Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
COVID-19 personal health mention detection from tweets using dual convolutional neural network

Linkai Luo a, Yue Wang a,⁎, Hai Liu b

a Department of Supply Chain and Information Management, The Hang Seng University of Hong Kong, Hong Kong Special Administrative Region
b Department of Computing, The Hang Seng University of Hong Kong, Hong Kong Special Administrative Region

ABSTRACT

Twitter offers extensive and valuable information on the spread of COVID-19 and the current state of public health. Mining tweets could be an important supplement for public health departments in monitoring the status of COVID-19 in a timely manner and taking the appropriate actions to minimize its impact. Identifying personal health mentions (PHM) is the first step of social media public health surveillance. It aims to identify whether a person’s health condition is mentioned in a tweet, and it serves as a crucial method in tracking pandemic events in real time. However, social media texts contain noise, many creative and novel phrases, sarcastic emoji expressions, and misspellings. In addition, the class imbalance issue is usually very serious. To address these challenges, we built a COVID-19 PHM dataset containing more than 11,000 annotated tweets, and we proposed a dual convolutional neural network (CNN) framework using this dataset. An auxiliary CNN in the dual CNN structure provides supplemental information for the primary CNN in order to detect PHMs from tweets more effectively. The experiment shows that the proposed structure could alleviate the effect of class imbalance and could achieve promising results. This automated approach could monitor public health in real time and save disease-prevention departments from the tedious manual work in public health surveillance.

1. Introduction

The outbreak of COVID-19 has caused an unprecedented impact on society (da Silva et al., 2021). It is predicted that the virus will continue to mutate and spread, and the whole world will remain in pandemic mode for a long time (Scudellari, 2020). In such a situation, it is critical to detect the outbreak of COVID-19 in a region as early as possible and take the appropriate action to minimize its impact. Various public health surveillance and epidemic intelligence systems have been implemented in different countries by the Centers for Disease Control (CDC) for early detection of large-scale contagious disease outbreaks. The system continuously collects formal public-health related information, such as hospital statistics, medical records, health encounters, disease registries, laboratory results, surveys, etc.; however, these traditional public health surveillance methods entail long processes of data collection, data validation, and data analysis. Thus, the system cannot respond timely to rapid-spreading diseases. As a result, such surveillance falls short of monitoring, assessing, and forecasting the trajectory of the spread of the disease. The prolonged response time may cause serious consequences (Paul and Dredze, 2017).

Due to the popularity of social media, a large amount of health-related information is shared online (Zheng et al., 2021; Rivadeneira et al., 2021). It has been reported that over 70% of adults in the USA use social media, and its popularity is increasing (Weissenbacher et al., 2019). Platforms such as Twitter provide real-time coverage of ongoing events that are dynamically and locally fluctuating, like the COVID-19 pandemic. As a result, data on Twitter offers extensive, valuable information on the current state of public health and the spread of disease in a region. In public health monitoring research, data from tweets has been an important source for early infectious disease detection, warnings, and interventions, including for cholera (Chunara et al., 2012), E. coli (Diaz-Aviles and Stewart, 2012), bubonic plague (Da’ar et al., 2016), seasonal conjunctivitis (Deiner et al., 2016), and Ebola (Joshi et al., 2020). The World Health Organization (WHO) has also begun to emphasize collecting epidemic information through social media. The WHO states that early detection can be found through social media data for more than 60% of epidemics (Joshi et al., 2020). Thus, as an important supplement to traditional health surveillance, social media monitoring can assist early detection of public health emergencies. The practice also has the potential to reduce the cost and speed of information acquisition and
A critical initial step for online health surveillance is to identify personal health mentions (PHM) in social media data. PHMs in posts indicate that either the poster or someone the poster knows is experiencing a health problem or relevant symptoms (Lamb et al., 2013). PHMs are usually classified into four categories: self-mention, other-mention, attention, and non-health. For example, the post “the cough is killing me” contains a self-mention PHM, as it indicates that the poster’s health condition. The post “cough is a symptom of COVID-19” is considered an awareness PHM, as it involves general information of COVID-19 and does not indicate a person’s health condition. “Corona s#*ls, drink German beer” is non-health PHM, as ‘Corona’ in this tweet means the beer brand, not the virus. Among the four classes, self-mention and other-mention PHMs are of particular interest, as they reflect the true public health condition. The personal health-relevant posts are fed into a smart epidemic informatics system to estimate the disease’s spatio-temporal spread, projected peak time, severity of outbreak, and duration of outbreak. Thus, accurately detecting PHMs is important and serves as a critical first step in online public health surveillance.

Academia also recognizes the importance of PHM identification on social media. The Social Media Mining for Health (SMM4H) Shared Task was first organized in 2016 at the flagship NLP conference, the Annual Conference of Association for Computational Linguistics (ACL). In 2019, the task of automatic PHM classification from tweets was added as a new task. Participants of the task classify influenza-related PHM tweets into two categories: being sick or vaccinated (Weissenbacher et al., 2019). During the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), a new shared task of identifying informative COVID-19 tweets was organized (Nguyen et al., 2020). The task can be considered a simple version of PHM identification, as each tweet must be classified as relevant to COVID-19 or not.

However, some challenges still need to be overcome in order to detect PHMs from tweets and other social media data. First, the tweet data are in free-text form; they are unstructured and subject to the variability of natural language. Informal and creative expressions, including emojis and idiomatic expressions, are very common in tweets. In addition, figurative or metaphorical expressions are also widely used (e.g., “When Paris sneezes, all Europe catches a cold”). Second, although a massive amount of tweets discuss the COVID-19 pandemic, only a small portion of them are self-mention and other-mention PHMs. However, in social media public health surveillance, it is of particular importance to identify these two types of PHMs, as they represent the actual disease-related health conditions. Additionally, the issues of class imbalance and data sparsity of annotated tweets pose great challenges to detecting disease-relevant PHMs, as existing machine learning methods usually perform worse on classes with fewer training samples. Finally, tweets are short text, which usually contain less information and lack sufficient contextual information to develop a data-driven approach that understands the complicated utterances in the text. Due to these factors, it has been acknowledged that the normal natural language processing methods perform worse with tweets than with standard texts.

To address these challenges, we propose a dual convolutional neural network (CNN) structure for PHM identification. We build a COVID-19 PHM dataset that contains 11,231 annotated tweets with the four PHM classes. To the best of our knowledge, this is the first COVID-19 dataset of this type; it is available at https://github.com/ym7721/PHM_COVID19. We hope that it can enable other researchers and promote research in this direction. Furthermore, we model the PHM identification as a text classification task. A CNN-based network structure is used to classify tweets into the four PHM classes, as CNN has been found to perform better in short-text classification tasks (Yu et al., 2020). In addition, a dual CNN structure is used to mitigate the class imbalance issue. The network consists of a primary CNN (P-Net) that performs the PHM identification task and an auxiliary CNN (A-Net). The A-Net categorizes the tweets into the dominating class, which has the most samples, and the dominated class, which is the combination of the other classes. The A-Net is combined with the P-Net to provide the information of whether the inputted tweets are from the dominating class or not. The experiment results show that the proposed network structure can significantly improve the performance of PHM identification with a small sample size. Thus, the dual network structure provides a more effective method for social media health surveillance.

2. Relevant work

Social media data has been examined in relation to health monitoring and epidemic intelligence (Barnes et al., 2021). The research method can be generally classified as knowledge-driven and data-driven. Since building and annotating a PHM dataset is costly and time-consuming, knowledge-driven methods were among the methods in initial exploration of health surveillance based on social media data. Knowledge-driven approaches rely on a knowledge base or ontology to provide relevant health and medical information. A knowledge base typically contains the relationship between medical entities (e.g., diabetes is a cause of heart disease, and diarrhea is a symptom of COVID-19). It also includes medical rules extracted from experience to predict health events of interest. Using information extraction techniques, Collier et al. (2020) built a medical knowledge base, BioCaster. They extracted and classified health-related news into different health domains. The techniques of named entity recognition and semantic role labeling are applied to extract the relationship between entities in BioCaster. In the following steps, a BioCaster-based system was used to forecast emerging public health events by monitoring news using the relevant keywords. Huang et al. (2016) deployed a medical knowledge base to predict whether a tweet contains relevant entities, using the associative relationship between medical concepts to map the extracted entities to the features of a classifier with which new illness tweets can be identified.

With the emergence of massive amounts of medical-related social media data, data-driven approaches have prevailed in identification of PHMs in tweets. It has been acknowledged that tweets’ contents are good indicators of personal health reports and have been successfully used as syntactic and lexical features (Lamb et al., 2013). Specifically, most data-driven approaches apply machine learning algorithms for PHM identification. Traditional methods (non-deep learning-based methods) usually consist of feature extraction and training classifiers for PHM identification. Features are usually elicited manually or semi-automatically using medical expertise. Paul and Dredze (2011) proposed the Aspect Topic Ailment Model (ATAM) to discover topics related to diseases of interest. ATAM consolidates the symptoms and the corresponding treatments into various topics. The combination of keywords and associated topics is then utilized to identify the disease from tweets. The proposed model was initially used for flu PHM detection but was later extended to other diseases. Using a similar idea to ATAM, Chen et al. (2016) proposed the Hidden Flu-State Tweet Model (HFSTM) to detect flu mentions from tweets. The model uses tweet-level symptom variables to construct health-related topics from the text corpus. It is capable of classifying the person’s health condition into the categories of ‘healthy’, ‘exposed’, and ‘infected’ based on the tweets that the person posted. Jiang et al. (2016) trained classifiers using decision trees, KNN, and MLP to predict whether a tweet contains personal health experience. Various features, such as emotion keywords and user mentions, are used as features to learn the classifier. Genualdo et al. (2013) examined tweets to detect influenza-like illnesses. Since health mentions in tweets are in layman’s terms, the researchers related all the layman expressions to the technical jargon related to influenza, which are defined by the CDC in European countries. Then, they trained a model based on the data pairs of jargon-layman terms to detect influenza-like illness mentions on internet. Besides the aforementioned features, other typical features of
these approaches include unigram and bigram features extracted from tweets (Aramaki et al., 2011), and semantic frame features (Chapman et al., 2005). The classifiers used include AdaBoost, Naïve Bayes, and support vector machine (Olszewski, 2003; Aramaki et al., 2011). In summary, this group of methods requires human-engineered features that depend on expertise. Some researchers have argued that such feature engineering requiring much human intelligence can be challenging, due to the various formats of PHMs in tweets.

Researchers have also explored deep learning-based approaches to identify PHMs in tweets. Unlike traditional machine learning approaches, the human-involved feature engineering can be skipped. The token sequences in the tweets are encoded with word embeddings and inputted to a neural network structure to identify PHMs. Jiang et al. (2018) used pre-trained word embeddings to encode tweets. A long short-term memory network (LSTMs) takes the embeddings as input and output, whether the tweets are health mentions or not. Using a similar idea, Wang et al. (2017) used GloVe-based word embeddings to encode the words in tweets and fed them to a bidirectional LSTM (BiLSTM) network to classify influenza-related tweets. Some researchers have also put effort into understanding the metaphorical or figurative expressions in tweets in order to better identify PHMs. Iyer et al. (2019) proposed a method to determine whether disease keywords were used figuratively. In this method, the output is passed to a CNN to detect whether the corresponding tweets contain PHMs or not. However, the authors noted that the figurative features cannot improve the performance of PHM identification without other features like word embeddings, especially for widely-used figurative expressions of disease like ‘heart attack’. Biddle et al. (2020) leveraged sentiments to detect figurative health mentions. The words in a tweet are encoded with word embeddings, which are fed into a BiLSTM and a sentiment detection module, respectively. The outcomes of BiLSTM and sentiment modules are concatenated to pass a softmax classifier, which indicates whether the inputted tweets contain figurative health mentions. However, the authors noticed that the model struggles to extract the complete contextual information of the disease word, and the overall performance is not improved. To better encode the contextual information in tweets, context-based word embeddings have also been investigated. Joshi et al. (2019) found that context-based representation performs better than word-based representation in identifying influenza classification and PHM detection.

To summarize, the research on PHM detection from tweets has drawn much attention. However, no research has been conducted on COVID-19 PHM detection, due to the lack of annotated datasets in the widely-accepted four classes of PHMs. In addition, COVID-19 PHMs are more unbalanced than other diseases, due to the massive amount of tweets in the ‘awareness’ category. Thus, previous methods may not fit well to the scenario of the research topic studied in this paper.

3. PHM identification based on dual CNN model

3.1. Problem definition

Let \( T \in \mathcal{T} \) be a tweet in a tweet space \( \mathcal{T} \). Let \( C = \{c_1, c_2, ..., c_3\} \) be the set of labels corresponding to the four PHM classes—i.e., self-mention, other-mention, awareness, and non-health. Let \( D \) be the training set, which consists of the data pairs of tweets and the corresponding PHM labels—i.e., \( (T, c) \in \mathcal{T} \times C \). For example, one data pair in the training set is \( <”had my COVID 19 nasal swab Saturday. Got a call last night from CDC test was positive!”, self-mention> \). The identification of PHMs in tweets is modeled as a text classification task. Given the training set \( D \), we seek to learn a model \( f : \mathcal{T} \rightarrow C \), a mapping from the tweet space to the label space. The mapping can be used to determine the label of a new tweet text \( T^* \) through the application of \( f(T^*) \).

3.2. Dual CNN structure

We adopt a deep learning-based approach to learn the mapping from the tweet domain to the label domain, due to the approach’s superior capability to text-extract the semantic information from text. Complicated pre-trained language models like BERT have accomplished state-of-the-art performance in many NLP tasks; however, they are resource-intensive and require a large amount of training data. The annotated tweets’ set size is moderate (11,231), and the tweets’ lengths are not long; thus, the complicated language model may not function well. This paper plans to use CNN models, due to their strong performance in extracting local and global contextual information from text (Kim, 2014; Wang et al., 2020).

In addition, the distribution of the tweets among the four classes are highly imbalanced in nature, as only a small portion of Twitter users share COVID-19-related health mentions of themselves and the people they know. In machine learning research, it has been acknowledged that the class imbalance issue significantly affects the reliability and quality of the results of machine learning tasks. The classifier trained using imbalanced data sets tends to favor classes with large number of training samples (Wang and Li, 2021). For example, in the PHM identification task, the samples in self-mention class are more likely to be classified into awareness class, as sample size of awareness class is significantly larger than the other three classes.

To tackle this issue, we propose a dual CNN structure that consists of a primary network (P-Net), which classifies the tweets into four desired PHM classes. Since awareness-class PHM samples are the greatest contributor to the whole PHM dataset, we further consolidate the annotated PHM dataset and combine the classes of self-mention, other-mention, and non-health in order to balance the distribution of sample size among the classes. An auxiliary network (A-Net) is trained using the consolidated data set to classify tweets into two classes: a majority class (i.e., the awareness class) and the remaining class. With the more balanced data set, the samples in self-mention class are less likely to be classified into awareness class. The output of A-Net is linked with P-Net to provide extra information on the data distribution among the classes. It should be noted that the P-Net and the A-Net have the same network structure in this paper, but they use different training sets and perform different classification tasks.

To summarize, the A-Net could alleviate the classification bias caused by class imbalance. The function of the A-Net is to balance the P-Net such that the P-Net will not be biased to the majority class. The structure of the dual CNN is shown in Fig. 1. The details of each layer in the structure are elaborated as follows.

- **Embedding layer (part 1 in Fig. 1)**

Word embeddings-based encoding has prevailed in NLP research in transforming words into lower-dimensional numerical vectors. This dense representation not only reduces the computational cost but also encodes semantic information in words. Currently, many pre-trained word embeddings are available, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). This paper uses 300-dimension GloVe to encode the words in tweets. For example, the word *apple* can be encoded into a 300-dimension vector \( \langle 0.52042001, -0.83139998, 0.49961001, 1.28929996, 0.1151, 0.057521, ... \rangle \). In this manner, each word’s embedding can be obtained using GloVe.

Let \( T = (t_1, t_2, ..., t_n) \) be a tweet where \( t_i \) is the i-th word and \( n \) is the number of words in the tweet. Let \( v_0 \) be the embedding vector of \( t_i \) using GloVe. By concatenating all the words’ embedding vectors, we can get the embedding matrix of tweet \( V: V = v_1 \oplus v_2 \oplus ... \oplus v_n \in \mathbb{R}^{m} \), where \( m \)
is the dimension of word embeddings. The procedure can be illustrated by Fig. 2. In this research, we set the length of all the tweets to be 64 words. If a tweet has fewer than 64 words, we paddle the sequence with 0 so that all the tweets have the same length. Thus, in this paper, \( m = 300 \) and \( n = 64 \).

- Convolutional layer (part 2 in Fig. 1)

A convolutional layer takes the embedding matrices as the input to extract the high-level semantic features from the text. We use convolutional kernels—that is, a matrix to perform the convolutional operation—with different sizes, to capture the different kinds of features. The width of the kernel is set to be \( n \), which is the same as the embedding dimension, and the height of the kernel is \( h \). Given the kernel \( \omega \in \mathbb{R}^{h \times n} \), the feature obtained from the convolution operation between the kernel and the inputted embedding matrix \( V \) is calculated by \( s_i = f(\omega^T v_{i:i+h-1} + b) \), where \( v_{i:i+h-1} \) is the concatenated word embeddings from word \( i \) to word \( i + h - 1 \), \( * \) is the convolutional operation between two matrices, and \( f \) is the ReLU function, or \( f(x) = \max(0, x) \). The output of the convolutional operation is a feature map: \( s = (s_1, s_2, ..., s_{m \times h}) \). To extract the semantic features with different granularities, we use three convolutional kernels with the values of \( h \) set to 3, 4, and 5; these can capture the 3-, 4-, and 5-gram semantic information in the tweets.

- Max-pooling and dropout layers (part 3 in Fig. 1)

A max-pooling layer follows the convolutional layer to reduce the redundant information in the feature map produced by the convolutional layer. A max-over-time pooling is performed on each feature map, \( s^* = \max(s_i) \). Previous research has shown that the highest value on each dimension of the feature maps can capture the most important features (Kim, 2014). After the max-pooling operation, the features maps with different sizes calculated in the previous step are concatenated as the output of the max-pooling layer: \( z = (z_1, z_2, ..., z_l) \), where \( l \) is the number of kernels. In this paper, \( l = 3 \). Compared to the output of the convolutional layer, the max-pooling layer can keep the most important features and greatly improve the efficiency of the network.

We add a dropout layer to prevent model overfitting. The idea is to set certain dimensions of the max-pooling’s output to be zero: \( y = z^* r \), where \(^*\) is the Hadamard product operation of two vectors and \( r \) is a Bernoulli vector in which each element takes value 0 with probability \( p \) and takes value 1 with probability \( 1-p \).

- Connection of P-Net and A-Net (part 4 in Fig. 1)

As mentioned previously, the P-Net and the A-Net have the same network structure but with different training sets and performance of

---

1 Statistics show that the average length of tweets is 33 characters, and only 12% of tweets are longer than 140 characters (Wang et al., 2021b). 64 words is longer than most, if not all, tweets.
different classification tasks. Suppose after the embedding, convolutional, max-pooling, and dropout layers, the outputs of P-Net and A-Net are \( y_1 \) and \( y_2 \), respectively. To fully leverage the information from the A-Net and thus mitigate the class imbalance issue, \( y_1 \) and \( y_2 \) are combined using the structure shown in Fig. 1: \( y^1 = y_1 + \sigma(y_2) \) and \( y^2 = y_2 + \sigma(y_1) \), where \( \sigma \) is sigmoid function and serves as the weight to control the contribution of the A-Net to the P-Net. The rationale of a dual CNN is that the P-Net is able to capture the semantic features from the PHM dataset. The A-Net extracts the features from the training set with binary labels, which is the consolidated PHM dataset. For the A-Net’s input, the class imbalance issue is not very serious, and thus the features extracted are not biased to the major class/label. Thus, the features, \( \sigma(y_2) \), provide more reliable information as to whether a tweet is personal health-related or not. This information could enable the P-Net to focus on health-related tweets for PHM identification and avoid the bias towards dominating classes caused by the class imbalance issue. Thus, we combine \( \sigma(y_2) \) with \( y_1 \) to form the comprehensive feature \( y^2 \) for PHM identification.

- Fully connected and softmax layers (part 5 in Fig. 1)

The dual CNN’s output in the previous step is processed by a fully connected layer—i.e., a fully connected network—so that the dimension of the outcome is transformed to the desired output size. The output of the fully connected layer is represented as \( FC(y^2) \). A softmax layer takes \( FC(y^2) \) as input and outputs the probability distribution among the PHM classes. Specifically, \( \bar{p}_i = \sum_{n \in r} \exp(\text{FC}(y^2)) \), where \( C \) is the number of PHM classes and \( \bar{p}_i \) represents the probability that the \( j \)-th tweet is classified to the correct PHM category (Lee and Chen, 2005; Chou et al., 2004).

The complexity of CNN is \( O(nKD) \) where \( n \) is the convolutional kernel’s size, \( K \) is the dimension of embeddings, \( d \) is number of convolutional kernels, and \( L \) is the length of the inputted text (Shen et al., 2018). In the proposed dual CNN network, the connection between P-Net and A-Net only deploys vectors multiplication operation. The corresponding complexity is much lower than the matrix calculation in the convolutional operations and can be ignored. Thus, the complexity of the dual CNN structure is \( O(nLD + nLKD) \) where the subscript of \( P \) and \( A \) represent P-Net and A-Net respectively. If P-Net and A-Net are of the same structure, the complexity of dual CNN is degenerated to \( O(nLKD) \).

3.3. Model training

For each network, we use cross-entropy to quantify the loss \( L(\hat{y}, y) \) where \( \hat{y} \) is the predicted label and \( y \) is the true label. Since our goal is to predict the real PHM label for each tweet but not the auxiliary network’s outputs, we use a weight \( \lambda \) to combine the two losses from two networks, \( L = (1 - \lambda)L(\hat{y}, y) + \lambda L(\hat{y}^2, y^2) \).

To further address the class imbalance, we apply class weight for PHM labels. The class weight for PHM label \( i \) is defined as \( w_i = N/n_i \), where \( N \) is the number of samples for training data and \( n_i \) is the number samples of PHM class \( i \). Then, the training loss is defined as \( L = \sum_{i=1}^{C} w_i L_i \), where \( L_i \) is the loss for class \( i \). Since the auxiliary CNN has more balanced data, the class weight is not used for the auxiliary CNN’s loss function. In summary, the final loss for the model is \( L = (1 - \lambda)\sum_{i=1}^{C} w_i L_i(\hat{y}^i, y^i) + \lambda L(\hat{y}^2, y^2) \). All the parameters in the dual CNN are estimated during the training stage by minimizing this defined loss function.

4. Experiments setting

4.1. Data set

A COVID-19 tweet corpus was built for the PHM identification tasks. For this, 11,231 tweets were crawled using hashtags such as “coronavirus”, “SARS”, and “COVID”. The tweets cover the period March–May 2020. Two annotators independently annotated the tweets with the following four labels.

- self-mention: The tweet mentions the poster’s health condition related to COVID-19, e.g., “had my COVID 19 nasal swab Saturday. Got a call last night from CDC test was positive!”
- other-mention: The tweet mentions the COVID-19-related health condition of a person other than the tweet poster, e.g., “Just found out today that my gran caught COVID 19. Bless her.”
- awareness: The tweet contains COVID-19-related keywords but is not related to a specific person that the poster knows, e.g., “to stay away from coronavirus? Wear your mask and wash your hands!”
- non-health: The tweet may contain COVID-19-related keywords but is not COVID related, e.g., “corona #ks, drink German beer”

The inter-annotator agreement is 0.76 measured by Fleiss kappa, indicating a substantial agreement. The annotation disagreements were solved by the majority voting between the two annotators and a third annotator.

After data annotation, we noticed that the tweets’ distribution was highly imbalanced across the four classes: the percentages of self-mention, other-mention, awareness, and non-health PHMs were 2.8%, 9.8%, 72.7%, and 14.7% respectively. Most of the tweets were awareness and non-health, and a relatively small portion of people mentioned the COVID-19 health condition of themselves or the people they know.

4.2. Performance measure

Precision, recall, and F1 score are the major performance metrics for classification tasks. For the i-th PHM, the corresponding precision is defined as \( P_i = \frac{TP_i}{FP_i + TP_i} \), where \( TP_i \) is the number of correctly classified true positive samples in the testing data set. \( FP_i \) stands for the number of false positive samples, i.e., the number of tweets that are incorrectly classified as PHM i. The overall precision for the PHM identification can be defined as \( P = \frac{\sum_{i=1}^{M} P_i}{\sum_{i=1}^{M} TP_i} \), where \( M \) is the number of PHMs, i.e., \( M = 4 \). Precision quantifies the proportion of correctly classified tweets for each class. Similarly, the recall of PHM class \( i \) can be defined as \( R_i = \frac{TP_i}{FN_i + TP_i} \), where \( FN_i \) is the number of false negative samples in the testing set, i.e., the number of tweets that should be classified as the i-th PHM but aren’t mapped to i-th PHM. Recall quantifies the proportion of tweets in each PHM class that have been correctly classified. F1 score is the harmonic mean of precision and recall, i.e., \( F_1 = \frac{2PR}{P + R} \). F1 is a combination of precision and recall and thus is used as the main measure of the overall performance of PHM classification.

5. Experiment results

5.1. Model training

The 300-dimension GloVe embeddings are used to encode each word in the tweets. During the training stage, the whole tweet set is randomly partitioned in the proportion of 8:1:1 as the training, validation, and testing sets, respectively (Ju and Liu, 2019). The hyperparameters are shown in Table 1.

During the training stage, an important parameter is \( \lambda \) in the loss function. We use a trial-and-error process to determine the optimal value by evaluating the classification performance in the validation set. We
notice that the best performance is attained when \( i \) is between 0.2 and 0.3. Thus, we set \( i \) to be 0.25. Fig. 3 shows the experimental result of the relationship between the F1 score and loss weight \( \lambda \).

### 5.2. Overall performance

We use three popular classification methods as the baselines to show the effectiveness of the dual CNN-based approach. The baselines are:

- **WE + SVM**: Let \( v_i \) be the embedding vector of \( t_i \) where \( t_i \) is the \( i \)-th word in a tweet \( T = (t_1, t_2, ..., t_n) \). Then the element-wise average of all the \( v_i \) is used as the embedding of the tweet \( T \), i.e., \( v_T = \frac{1}{n} \sum v_i \) (Wang et al., 2021c). \( v_T \) is fed into a SVM to train the classifier to identify PHMs (Ju et al., 2019). In this paper, 300-dimension GloVe is used to encode each word. We used this approach as a non-deep learning baseline.

- **LSTM network** (Hochreiter and Schmidhuber, 1997; Wang et al., 2021b): LSTM is a popular recurrent neural network (RNN) unit. It takes the sequence of the embeddings of the words in the tweets as the input. LSTM uses forget gate, input gate, and forget gate to determine the information flow in the sequence. Research has shown that LSTM could capture the contextual dependency in the text.

- **Gated recurrent unit (GRU)** (Wang et al., 2021a): GRU is also a type of RNN unit. It uses the reset gate and update gate to control the contextual information update. GRU is simpler than LSTM but with comparable performance in some classification tasks.

- **Bidirectional Encoder Representations from Transformers (BERT)** (Devlin et al., 2019): BERT and its variants have achieved state-of-the-art performance in many tasks. We use the pre-trained BERT to encode the tweets text. The representation is fed into a softmax to perform the classification task. BERT is fine-tuned on the training set and tested on the testing set.

Table 2 shows overall performance of the proposed dual CNN and the baselines. The best performances are highlighted with bold font. Dual CNN demonstrates the best overall performance for PHM identification. CNN structure leverages the convolutional kernels with different sizes (3, 4 and 5 in this paper) to capture the contextual information.

#### Table 1

| Hyperparameter       | Value          |
|----------------------|----------------|
| Dropout rate         | 0.5            |
| Batch size           | 64             |
| Learning rate        | 0.002          |

#### Table 2

| Model     | Self-mention | Other-mention | Awareness | Non-health | Prec.   | Recall  | F1 score |
|-----------|--------------|---------------|-----------|------------|---------|---------|----------|
| LSTM      | 0.5000       | 0.6250        | 0.8217    | 0.5581     | 0.4643  | 0.5906  | 0.4815   |
| Dual CNN  | 0.5389       | 0.8189        | 0.9060    | 0.6027     | 0.4815  | 0.5906  | 0.5823   |

Considering that tweets are not long, CNN structure is efficient at extracting the semantics from the tweets. Recurrent neural network-(RNN-) based structure, such as LSTM and GRU, could model the long-range contextual information in the text. Since tweets are not long text, RNN-based structures perform worse than CNN based ones. BERT has shown its superiority at various NLP tasks (Devlin et al., 2019; Luo and Wang, 2019). However, it usually requires big data set to fine-tune and adapted to the target tasks. The data set in this paper is not huge. Thus, the semantic difference between the training set of BERT and COVID-19 cannot be sufficiently bridged. This explains the worse performance of BERT comparing with dual CNN. If the data set is in a bigger scale (for example, hundreds of thousands or even bigger), we think the performance of BERT would be greatly improved. Overall, the proposed dual CNN with balancing treatment improves the overall classification of the labels.

In addition, we would also like to see the performance comparison between dual CNN and the baselines on the classification of each label. Due to the space limits, we only compare dual CNN with LSTM, as the latter shows the best performance of the baselines.

Based on Table 3, it can be observed that the performance of dual CNN in identifying self-mention PHMs and other-mention PHMs is much better than that of LSTM (0.5823 vs 0.4815 for self-mention and 0.6500 vs 0.6073 for other-mention). For awareness (class 3), the performance of dual CNN is slightly worse. This is because the performance of LSTM (and others) is governed by the performance of the dominant class, i.e., the awareness class under class imbalance. When training the models, the main objective is to optimize the performance on the dominant class. The proposed dual CNN with balancing treatment considers the classes with smaller sample size, and thus address the class imbalance problem. It sacrifices the performance for this dominant label and improves the performance of other labels. Since class 3 is not of primary interest for public health surveillance purposes, its relative worse performance is acceptable.

In addition, we argue that recall is more important than precision in the PHM identification task; recall measures the percentage of PHMs that have been correctly identified. Since PHM identification serves as an initial step in epidemic intelligence, it would be more important for a health surveillance system to collect all the relevant PHMs—i.e., self-mentions and other-mentions—without missing any; thus, a higher recall rate is desired. Although the number of false positive identifications of PHMs may be high (which means the precision is low), the subsequent steps in public health surveillance could involve further screening of the identified PHMs to eliminate false positives and provide reliable data for policymakers to make decisions. It is worth using a

#### Table 3

| Model     | Label      | Precision | Recall  | F1 score |
|-----------|------------|-----------|---------|----------|
| LSTM      | Self-mention | 0.5000    | 0.4643  | 0.4815   |
|           | Other-mention | 0.6250    | 0.5906  | 0.6073   |
|           | Awareness   | 0.8217    | 0.9016  | 0.8598   |
|           | Non-health  | 0.5581    | 0.3077  | 0.3967   |
| Dual CNN  | Self-mention | 0.5389    | 0.8189  | 0.5823   |
|           | Other-mention | 0.4510    | 0.8214  | 0.5823   |
|           | Awareness   | 0.9060    | 0.8080  | 0.8597   |
|           | Non-health  | 0.6027    | 0.5641  | 0.5828   |
moderate scarification of precision to attain a big improvement of recall, as seen in the dual CNN. The recalls of self-mention and other-mention are both higher than 0.8. For LSTM, the recalls on these two classes are much worse.

In summary, the proposed method uses an A-Net to mitigate the class imbalance issue. Although the classification performance of the dominating class is sacrificed, this improves the identification of the important PHMs for health surveillance—i.e., self-mentions and other-mentions.

5.3. The effect of A-Net on overall performance

In Fig. 4, it can be observed that class imbalance affects the overall performance of PHM identification. This may cause serious issues for the PHM identification task, since the relevant classes have fewer training samples. This paper proposes the dual CNN structure in an attempt to mitigate the class imbalance issue. We wanted to investigate whether the A-Net could provide support to the primary CNN structure; to do so, we conduct an experiment of PHM identification using a vanilla CNN without the A-Net structure. Weighted loss function is used during the training stage. The comparison of the performances of the dual CNN and the vanilla CNN is shown in Table 4.

In Table 4, we notice that dual CNN outperforms vanilla CNN in the classifications of self-mention, other-mention and non-health in terms of the overall F1 score. However, CNN slightly outperform the detection of awareness class, which is not of primary concern for public health surveillance task. The recall rates of self-attention and other-mention (which are of primary interest) are increased from 0.5357 to 0.8214 and from 0.5197 to 0.8189 with A-Net. By comparing the recalls of these four PHM classes, we can see that the performance of the dominating class’s identification is somewhat sacrificed, but the performances of the classes with fewer samples improve significantly. Thus, the A-Net provides an effective way of identifying PHMs by balancing the performances of all the classes. This is extremely useful when the relevant classes contain fewer samples, as in the PHM identification task.

5.4. The effect of training sample size on overall performance

Based on Table 3, we can see that class imbalance greatly influences the identification of PHMs. One direct consequence is that the performance is determined by the dominating class (in this study, awareness), which is not relevant to public health surveillance. The trained classifier also favors the dominating PHM class and may not perform well for the relevant PHM classes (in this study, self-mentions and other-mentions) as indicated in Table 3. The relevant classes cannot contribute much to the overall performance due to their small sample sizes. The dual CNN can compensate for this issue to some extent (see previous subsection).

A straightforward way to mitigate the effect of class imbalance is to expand the smaller-size classes. Since class 1, 2 and 4 tweets only account for a small proportion of tweets, it is not economically efficient to crawl these tweets from Twitter. In NLP research, various data augmentation approaches have been proposed to solve data sparsity issues. We wanted to see whether manually-augmented data could improve COVID-19 PHM identification performance. We apply the widely-used approaches of back translation, synonym replacement, and random swap to generate more text in classes 1 and 2 (Wei and Zou, 2019). After data augmentation, the sample sizes of class 1 and 2 triple. We use the augmented data set to retrain the dual CNN network and test it on the same testing set. All the settings remain the same in order to enable a fair comparison. The results are shown in Fig. 4. It can be observed that for class 1 the F1 score increases significantly from 0.5823 to 0.6500; for label 2, the F1 score increases from 0.65 to 0.6625. For label 3 and label 4, the F1 scores slightly increased, as data augmentation was not performed on them. Usually among the four classes, we are more concerned with the identification of class 1 (self-mention) as it indicates the true disease situation in the community. Data augmentation can efficiently improve the performance of PHM for class 1 without the sacrifice of other classes.

6. Conclusion

PHM identification can be a critical tool for social media public health surveillance. To the best of our knowledge, this paper is the first exploration of COVID-19 PHM detection in tweets. We built a COVID-19 dataset from social media to train our models. To deal with class imbalance, we propose an A-Net that provides support to the primary CNN structure; to do so, we conduct an experiment of PHM identification using a vanilla CNN without the A-Net structure. Weighted loss function is used during the training stage. The comparison of the performances of the dual CNN and the vanilla CNN is shown in Table 4. In Table 4, we notice that dual CNN outperforms vanilla CNN in the classifications of self-mention, other-mention and non-health in terms of the overall F1 score. However, CNN slightly outperform the detection of awareness class, which is not of primary concern for public health surveillance task. The recall rates of self-attention and other-mention (which are of primary interest) are increased from 0.5357 to 0.8214 and from 0.5197 to 0.8189 with A-Net. By comparing the recalls of these four PHM classes, we can see that the performance of the dominating class’s identification is somewhat sacrificed, but the performances of the classes with fewer samples improve significantly. Thus, the A-Net provides an effective way of identifying PHMs by balancing the performances of all the classes. This is extremely useful when the relevant classes contain fewer samples, as in the PHM identification task.

Table 4

The performances comparison between CNN and Dual CNN on each class.

| Model  | Label    | Precision | Recall | F1 score |
|--------|----------|-----------|--------|----------|
| CNN    | Self-mention | 0.5000   | 0.5357 | 0.5172   |
|        | Other-mention | 0.6286   | 0.5197 | 0.5690   |
|        | Awareness    | 0.8206   | 0.9114 | 0.8636   |
|        | Non-health   | 0.6163   | 0.3397 | 0.4380   |
| Dual CNN | Self-mention | 0.4510   | 0.8214 | 0.5823   |
|         | Other-mention | 0.5389   | 0.8189 | 0.6500   |
|         | Awareness    | 0.9060   | 0.8080 | 0.8597   |
|         | Non-health   | 0.6027   | 0.5641 | 0.5828   |

Fig. 4. The comparison of F1 scores with and without data augmentation for each label (class).
PHM dataset containing 11,231 annotated tweets. COVID-19 PHM identification was defined as a text classification task. We proposed a dual CNN structure to address the concerns. The dual CNN fully leveraged the auxiliary information extracted by the A-Net to mitigate the class imbalance issue in the dataset. The experiments showed the effectiveness of the dual CNN in identifying PHMs, especially the PHMs that are crucial to public health surveillance.

This work has its limitations. To start, we believe that there is much room for improvement of the PHM identification task’s performance. The dataset has been made available, and the proposed method in this paper could serve as a baseline for other researchers’ explorations into more effective methods. In addition, we only considered the semantic information from tweets; other information, such as the posting time of tweets and the region of the posters, was not considered. Thus, the method is only applicable to a PHM identification task. To address other issues in epidemic intelligence, such as disease trend forecasting, multilingual information needs to be considered. In the future, we hope to try different data augmentation methods to address the class imbalance issue. Data augmentation methods like back-translation have been proven effective for NLP tasks with low resources, and data imbalance can be partially solved by manual augmentation; we hope to explore whether it is an effective means for the PHM identification task. In addition, external domain knowledge related to COVID-19 could provide some insights into PHM identification; we hope to explore the effectiveness of integrating domain knowledge and various optimization techniques into a machine learning structure in future work (Liu et al., 2020; Ju et al., 2021).

CRediT authorship contribution statement

Linkai Luo: Investigation, Methodology, Validation. Yue Wang: Conceptualization, Funding acquisition, Investigation, project administration, Supervision, Writing - original draft. Hai Liu: Formal analysis/Resources, Supervision, Writing - review & editing.

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.

References

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1568-1576).

Barnes, S. J., Diaz, M., & Arnaboldi, M. (2021). Understanding panic buying during COVID-19. Expert Systems with Applications, 169, Article 115655.

Biddle, R., Joshi, A., Liu, S., Paris, C., & Guandong, X. (2020). Leveraging sentiment analysisResources, Supervision, Writing - review

Deiner, M., Lietman, T., McLeod, S., Chodosh, J., & Porco, T. (2016). Surveillance tools emerging from search engines and social media data for determining eye disease patterns. Jama Ophthalmology, 134(9), 1024-1030.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186).

Díaz-Aviles, E., & Stewart, A. (2012). Tracking Twitter for epidemic intelligence: Case study: EHHC/HUS outbreak in Germany. Proceedings of Web Science Conference, 27-32.

Gonzalez, A., Stilo, G., Gonfiantini, M. V., Pandolfi, E., Velard, P., & Tozzi, A. E. (2013). Influenza-like illness surveillance on Twitter through automated learning of naive language. PloS ONE, 8(12), Article e82489.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

Huang, P., Mackinlay, A. & Yepes, A. J. (2016). Syndromic surveillance using generic medical entities on Twitter. Proceedings of the Australasian Language Technology Association Workshop, 35–44.

Iyer, V., Joshi, A., Karimi, S., Sparks, R., & Paris, C. (2019). Figurative usage detection of symptom words to improve personal health mention detection, arXiv:1906.05466.

Jiang, K., Calix, R., & Gupta, M. (2016). Construction of a personal experience tweet corpus for health surveillance. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing (pp. 128-135).

Jiang, K., Feng, S., Song, Q., Calix, R. A., Gupta, M., & Bernard, G. R. (2018). Identifying tweets of personal health experience through word embedding and LSTM neural network. BMC Bioinformatics, 19(8), 210.

Joshi, A., Sparks, R., Karimi, S., Yan, S.-L., Chughtai, A., Paris, C., & MacIntyre, C. R. (2020). Automated monitoring of tweets for early detection of the 2014 Ebola epidemic. PloS ONE, 15(3), Article e0230322.

Joshi, A., Karimi, S., Sparks, R., Paris, C., & MacIntyre, C. R. (2019). A comparison of word2vec based and contextualized representations for classification problems in health informatics. In Proceedings of Workshop on Biomedical Natural Language Processing (pp. 135–141).

Ju, X., Chen, V. C. P., Rosenberger, J. M., & Liu, F. (2021). Fast knot optimization for multivariate adaptive regression splines using hill climbing methods. Expert System with Applications, 171, Article 115655.

Ju, X., & Liu, F. (2019). Wind farm layout optimization using self-informed genetic algorithm with information guided exploitation. Applied Energy, 248, 429-445.

Ju, X., Liu, F., Wang, L., & Lee, W.-J. (2019). Wind farm layout optimization based on support vector regression guided genetic algorithm with consideration of participation among landowners. Energy Conversion and Management, 196(9), 1267-1281.

Kim, Y. (2014). Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746–1751.

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on Twitter. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 789-795).

Lee, T.-S., & Chen, L.-F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, 28(4), 743-752.

Liu, F., Ju, X., Wang, N., Wang, L., & Lee, W.-J. (2020). Wind farm macro-siting optimization with insightful bi-criteria identification and relocation mechanism in genetic algorithm. Energy Conversion and Management, 217, Article 112964.

Luo, L., & Wang, Y. (2019). EmotionX-HSU: Adopting pre-trained BERT for emotion classification, arXiv:1907.09669.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in Vector Space, arXiv:1301.3781.

Nguyen, D. Q., Vu, T., Rahimi, A., Dao, M. H., Nguyen, L. T. & Doan, L. (2020). WNWUT-2020 task 2: Identification of informative COVID-19 English tweets, Proceedings of the Sixth Workshop on Noisy User-generated Text, 314–318.

Oleszkowski, R. T. (2006). Bayesian classification of triage diagnoses for the early detection of epidemics. In Proceedings of the International Florida Artificial Intelligence Research Society Conference (pp. 412-416).

Paul, M. J., & Dredze, M. (2011). What are you tweet: Analyzing Twitter for public health. In Proceedings of the Fifth International Conference on Weblogs and Social Media (pp. 258–272).

Paul, M. J., & Dredze, M. (2017). Social monitoring for public health. Synthesis Lectures on Information Concepts, Retrieval, and Services, 9(5), 1-183.

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1532–1543.

Rivadeneira, L., Yang, J., & López-Ibáñez, M. (2021). Predicting tweet impact using a novel evidential reasoning prediction method. Expert Systems with Applications, 169, Article 114400.

Scudellari, M. (2020). How the pandemic might play out in 2021 and beyond. Nature, 584, 22-25.

Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y.,...Carin, L. (2018). Baseline models must love: On simple word-embedding-based models and associated pooling mechanisms. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 440–450).

Wang, C.-K., Singh, O., Tang, Z.-L., & Dai, H.-J. (2017). Using a recurrent neural network model for classification of tweets conveyed influenza-related information. In...
Proceedings of the International Workshop on Digital Disease Detection Using Social Media (pp. 33–38).
Wang, P., Li, M., Li, X., Zhou, H., & Hou, J. (2021a). A hybrid approach to classifying Wikipedia article quality flaws with feature fusion framework. Expert Systems with Applications, 181, Article 115089.
Wang, Y., & Li, X. (2021). Mining product reviews for needs-based product configurator design: A transfer learning-based approach. IEEE Transactions on Industrial Informatics, 17(9), 6192–6199.
Wang, Y., Li, X., & Mo, D. Y. (2021b). Knowledge-empowered multitask learning to address the semantic gap between customer needs and design specifications. IEEE Transactions on Industrial Informatics, 17(12), 8397–8405.
Wang, Y., Li, X., Zhang, L., & Mo, D. Y. (2021c). Configuring products with natural language: A simple yet effective approach based on text embeddings and multilayer perceptron. International Journal of Production Research, accepted, https://doi.org/10.1080/00207543.2021.1957508
Yu, S., Liu, D., Zhu, W., Zhang, Y., & Zhao, S. (2020). Attention-based LSTM, GRU and CNN for short text classification. Journal of Intelligent & Fuzzy Systems, 39(1), 333–340.
Zheng, L., He, Z., & He, S. (2021). An integrated probabilistic graphic model and FMEA approach to identify product defects from social media data. Expert Systems with Applications, 178, Article 115030.