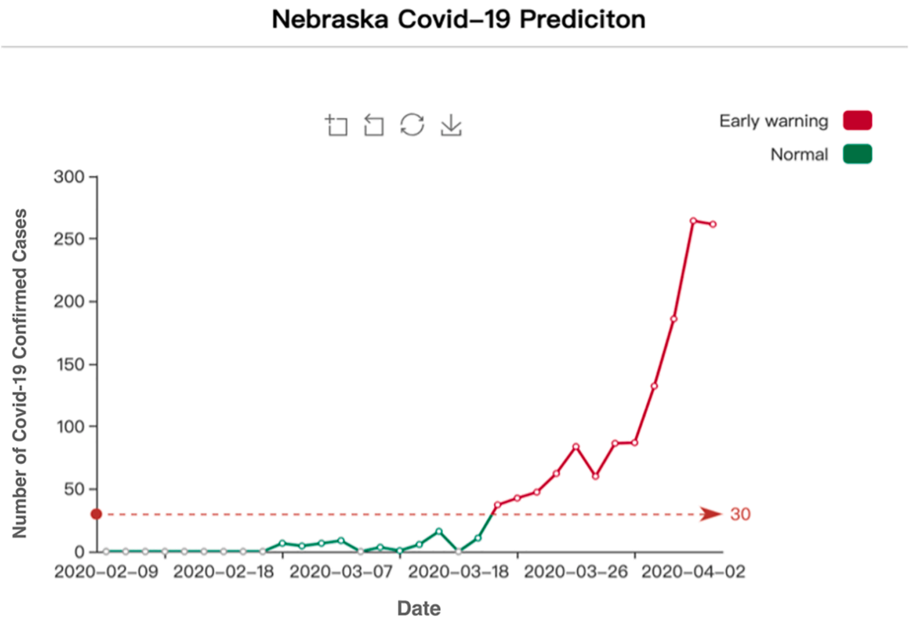


Fig. 10. Prediction chart UI in the early warning system.

can be used for COVID-19 surveillance. We also employ fine-tuning BERT as a tweet classification method and achieve a 0.98 F1 score on the COVID-19 Twitter dataset. To our best knowledge, this is the first work on epidemic surveillance using BERT. Another notable contribution of this paper is to propose a Twitter-based linear regression model to detect early warning of the outbreak. We also deploy an early warning system web application that has visualized these results and allowed users to interact in the web app. The evaluation of our results has indicated that it is possible to apply Twitter to achieve COVID-19 surveillance in the US and predict in advance to provide helpful support for health departments and public health officials to make decisions.

Our work also has at least two limitations. First, we only conduct the experiments with Twitter data, but many types of data may be involved in the study to improve the performance. Second, the size of the annotation dataset for text classification is relatively small, and more data are still needed for future study.

For the future work directions that could be investigated, we plan to involve other data types such as Google search data and other social media data for our COVID-19 surveillance. Furthermore, we can collect Twitter data in an extended period and annotate additional data to increase the size of the dataset.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research is supported by the NCHI Grant NCHI-I01200300015.

References

Chen, L., Tozammel Hossain, K. S. M., Butler, P., Ramakrishnan, N., & Prakash, B. A. (2016). Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery, 30*(3), 681–710. https://doi.org/10.1007/s10618-015-0434-x

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. *ArXiv Preprint.* http://arxiv.org/abs/1007.4748.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics. *Human Language Technologies – Proceedings of the Conference, 1*(Mlm), 4171–4186.

Di Martino, S., Romano, S., Bertolotto, M., Kanhabua, N., Mazzeo, A., & Nejdl, W. (2017). Towards Exploiting Social Networks for Detecting Epidemic Outbreaks. *Global Journal of Flexible Systems Management, 18*(1), 61–71. https://doi.org/10.1007/s40171-016-0148-y

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. In *The Lancet Infectious Diseases* (Vol. 20(5, pp. 533–534). Lancet Publishing Group. https://doi.org/10.1016/S1473-3099(20)30120-1.

Fabian, P., Varoquaux, G., Cournapea, A. G., Vincent, M., Thirion, B., Olivier, G. D., … Passos, A. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12(85)x*, 2825–2830.

Freedman, D. A. (2009). *Statistical models: Theory and practice*. cambridge University Press.

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association, 15* (2), 150–157. https://doi.org/10.1197/jamia.M2544

Gharavi, E., Nazemi, N., & Dadgostari, F. (2020). Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: COVID-19 a Case Study in the US. *ArXiv Preprint.* http://arxiv.org/abs/2005.00475.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457* (7232), 1012–1014. https://doi.org/10.1038/nature07634

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *Proceedings of the 3rd International Web Science Conference, WebSci 2011*. 10.1145/2527031.2527049.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics, 28*(1), 100. https://doi.org/10.2307/2346830

Ji, X., Chun, S. A., & Geller, J. (2012). Epidemic outbreak and spread detection system based on twitter data. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-642-29361-0_19

Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. *ACL 2017 – 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 562–570. 10.18653/v1/P17-1052.

Jordan, S. E., Hovet, S. E., Fung, I. C. H., Liang, H., Fu, K. W., & Tse, Z. T. H. (2019). Using twitter for public health surveillance from monitoring and prediction to public response. *Data, 4*(1), 1–20. https://doi.org/10.3390/data4010006

Keller, J. M., & Gray, M. R. (1985). A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Transactions on Systems, Man and Cybernetics, SMC-15(4)*, 580–585. https://doi.org/10.1109/TSMC.1985.6313426

Lee, K., Agrawal, A., & Choudhary, A. (2013). Real-time disease surveillance using Twitter data. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1474–1477. 10.1145/2487575.2487709.

Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., … Chen, W. (2018). ECharts: A declarative framework for rapid construction of web-based visualization. *Visual Informatics, 2*(2), 136–146. https://doi.org/10.1016/j.visinf.2018.04.011

Lopez, C. E., Vasu, M., & Gallemore, C. (2020). Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. *ArXiv Preprint.* http://arxiv.org/abs/2003.10359.

Marcus, A., Bernstein, M. S., Badar, O.,x Karger, D. R., Madden, S., & Miller, R. C. (2011). TwitInfo: Aggregating and visualizing microblogs for event exploration. *Conference on Human Factors in Computing Systems – Proceedings*, 227–236. 10.1145/1978942.1978975.

Masri, S., Jia, J., Li, C., Zhou, G., Lee, M. C., Yan, G., & Wu, J. (2019). Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. *BMC Public Health, 19*(1), 1–14. https://doi.org/10.1186/s12889-019-7103-8

Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., … da Silva Sousa, L. (2016). Tracking dengue epidemics using twitter content classification and topic modelling. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-46963-8_7

Modu, B., Polovina, N., Lan, Y., Konur, S., Taufiq Asyhari, A., & Peng, Y. (2017). Towards a predictive analytics-based intelligent malaria outbreakwarning system. *Applied Sciences (Switzerland), 7*(8), 1–20. https://doi.org/10.3390/app7080836

Mohammad, H., & Md Nasir, S. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Odlum, M., & Yoon, S. (2015). What Can We Learn about the Ebola Outbreak from Tweets? *American Journal of Infection Control, 176*. https://doi.org/10.1016/j.ajic.2015.02.023.What

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases, 47*(11), 1443–1448. https://doi.org/10.1086/593098

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLoS Computational Biology, 11*(10), 1–15. https://doi.org/10.1371/journal.pcbi.1004513

Santos, J. C., & Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling, 11*(SUPPL.1), 1–11. https://doi.org/10.1186/1742-4682-11-S1-S6

Şerban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing and Management, 56*(3), 1166–1184. https://doi.org/10.1016/j.ipm.2018.04.011

World Health Organization. (2020). WHO Coronavirus (COVID-19) Dashboard. *WHO.* https://covid19.who.int/.

Yousefinaghani, S., Dara, R., Poljak, Z., Bernardo, T. M., & Sharif, S. (2019). The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study. *Scientific Reports, 9*(1), 1–17. https://doi.org/10.1038/s41598-019-54388-4

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92*, 144–152. 10.1145/130385.130401.

Zhang, L., Tao, Y., Wang, J., Ong, J. J., Tang, W., Zou, M., Bai, L., Ding, M., Shen, M., Zhuang, G., & Fairley, C. K. (2020). Early characteristics of the COVID-19 outbreak predict the subsequent epidemic scope. *International Journal of Infectious Diseases*, *97*, 219–224. x10.1016/j.ijid.2020.05.122.