Original Paper

Linguistic Pattern–Infused Dual-Channel Bidirectional Long Short-term Memory With Attention for Dengue Case Summary Generation From the Program for Monitoring Emerging Diseases–Mail Database: Algorithm Development Study

Yung-Chun Chang[1,2], PhD; Yu-Wen Chiu[1,3], BA; Ting-Wu Chuang[3], PhD

[1]Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan

[2]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

[3]Department of Molecular Parasitology and Tropical Diseases, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Ting-Wu Chuang, PhD
Department of Molecular Parasitology and Tropical Diseases
School of Medicine, College of Medicine
Taipei Medical University
No 250
Wu-Hsing Street
Taipei, 110
Taiwan
Phone: 886 27361661 ext 3123
Email: chtingwu@tmu.edu.tw

Abstract

Background: Globalization and environmental changes have intensified the emergence or re-emergence of infectious diseases worldwide, such as outbreaks of dengue fever in Southeast Asia. Collaboration on region-wide infectious disease surveillance systems is therefore critical but difficult to achieve because of the different transparency levels of health information systems in different countries. Although the Program for Monitoring Emerging Diseases (ProMED)–mail is the most comprehensive international expert–curated platform providing rich disease outbreak information on humans, animals, and plants, the unstructured text content of the reports makes analysis for further application difficult.

Objective: To make monitoring the epidemic situation in Southeast Asia more efficient, this study aims to develop an automatic summary of the alert articles from ProMED-mail, a huge textual data source. In this paper, we proposed a text summarization method that uses natural language processing technology to automatically extract important sentences from alert articles in ProMED-mail emails to generate summaries. Using our method, we can quickly capture crucial information to help make important decisions regarding epidemic surveillance.

Methods: Our data, which span a period from 1994 to 2019, come from the ProMED-mail website. We analyzed the collected data to establish a unique Taiwan dengue corpus that was validated with professionals' annotations to achieve almost perfect agreement (Cohen κ=90%). To generate a ProMED-mail summary, we developed a dual-channel bidirectional long short-term memory with attention mechanism with infused latent syntactic features to identify key sentences from the alerting article.

Results: Our method is superior to many well-known machine learning and neural network approaches in identifying important sentences, achieving a macroaverage F1 score of 93%. Moreover, it can successfully extract the relevant correct information on dengue fever from a ProMED-mail alerting article, which can help researchers or general users to quickly understand the essence of the alerting article at first glance. In addition to verifying the model, we also recruited 3 professional experts and 2 students from related fields to participate in a satisfaction survey on the generated summaries, and the results show that 84% (63/75) of the summaries received high satisfaction ratings.

Conclusions: The proposed approach successfully fuses latent syntactic features into a deep neural network to analyze the syntactic, semantic, and contextual information in the text. It then exploits the derived information to identify crucial sentences in the ProMED-mail alerting article. The experiment results show that the proposed method is not only effective but also outperforms the compared methods. Our approach also demonstrates the potential for case summary generation from ProMED-mail alerting

articles. In terms of practical application, when a new alerting article arrives, our method can quickly identify the relevant case information, which is the most critical part, to use as a reference or for further analysis.

(JMIR Public Health Surveill 2022;8(7):e34583) doi: 10.2196/34583

KEYWORDS

ProMED-mail; natural language processing; dengue; dual channel; bidirectional long short-term memory

Introduction

Background

Globalization and climate change have exacerbated the frequency and virulence of infectious diseases worldwide [1-3]. Climate and environmental changes play an undeniable role in changing disease ecology and transmission dynamics [4-6], with transboundary transmission also being frequently linked to international transportation [7]. Monitoring the region-wide or global infectious disease transmission patterns relies on intercountry collaborations to share disease surveillance information. The Program for Monitoring Emerging Diseases (ProMED)–mail [8] was launched by the International Society for Infectious Diseases in 1994 to collect global disease outbreak information on humans, animals, and plants [9,10]. Currently, ProMED-mail is the largest unofficial infectious diseases platform based on volunteer reporting, and it receives disease outbreak reports or research findings from different users (including individual scientists and governmental agencies) around the world. The EpiCore program, involving a worldwide network of public health professionals, was added in 2014 to scrutinize and verify the reported information [11].

Each report's quality is enhanced through an expert-review process that includes reducing data redundancy and errors, which are common in social media reports. Reports from social media platforms such as Twitter and Facebook have been used to detect disease outbreaks in previous works, with Google Trends being a good example of web data providing early warning messages regarding influenza [12]. However, social media reports have at least three main limitations. The first involves unclear definitions. Many infectious diseases might share very similar clinical symptoms, making it difficult to differentiate them from simple keyword searches by users. Second, with the passage of time, the attention that people pay to disease outbreaks wanes. The third limitation relates to social media accessibility in different countries or groups of people: when researchers conduct long-term pattern analysis or multinational analysis of disease outbreaks, social media will introduce bias. These issues have been addressed by a few studies [13,14]. Social media and ProMED-mail might play different roles regarding dengue detection and analysis: whereas social media can be used to detect the emergence of dengue in the early stage, ProMED-mail can provide richer, more correct, and reliable epidemiological information, as well as continuous monitoring of that information [15-17].

Well-known contributions of ProMED-mail are the early reports of suspected cases of severe acute respiratory syndrome in China in 2002 and the Middle East respiratory syndrome coronavirus in Saudi Arabia in 2012 [11,18]. More recently, ProMED-mail data have been used to analyze a cholera outbreak in Africa, a vector-borne disease outbreak amid violent conflict in Syria, and global avian influenza outbreaks [19-21]. Thanks to >25 years of effort, huge amounts of disease outbreak information have been accumulated in the ProMED-mail database; however, the unstructured text format of the ProMED-mail report hampers the efficiency of scientific analysis. Most previous studies using the ProMED-mail database usually relied on labor-intensive review processes, making it difficult to analyze multiple diseases and broader study areas. However, natural language processing (NLP) can help because it is a powerful technique for extracting information from unstructured clinical or health records [22-24].

With the outbreak of COVID-19, sources of epidemic surveillance have received more and more attention, prompting the publication of several research papers. Nonetheless, few studies have taken advantage of NLP technology for the development or analysis of data from ProMED websites. Carrion and Madoff [11] have noted that every season, ProMED would publish a word cloud of epidemics in various regions to show epidemics that have been particularly severe in each region in the current season. Taking 2016 as an example, alerting reports from all over the world were processed by NLP technology to produce a visualized word cloud in which dengue is the keyword for entire Southeast Asia. In addition, Kim et al [25] developed a deep learning approach to automatically recognize the relevant information that is necessary to deal with potential disease outbreaks; this is consistent with our view. Their study used 2 approaches—convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) [26,27]—to classify the sources of the texts about infectious diseases in the alerting articles, and they achieved an overall accuracy rate of 92.9%.

In recent years, factors such as climate, weather, and culture have stimulated global epidemic monitoring, from which it has become clear that dengue fever in Southeast Asia still poses a serious threat [28,29]. Dengue incidence has increased significantly around the world in the last 2 decades [30], with an estimated 390 million infections per year, of which 96 million exhibit clinical symptoms. Each year, approximately 3.9 billion people are at risk of infection with dengue viruses, with 2 million severe cases and 2100 deaths [31]. Although dengue infection is prevalent in 129 countries, 70% of all cases are located in Asia. The reported number of dengue cases has increased >8-fold over the last 2 decades, and most deaths have occurred in younger age groups. As there is currently no effective vaccine or treatment available for dengue infection, dengue surveillance information is imperative for disease control and prevention.

For these reasons, our research focuses on Southeast Asia. As the alerting articles from ProMED-mail are lengthy (an average of 1872 words and 82 sentences in Asian-related dengue fever

alerting articles), it is important to develop a summary generation system to assist relevant researchers to become more proficient in monitoring the pandemic. In general, case information is related to the outbreak location, time, and patient [32]. The combination of these 3 types of information constitutes the coincidence of occurrence of case information. However, although NLP research for information extraction has flourished [33-36], it is not easy to determine whether all the important case information is contained in a single ProMED-mail alerting article. For example, as digits usually represent the number of cases, the appearance of important case information is often accompanied by the appearance of numbers. However, in the sentence "CDC deputy director-general Chuang Jen-Hsiang said on Wednesday [August 7, 2019] that the patient also has underlying diseases, which is why he was only diagnosed with dengue after 2 screenings and multiple hospital visits," the digit "2" represents frequency, rather than the number of cases. It thus cannot become one of the sentences in the summary, although it contains dengue-related keywords such as disease, dengue, and diagnose. Moreover, location and time are also relevant information for important cases. We thus assume intuitively that if the location and time are mentioned in the same sentence, it may more likely describe relevant information about the case. However, although the sentence "Since 18 Nov [2006], Kaohsiung County City health authorities reinforced implementation of the mosquito-elimination campaign" mentions location and time, the content obviously does not contain information related to any infection outbreak.

Objectives

The specific aim of this research was therefore to extract abstract sentences that contain important epidemiological information on dengue incidence in Southeast Asia. Our proposed method can enhance the decision-making efficiency of epidemic monitoring units by quickly and automatically generating summaries of alerting articles. This is particularly important during the current COVID-19 pandemic. Specifically, our method first decomposes an alerting article from ProMED-mail into sentences. Next, the sentences are classified into 2 categories in accordance with their syntactic characteristics. Finally, the proposed deep neural network–based method integrates linguistic patterns and latent syntactic features to identify important sentences as the basic unit of summary generation. The experiment results based on real-world data sets demonstrate that the proposed method successfully exploits the syntactic, semantic, and contextual information relevant to epidemiological information on dengue. Consequently, our method not only outperforms many well-known information extraction methods but also achieves a satisfaction rating of 84% for the abstractive sentence summarization of dengue alerting article data.

Methods

Data Corpus

The aim of this study was to develop a method to extract sentences that convey dengue case information in ProMED-mail alerting articles and then automatically generate summaries. However, to the best of our knowledge, there is no official data set for crucial sentences extraction with regard to dengue. For this reason, we compiled our own data corpus for method development and performance evaluation. To do this, we first collected all the articles from 1994 to 2019 on the ProMED-mail website as a preliminary corpus. Next, we used the country name to further extract Southeast Asia articles and then verified that the title contained "Dengue/DHF (dengue hemorrhagic fever) updates" to indicate a series of dengue fever alerting articles. To facilitate the efficiency of the corpus construction, we sampled 15% of the instances from the data set for annotation for a total of 129 articles that contained 965 sentences. Next, the data set was annotated by 2 experts who are medical professionals with high English proficiency. Before the annotation stage, we conducted a training to ensure that the annotators had a common understanding of what defines crucial sentences in a dengue summary. We provided each annotator with 20 instances of both positive and negative cases during the training stage. A third expert acted as an arbiter for verifying the annotations. During the actual labeling process, the annotators labeled 246 instances as crucial sentences (Cohen κ=0.895, which indicates that the interobserver agreement of our annotated data corpus is reliable [37-39]). The final annotation results were used for the performance evaluation of the proposed model. It is worth noting that 80.1% (197/246) of the crucial sentences were composed of multiple clauses with complex syntactic structures (Table 1). Furthermore, 34.6% (85/246) of the sentences contained digits that did not convey case information (eg, "Dengue virus circulating type 1" and "20-34 mosquitoes or mosquito larvae found in every 100 households"). In light of these sentence statistics, it is clearly a challenging task to extract sentences that convey case information. The corpus has been released to promote further research and is available at the Taipei Medical University Dengue Case Corpus [40].

Table 1. The statistics of our corpus (N=965)[a].

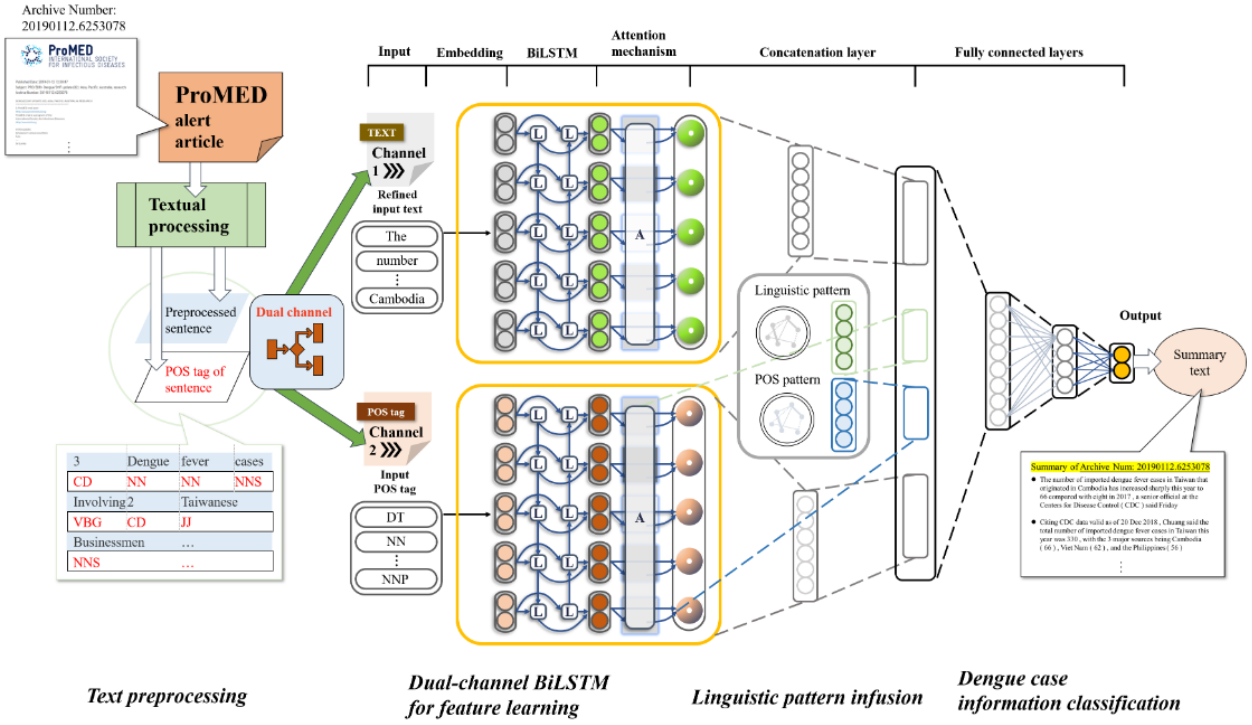
	Number of sentences, n (%)	
	Single-clausal sentences	Multiclausal sentences
Dengue case sentences (n=246)	49 (19.9)	197 (80.1)
Non-dengue case sentences (n=719)	407 (56.6)	312 (43.4)

[a]Number of paragraphs: 129, number of sentences: 965, number of single-clausal sentences: 456, and number of multiclausal sentences: 509.

Figure 1 presents the overview of the proposed framework, which automatically detects summary sentences that contain case-related information in a set of documents of the dengue alerting series written in English. In this research, we used a

linguistic pattern–infused BiLSTM with an attention mechanism neural network for dengue case information extraction. That is, to extract sentences that convey dengue case information from alerting articles from ProMED-mail, we treated dengue case information extraction as a binary classification problem that can be formulated as follows. Let $W = \{w_1,\ldots, w_k\}$ be a set of words and $S = \{s_1,\ldots, s_m\}$ be a set of sentences from a set of alerting articles. Each sentence s comprises a set of words such that sW. The goal of this task was to decide whether a sentence s_j expresses dengue case information. Our framework consisted of 4 main procedures: text preprocessing, dual-channel BiLSTM (DuBiLSTM), linguistic pattern infusion, and dengue case information classification. Further details of each procedure are provided in the following sections.

Figure 1. Overview of the proposed framework. A: attention layer; BiLSTM: bidirectional long short-term memory; CD: cardinal number; DT: determiner; JJ: adjective; L: forward long short-term memory layer and backward long short-term memory layer; NN: noun, singular or mass; NNS: noun, plural; POS: parts of speech; ProMED: Program for Monitoring Emerging Diseases; VBG: verb, gerund, or present participle.



Text Preprocessing

Preprocessing is a critical task that needs to be performed before feeding data into a neural network. When a report document d_n is entered, we first decompose it into a set of paragraphs $P = \{p_1,..., p_i\}$ and obtain the sentence collections $S = \{s_1,..., s_j\}$ through sentence segmentation for each paragraph. Next, we break a sentence into tokens and tag the parts of speech (POS) $T = \{t_1,..., t_k\}$ using the Natural Language Toolkit package [41]. As dengue case information can be narrated in a sequence of clauses, we recognize 2 types of sentences, namely single-clausal sentences and multiclausal sentences. Moreover, as frequently used words are generally not helpful for identifying dengue case information, we removed the stop words as well as punctuation marks (commas and semicolons) in the sentences.

DuBiLSTM With Attention Mechanism

In this research, we developed a DuBiLSTM with an attention mechanism neural network to learn latent semantic features behind both alert article texts and shallow paring information. The embedding layer is first used to transform the input tokens and POS tags of a sentence into 300D vectors. We used Global Vectors for Word Representation pretrained word embeddings (ie, glove.6B) to transform sentences of the alerting articles into 300D vectors. The POS embeddings are learned from the embedding layer of our model with continuous bag-of-words mode. Specifically, a sentence is represented by $s = \{w_1,..., w_k\}\{pos_1,..., pos_k\}$, its corresponding word vector is $v_w = \{v_{w\_1},..., v_{w\_n}\}$, and its POS vector is $v_{pos} = \{v_{pos\_1},..., v_{pos\_n}\}$, which are the inputs of the model.

Compared with the original recurrent neural network, the reason for the improvement of long short-term memory (LSTM) is its special design. LSTM defines and maintains an internal memory cell state throughout the life cycle to establish temporal connections. This internal memory cell state is the most important element of LSTM's structure. The LSTM model consists of a series of identical timing modules. In addition to the original input, LSTM has 3 designs—forget gate, input gate, and output gate—that determine whether the input is important enough to be remembered and whether it can be output.

The details are described herein. Suppose there are 3 element-wise functions that help to calculate the next moment by the previous moment and this moment where σ(.) is a sigmoid function, tanh(.) is a hyperbolic tangent function, and $\odot$ is the product. We also have $x_t \in \mathbb{R}^d$ and $h_t \in \mathbb{R}^h$ denoting the input vector and the hidden state vector at moment t, respectively,

whereas $U \in \mathbb{R}^{h \times h}$ and $W \in \mathbb{R}^{h \times d}$ indicate the weight metrices of gates or cells for input vector x_t and hidden state vector h_t, respectively, and $b \in \mathbb{R}^h$ indicates the weight metrices of gates or cells for the bias vector, where the superscripts *d* and *h* refer to the number of input features and number of hidden units, respectively. The forget gate at the moment *t* $f_t \in \mathbb{R}^h$ determines the information to be forgotten by outputting a number in (0, 1), in line with the following equation:

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \quad \textbf{(1)}$$

With regard to the second mechanism, the input gate of LSTM then decides what new information input should be kept by calculating $i_t \in \mathbb{R}^h$ and $\tilde{c}_t \in \mathbb{R}^h$ and combining the 2 parameters in the light of the following equations:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \quad \textbf{(2)}$$

$$\tilde{c}_t = \tanh(W_c h_{t-1} + U_c x_t + b_c) \quad \textbf{(3)}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad \textbf{(4)}$$

The third special mechanism is the output gate. This represents which parts of the cell state should be outputted based on the following equations, where $h_t \in \mathbb{R}^h$ represents the hidden state vector, also known as the output vector:

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \quad \textbf{(5)}$$

$$h_t = o_t \odot \tanh(c_t) \quad \textbf{(6)}$$

However, the information in the LSTM network is a 1-way transmission, and LSTM can only use past information, not future information. BiLSTM can consider both past and future data information by connecting 2 LSTM networks with opposite timings in the same output. The forward andbackward LSTMs can obtain, respectively, the past and future data information of the input sequence. The hidden state H_t of BiLSTM at time *t* includes forward $\vec{h}_t$ and backward $\overleftarrow{h}_t$:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}(h_{t-1}, x_t, c_{t-1}), t \in [1, T] \quad \textbf{(7)}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(h_{t+1}, x_t, c_{t+1}), t \in [T, 1] \quad \textbf{(8)}$$

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad \textbf{(9)}$$

In addition, as the attention mechanism can allocate more attention to important information and less to other information, the receiving sensitivity and processing speed of information in the focused attention area are greatly improved. The attention function can softly map the combination of query *Q* and a set of key-value pairs {*K*,*V*} to some notable outputting results, where $Q = \{Q_1, ..., Q_N\}$ and $\{K,V\} = \{(K_1,V_1), ..., (K_M,V_M)\}$. Furthermore, the multihead attention mechanism would convert *Q*, *K*, and *V* into *H* subspaces in the first step, with various and learnable linear projections, as the following transforming equation shows:

$$Q^h, K^h, V^h = QW_h^Q, KW_h^K, VW_h^V \quad \textbf{(10)}$$

where $\{Q^h, K^h, V^h\}$ are the input query, key, and value of the *n*th head, respectively; $\{W_h^Q, W_h^K, W_h^V\} \in \mathbb{R}^{d \times d_k}$ represents the parameter matrices at the same time; and *d* and d_k indicate, respectively, the dimension of the model and its subspace. In the second step, *H* attention functions are implemented in parallel to generate the output states $O = \{O^1, ..., O^H\}$, where any O^h in *O* is defined in the following equation:

$$O^h = Att^h V^h \text{ with } Att^h = softmax\left(\frac{Q^h K^{hT}}{\sqrt{d_k}}\right) \quad \textbf{(11)}$$

where Att^h is the attention distribution that comes from the *h*th attention head. Finally, the output states *O* are concatenated with each other and then connected with distinctively generated features for the next stage.

Linguistic Pattern Infusion

The human perception of a dengue case information alert involves identifying a relevant lexicon or semantic content to rapidly narrow down the scope of possible candidates. For instance, when an expression contains strongly correlated words such as “dengue” and “total” at the same time, it is natural to conclude that this is probably an expression about a dengue case. These lexical indicators can help explain how humans can skim through an article to quickly capture the dengue case information expressions. Therefore, we used log-likelihood ratio (LLR); it is an effective feature selection method that can generate representative patterns from sequences of dengue case information expressions [42-44]:

$$LLR(wp, D) = 2\log\frac{p(wp|D)^k\left(1-p(wp|D)\right)^m p(wp|\neg D)^l\left(1-p(wp|\neg D)\right)^n}{p(wp)^{k+l}\left(1-p(wp)\right)^{m+n}} \quad \textbf{(12)}$$

Given a training data set composed of binary labels for representing sentences that describe dengue case information (*D*) or not (¬*D*), we pair words from sentences to generate a set of co-occurring word pairs WP = $\{wp_1, ..., wp_f\}$ and POS pairs PP = $\{pp_1, ..., pp_g\}$. The LLR uses the following mechanism to calculate the likelihood that the occurrence of a word pair and a POS pair in the dengue case information is not random. To illustrate, we take the LLR calculation for a word pair, where *N*(*D*) and *N*(¬*D*) are the numbers of positive and negative sentences, respectively. *N*(*wp*^D), which is denoted as *k*, is the number of alert articles containing *wp* and *D* simultaneously, whereas *N*(*wp*^¬*D*), which is denoted as *l*, is the number of negative sentences that include *wp*. To further simplify the formula, we also define $m = N(D) - k$ as the number of sentences containing *D* without the word pair *wp* and $n = N(\neg D) - l$, which means the number of sentences with neither *D* nor *wp*. A maximum likelihood estimation is conducted to obtain probabilities *p*(*wp*), *p*(*w*|*D*), and *p*(*w*|^*D*) by calculating the log-likelihood of the hypothesis that the presence of *wp* in set *D* is not random. A word pair with a large LLR value is therefore closely associated with the expression of dengue case information. We rank all the word pairs by every LLR value in the training data, and the top 50 word pairs that describe dengue case information and those that do not are selected as linguistic patterns for positive and negative sentences, respectively. The

same procedure is adopted to calculate the LLR value for the POS pairs for the compilation of POS patterns.

Next, we integrate the generated linguistic patterns and POS patterns into a DuBiLSTM with an attention mechanism by concatenating both positive and negative vectors (which are composed of 60 and 25 dimensions, respectively) with the LLR value of the matched patterns (ie, 170D pattern vectors). As the LLR values of linguistic patterns and POS patterns indicate the weight associated with positive and negative sentences, merging the linguistic pattern and POS pattern features into a DuBiLSTM with an attention mechanism is discriminative.

Finally, the direct splicing strategy is used to fuse pattern features with latent semantic features from a DuBiLSTM with an attention mechanism neural network. The calculation formula for this is as follows:

$$F = PF \oplus LSF \quad \textbf{(13)}$$

where PF=pattern features and LSF=latent semantic features.

Dengue Case Information Classifier

The final step in our framework involves constructing a classifier to predict the labels through the 3 fully connected layers and the activation layer and then to output the distribution probability of the labels. In the fully connected layer, the model maps the fused feature vector to the instance label space. In the output layer, the softmax function is used for normalization, and the output of the fully connected layer is converted into the approximate probability value y for each category. The calculation formula is as follows:

$$\hat{y} = softmax(MF^T + b) \quad \textbf{(14)}$$

where M is the parameter matrix of the connection layer, F is the characterization of the fusion-distributed characteristics, b is the bias, and softmax is a normalization function. Although the 3 fully connected layers increase the computational cost, the classifier efficiently learns weights through the neuron layer [45-47]. The neurons in each layer will be connected to the neurons in the next layer. Considering the convergence rate, the rectified linear unit (ReLU) function is used as an activation function for nonlinear operation. This can easily cause overfitting in model learning; therefore, to avoid this, we use the dropout mechanism to correct for overfitting [48]. The 2 probabilities are predicted for negative and positive, and the larger probability will be taken out to become the final prediction result through the softmax function.

In addition, the Adam optimizer [49] was chosen to optimize the loss function of the network. The model parameters are fine-tuned by the Adam optimizer, which has been shown to be an effective and efficient backpropagation algorithm. We use the cross-entropy function as the loss function because it can reduce the risk of a gradient disappearance during the process of stochastic gradient descent; this is why it often performs better than the classification error rate or the mean square error [50]. The loss rate of the model can be calculated using the following equation:

$$Loss = -\sum_{i=1}^{N} y_i \times \log \hat{y}_i + (1 - y_i) \times \log(1 - \hat{y}_i) \quad \textbf{(15)}$$

where N is the number of training samples, y is the label of the sample, and $\hat{y}$ is the output of the model.

Comparative Analysis Models

To conduct a comprehensive evaluation of the proposed method, we also developed baselines of the machine learning model and deep neural networks to estimate the significance of our approach and the performance variation in different classification systems. Our first baseline uses a tokenized representation evaluated on the radial basis function kernel–based support vector machine (SVM) [51]. This system learns the statistical relevance of each token in a clinical record within different syntactic and semantic contexts. Next, 2 ensemble learning approaches were also implemented for comparison. The first is random forest (RF), an ensemble learning method for classification that constructs a multitude of decision trees (DTs) adopting term frequency–inverse document frequency text representation. The other model is the classifier extreme gradient boosting (XGB), which is a gradient boosting DT that integrates multiple learners for classification problems [52]. We included XGB in this study because it has been validated on real-life large-scale imbalanced data sets and solves many data science–related problems in a fast and accurate way [53]. Finally, the 3 deep neural networks were compared for performance evaluation. The first deep learning model is a class of feedforward artificial neural networks called multilayer perceptron (MLP) [54]. We constructed an MLP through 3 fully connected dense layers afterward the input layer (the model was constructed by adding 3 hidden layers between the input layer and the output layer, where hidden layers are responsible for feature extraction to help output layer to do classification work). In addition, we also adopted a whole-text multikernel CNN model using static word embeddings [55,56] of instances (CNN for text) as another baseline. The last deep learning model that we included in the comparative analysis is the LSTM recurrent neural network, which is capable of learning order dependence in a text sequence and is widely used in NLP research.

To examine the incremental performance that benefits from the proposed method, the BiLSTM and DuBiLSTM are also listed as comparators. To serve as a basis for comparison, we also included naïve Bayes [51,57-59] and DT [51] as baselines.

Results

Evaluation Metrics

In our experiments, the performance evaluation metrics included precision, recall, and F1 score. In general, there is a trade-off between precision and recall. As the 2 metrics evaluate system performance from different perspectives, a single metric that balances (ie, averages) the trade-off is essential. The F1 score is the harmonic mean of precision and recall, and as it is generally close to the minimum of the 2 values, it can be considered an attempt to find the best possible compromise (balance) between precision and recall [51]. The F1 score is

also deemed a conservative metric that prevents the possible overestimation of system performance because the harmonic mean is always less than, or equal to, the arithmetic mean and geometric mean. For this reason, the F1 score is extensively used to judge the superiority of information systems [51]. We used the macroaverage to compute the average performance, and to obtain reliable verification results, we adopted a 10-fold cross-validation approach [60]. Our model was implemented with Keras [61] under the following configurations: the dropout probability was set at 0.35 after each layer, loss function was categorized as cross-entropy, ReLU activation was applied to the dense layer, and training was set at 40 epochs.

Model Comparisons

The model performance is shown in Table 2. First, the naïve Bayes classifier, which is a conditional probability–based approach using bag-of-words feature space with term frequency–inverse document frequency term weighting, only achieved a mediocre performance. As this classifier only considers surface word weightings, it has difficulty representing interword relations, and its overall F1 score is only 70.34%. By contrast, the DT further learns keyword weighting for representation through an entropy-based feature selection method. Hence, the DT is able to obtain significant improvement. The XGB obtained better performances because it is able to integrate multiple machine learning algorithms using the gradient boosting mechanism to optimize the loss function. Likewise, the RF integrates multiple DTs through ensemble learning to optimize prediction results. Therefore, the overall performance of the XGB and RF are similar, with both achieving F1 scores of approximately 85%. It is worth mentioning that the prediction performance of an SVM achieves an F1 score of approximately 90%. This is because the SVM can solve nonlinear obstacles and build models based on learned word and phrase correlations in context, thereby enhancing classification performance. When comparing the deep learning approaches, the performance of an MLP is similar to the ensemble learning–based approaches. The redundancy and inefficiency might be caused by the large number of parameters in the fully connected neural network structure. The learned weighting is therefore unreliable and leads to the lowest performance among all neural models considered in this study. In contrast to the MLP, CNN for text and LSTM have better performances that achieve F1 scores of approximately 90%. This indicates that both deep neural network models efficiently represent textual information and learn the context of the ProMED alerting articles to identify dengue case information. It is worth noting that our method can extract latent linguistic features from ProMED alerting reports because learned word and POS embeddings are adopted to represent syntactic and context relations. Moreover, we used discriminative patterns to encode the characteristics of collocation relationships to capture descriptors of dengue within the alert reports. Consequently, our method achieves the best overall precision, recall, and F1 score among the compared methods.

Table 2. The performance results of the compared methods.

System	Negative, precision; recall; F1 score (%)	Positive, precision; recall; F1 score (%)	Macroaverage, precision; recall; F1 score (%)	*P* value
NB[a]	81.69; 99.30; 89.64	94.51; 34.96; 51.04	88.10; 67.13; 70.34[b]	<.001
DT[c]	95.79; 85.40; 90.29	67.59; 89.02; 76.84	81.69; 87.21; 83.57[b]	<.001
RF[d]	96.53; 88.87; 92.54	73.60; 90.65; 81.24	85.06; 89.76; 86.89[b]	<.001
SVM[e]	94.12; 95.69; 94.90	86.75; 82.52; 84.58	90.43; 89.10; 89.74[b]	<.001
XGB[f]	92.55; 91.52; 92.03	75.98; 78.46; 77.20	84.26; 84.99; 84.61[b]	<.001
MLP[g]	94.64; 90.82; 92.69	76.00; 84.96; 80.23	85.32; 87.89; 86.46[b]	<.001
CNN[h] for text	94.47; 94.99; 94.73	85.12; 83.74; 84.43	89.80; 89.37; 89.58[b]	<.001
LSTM[i]	94.72; 94.85; 94.79	84.90; 84.55; 84.73	89.81; 89.70; 89.76[b]	<.001
BiLSTM[j]	95.74; 93.88; 94.80	83.08; 87.80; 85.38	89.41; 90.84; 90.09[k]	.94
DuBiLSTM[l]	95.89; 94.16; 95.02	83.78; 88.21; 85.94	89.84; 91.18; 90.48[k]	.95
Our method	97.72; 95.27; 96.48	87.12; 93.50; 90.20	92.42; 94.38; 93.34	—[m]

[a]NB: naïve Bayes.

[b]*P*<.001 (a chi-square test was applied to determine whether our method significantly improves performance in comparison with other methods).

[c]DT: decision tree.

[d]RF: random forest.

[e]SVM: support vector machine.

[f]XGB: extreme gradient boosting.

[g]MLP: multilayer perceptron.

[h]CNN: convolutional neural network.

[i]LSTM: long short-term memory.

[j]BiLSTM: bidirectional long short-term memory.

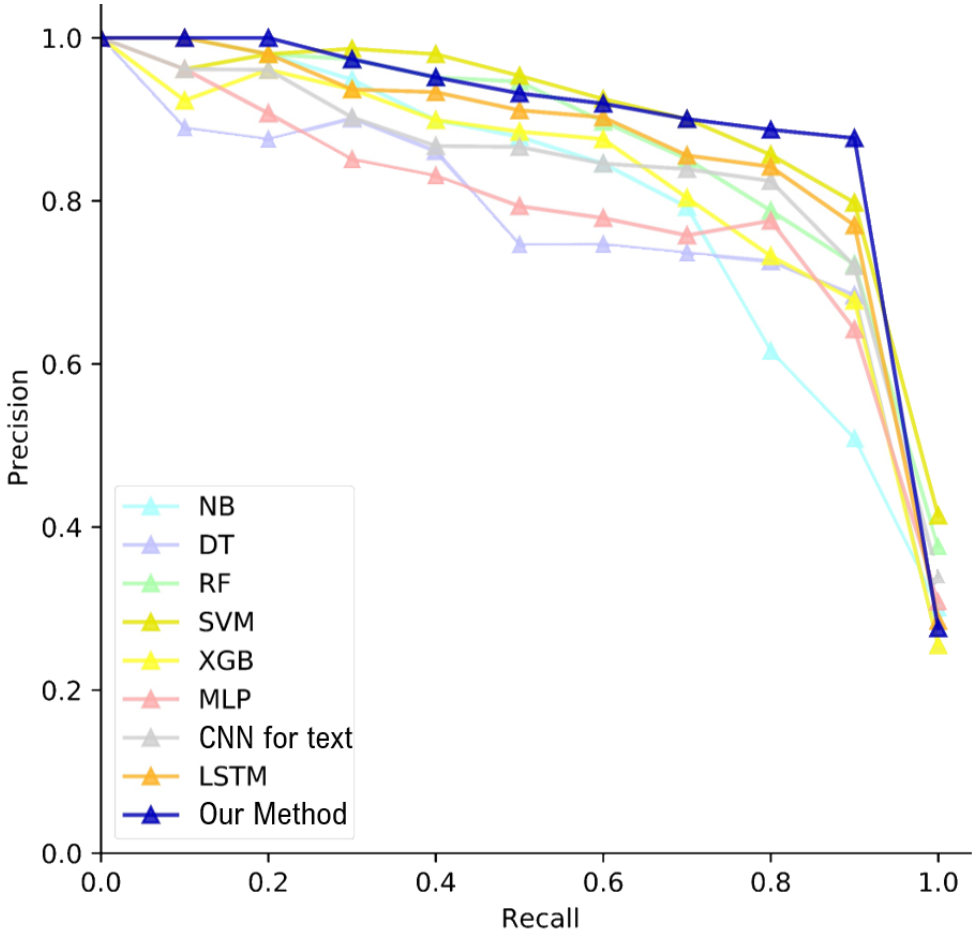
[k]*P*>.05 (a chi-square test was applied to determine whether our method significantly improves performance in comparison with other methods).

[l]DuBiLSTM: dual-channel bidirectional long short-term memory.

[m]Not available.

Finally, we evaluated the performances of the compared methods using 11-point precision recall curves [62]. To plot these curves, the evaluated sentences were sorted according to their prediction scores. Figure 2 shows that the precision scores of our method at the 11 recall levels are superior to those of the compared methods. In other words, our method is able to most accurately extract sentences that convey dengue case information.

Figure 2. The precision recall curves of the compared methods. CNN: convolutional neural network; DT: decision tree; LSTM: long short-term memory; MLP: multilayer perceptron; NB: naïve Bayes; RF: random forest; SVM: support vector machine; XGB: extreme gradient boosting.



To summarize, the DuBiLSTM is able to learn the latent syntactic and semantic information of a text, and the attention mechanism can further highlight the important elements in context. The linguistic patterns are successfully integrated into the neural network to capture discriminative collocation of latent features. Consequently, our method significantly outperforms the compared methods and achieves a remarkable dengue case information extraction performance.

Discussion

Principal Findings

This study describes a new method for identifying dengue case information by using BiLSTM with an attention mechanism enriched with linguistic patterns. As the results show, BiLSTM can consider contextual information more efficiently. Through its bidirectional mechanism, the output for any current moment is not only related to a previous state but may also be related to a future state. The DuBiLSTM can yield an even more slightly improved overall performance because of the benefits accruing from the enhancement of the precision and recall of both positive and negative categories. This indicates that the dual-channel framework is able to generate more shallow linguistic features for BiLSTM. It is noteworthy that our method achieves the best performance. As the generated linguistic pattern can examine the content of sentences to identify dengue case information, it does not conflict with the DuBiLSTM, which analyzes syntactic and semantic information in the sentences.

As a consequence, combining BiLSTM and DuBiLSTM improves the system performance and achieves a remarkable performance on the Taipei Medical University Dengue Case Corpus. The high proportion of dengue case information expressions can be identified by the generated linguistic patterns. For instance, the positive sentence “Taiwan recorded another 7 cases of dengue fever [Wednesday, September 5, 2018], bringing the total number so far this year to 81 and prompting stronger calls by the relevant authorities for greater public cooperation to prevent the spread of the mosquito-borne disease in the peak season.” is correctly detected as dengue case information sentence through the successful match of the generated pattern *[total]-[dengue]* and *[reported]-[dengue]*. It shows that identifying sentences with matched patterns can enhance the performance to discriminate the dengue case information extraction.

As shown in Figures 3 and 4, we visualized the collocation of words and POS for further observation, where nodes and edges represent the linguistic pattern, with the depth of the edge denoting the weight value (ie, LLR) of the collocation. We can observe the appearance of linguistic patterns in Figure 3, such as *[dengue]-[reported]* and *[locally]-[acquired]*, indicating that the sentence is more likely to be crucial. In addition, we also noticed that the word *case* has radial edges, which suggests

that many discriminative linguistic patterns are composed of the word *case*. This is because case-related information often mentions the term *case*. For instance, "Dengue [reported] 100 cases locally acquired; Municipality most affected: Kaohsiung City," which is a very typical example containing accurate case information. In addition, from Figure 4, we observe from the POS pattern network that the more important POS collocations are *[JJS]-[CD]*, where *JJS* stands for *adjective, superlative* and *CD* stands for *cardinal number*, and *[JJS]-[NN]*, where *NN* stands for *noun, singular or mass* (the detailed meanings of POS tags are provided in the URL [63]. This is because the case numbers mostly occur in digit form, and in this situation, the POS tag belongs to *CD*, which is then combined with adjectives and nouns to complete the description of the case information.

Table 3 lists the errors in the sentences of single-clausal and multiclausal types. As shown in the table, the total error rate of the proposed method is 6% (58/965). The individual error rates of single-clausal and multiclausal sentence types are 3.1% (14/456) and 8.6% (44/509), respectively. This indicates that dengue case information in multiclausal sentences is difficult to detect. This is because the syntactic structures of multiclausal sentences are so complex that they confuse the pattern-matching process. As a result, the matched linguistic and POS patterns are prone to errors that affect the correctness of pattern representation and the performance of the corresponding detection. We also observed from the results that a vast proportion of false positives, that is, negative instances incorrectly identified as positive, occurred because the sentences expressed global dengue case information instead of expressing information from the observed country of the alert. For example, our model incorrectly classified "This alarming, particularly considering 96 million cases of symptomatic dengue year worldwide." as a positive sentence. However, the sentence conveys the aggregation of an annual and global pandemic situation, rather than representing a piece of detailed case information, although it contains the highly associated words *case* and *dengue*. For the false negatives, we noticed that a sentence may be split into several fragments because of the writing style. This can result in an incomplete context and thus unclear semantics behind the text. For instance, the sentence "Locality affected: Tainan 32 past week," the digit "32" actually represents the number of cases. However, because the word *cases* is omitted from the text, our model missed the positive sentence, which increased the false-negative rate.

Figure 3. The network visualization for generated linguistic patterns. CDC: Taiwan Centers for Disease Control.

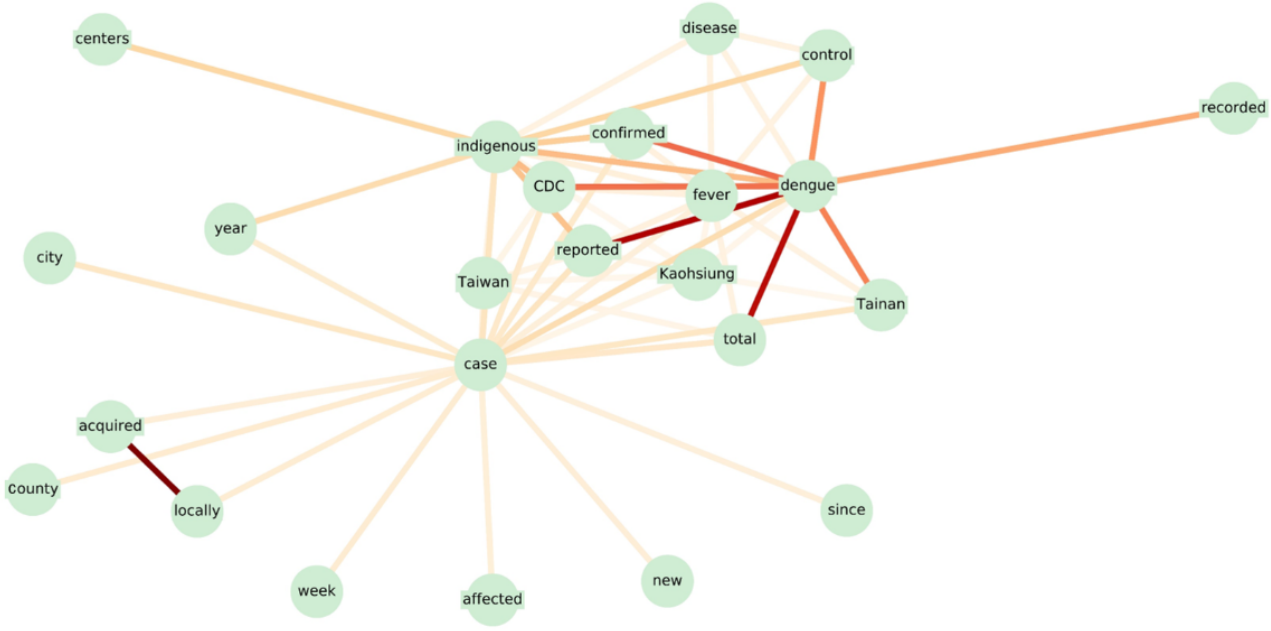


Figure 4. The network visualization for generated parts-of-speech (POS) patterns. CC: coordinating conjunction; CD: cardinal number; DT: determiner; EX: existential there; IN: preposition or subordinating conjunction; JJ: adjective; JJS: adjective, superlative; NN: noun, singular or mass; NNP: proper noun, singular; NNPS: proper noun, plural; NNS: noun, plural; PRP$: possessive pronoun; RB: adverb; RBS: adverb, superlative; TO: to; VBD: verb, past tense; VBG: verb, gerund, or present participle; VBN: verb, past participle; VBP: verb, nonthird person singular present; VBZ: verb, third person singular present; WDT: wh-determiner.

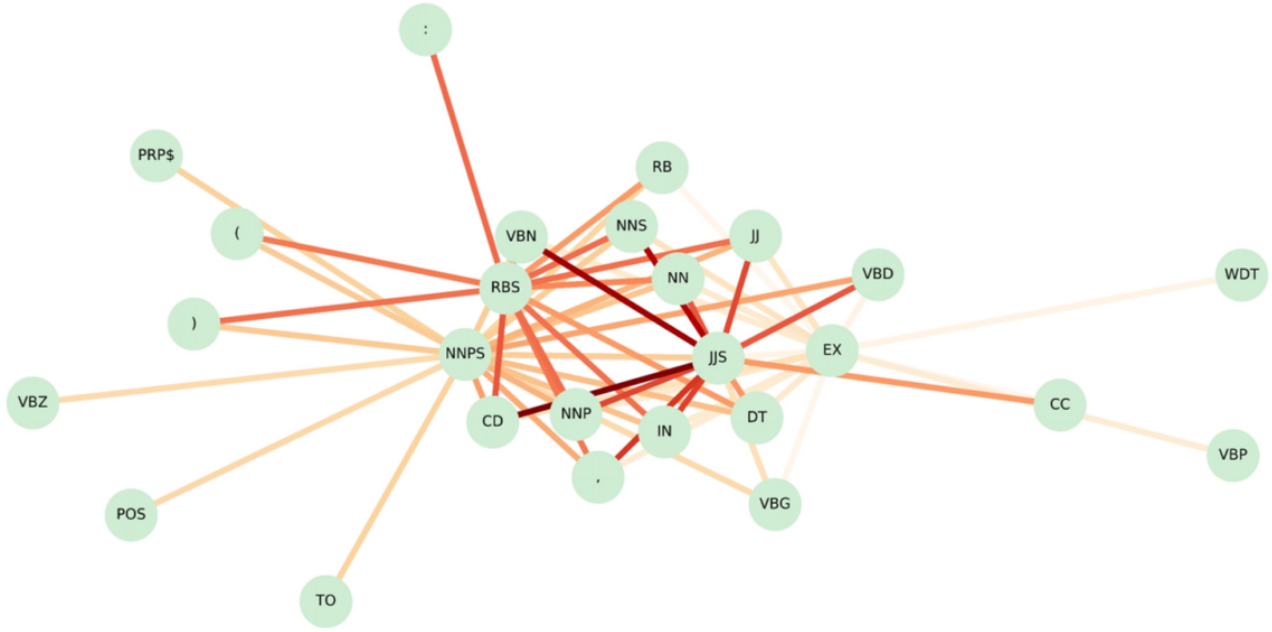


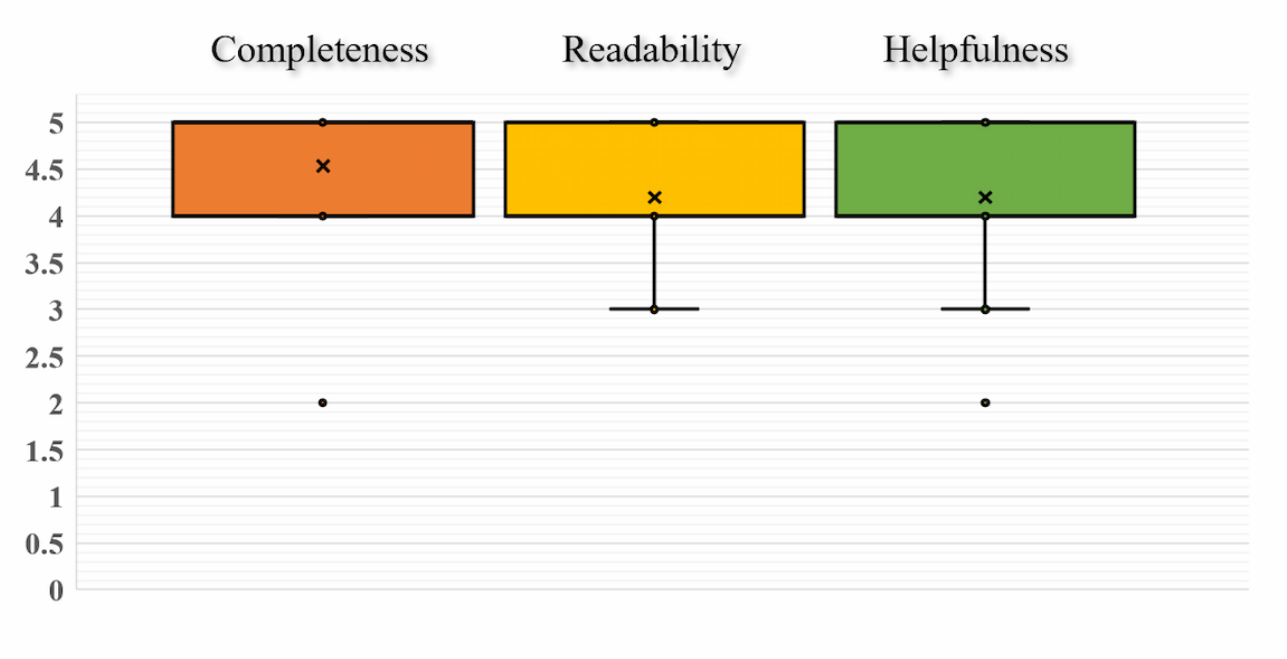
Table 3. Error distribution of dengue case information detection.

Clause type	False positive, n (%)	False negative, n (%)	Error rate, n (%)
Single-clausal (n=456)	7 (1.5)	7 (1.5)	14 (3.1)
Multiclausal (n=509)	29 (5.7)	15 (2.9)	44 (8.6)
Corpus (n=965)	36 (3.7)	22 (2.3)	58 (6)

The goal of this research was to automatically generate a summary of the alerting articles from ProMED-mail to help researchers reduce reading effort and more quickly comprehend the main topic. Given an alerting report, our model can extract crucial sentences that express dengue case information, and these extracted sentences can be combined to form a summary. To estimate the practicality of the proposed model for epidemic monitoring, we conducted a satisfaction analysis experiment to assess the acceptability of the summaries by end users from the medical science field. The survey participants were 2 female students from the School of Public Health, Taipei Medical University; 2 male faculty members from the department of parasitology and tropical diseases, Taipei Medical University; and 1 male internal medicine physician from a Taipei Medical University–affiliated hospital; their ages ranged from 23 to 55 years. They evaluated on a 5-point Likert scale the quality of the summaries generated by our method. The summaries were randomly sampled from 25% (5/20) of the Southeast Asia dengue alerting reports from January 2019 to December 2020. We estimated the average number of words and sentences from both the original alerting reports and the generated summaries and then derived the compression rates at the word level (2.8) and sentence level (3.7). We included three questionnaire items: (1) completeness—*completeness of the generated summary content*, (2) readability—*fluent and easy to read*, and (3) helpfulness—*helps to improve analysis efficiency or reduce text reading time*.

To analyze the bias of the scoring distribution, we used a box plot to illustrate the distribution of the scores in the satisfaction survey (Figure 5). The extended range of the boxed image (including box whiskers) represents the highest to lowest distribution of the 5 scores, and the symbol *x* in the box indicates the mean value of the scores. As we can see, the item *Completeness* has the highest score, which indicates that all the epidemiologists were satisfied with the quality and accuracy of the summary text. The average scores of the remaining 2 items are also >4.2, which demonstrates that our automatic summaries are of high quality. These findings are evidence of the usefulness of this research. Nevertheless, there are still a few outliers with scores of 2 in the satisfaction questionnaire, which means that the epidemiologists were not entirely satisfied with these case summaries. On the basis of further analysis of the results, it was found that this was mainly due to incorrect results of our prediction of multiclausal sentences, which is the main type of error in the proposed model. However, multiclausal sentences typically entail rich information. Therefore, as the quality of the summary is sensitively affected by the incorrect identification of multiclausal sentences, one of our directions for future research is to improve the accuracy of multiclausal sentence prediction.

Figure 5. Box plot of expert assessment on a 5-point Likert scale of the quality of generated summaries.



To summarize, the experiment results from the satisfaction analysis demonstrate that our summary system is helpful for experts and scholars to quickly read and effectively analyze a large number of briefings.

Limitations and Future Directions

This study includes some limitations. The approaches developed in this study mainly focus on extracting sentence information for summarizing or, more specifically, extracting qualitative information from unstructured content. However, this approach is currently unable to acquire precise quantitative information. For instance, the number of incidence cases (newly infected) and cumulative cases cannot be reliably identified using current algorithms. Our future work will therefore focus on this issue by integrating date, location, and identification of the number of cases to retrieve important quantities of disease information. This information can be applied to more advanced spatial and temporal analyses in the future.

The second limitation is that the reporting effort and frequency in ProMED-mail are not consistent because of its volunteer-oriented design. This problem could be overcome by integrating other outbreak-reporting platforms such as the HealthMap project, which provides a visualized platform for various disease alerts [53,64]. Collecting epidemiological surveys from the scientific literature is another approach that can be used to enrich the data set.

Conclusions

The combination of high rates of international travel and rapid environmental changes makes region-wide collaboration in monitoring emerging and re-emerging infectious diseases necessary. The current COVID-19 pandemic has also had a huge impact on the surveillance and control of other infectious diseases [65]. In addition, because ProMED-mail records and follows up undiagnosed diseases in different countries [66], this abundant disease surveillance information is unstructured and is thus not able to be efficiently used by public health workers or scientists. Our proposed deep neural network provides a good way to extract outbreak information from unstructured text, which can then be further analyzed.

In summary, our study built a prototype of an NLP algorithm to retrieve sentence summarizations from the ProMED-mail database. This approach can help medical scientists and public health workers to save more time on content summarization and analysis. Further work will continue to optimize the algorithm to extract more important quantification information.

Acknowledgments

This research was supported by the Ministry of Science and Technology of Taiwan (MOST 108-2638-H-002-002-MY2 and MOST 109-2410-H-038-012-MY2).

Authors' Contributions

YCC and TWC helped to design the study and conceived the research question. YCC, TWC, and YWC conducted the experiments and statistical analyses and reviewed and interpreted the findings. YCC, TWC, and YWC wrote the manuscript, reviewed it, and noted the points of revision.

Conflicts of Interest

None declared.

References

1. Johansson MA, Cummings DA, Glass GE. Multiyear climate variability and dengue--El Niño southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: a longitudinal data analysis. PLoS Med 2009 Nov;6(11):e1000168 [FREE Full text] [doi: 10.1371/journal.pmed.1000168] [Medline: 19918363]
2. Kovats RS, Bouma MJ, Hajat S, Worrall E, Haines A. El Niño and health. Lancet 2003 Nov 01;362(9394):1481-1489. [doi: 10.1016/S0140-6736(03)14695-8] [Medline: 14602445]
3. Petersen LR, Jamieson DJ, Powers AM, Honein MA. Zika Virus. N Engl J Med 2016 Apr 21;374(16):1552-1563. [doi: 10.1056/NEJMra1602113] [Medline: 27028561]
4. Patz JA, Olson SH. Malaria risk and temperature: influences from global climate change and local land use practices. Proc Natl Acad Sci U S A 2006 Apr 11;103(15):5635-5636 [FREE Full text] [doi: 10.1073/pnas.0601493103] [Medline: 16595623]
5. Ebi KL, Nealon J. Dengue in a changing climate. Environ Res 2016 Nov;151:115-123 [FREE Full text] [doi: 10.1016/j.envres.2016.07.026] [Medline: 27475051]
6. Chuang T, Soble A, Ntshalintshali N, Mkhonta N, Seyama E, Mthethwa S, et al. Assessment of climate-driven variations in malaria incidence in Swaziland: toward malaria elimination. Malar J 2017 Jun 01;16(1):232 [FREE Full text] [doi: 10.1186/s12936-017-1874-0] [Medline: 28571572]
7. Wesolowski A, Qureshi T, Boni MF, Sundsøy PR, Johansson MA, Rasheed SB, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc Natl Acad Sci U S A 2015 Sep 22;112(38):11887-11892 [FREE Full text] [doi: 10.1073/pnas.1504964112] [Medline: 26351662]
8. ProMED homepage. ProMED. URL: https://promedmail.org/ [accessed 2022-07-04]
9. Morse SS, Rosenberg BH, Woodall J. ProMED global monitoring of emerging diseases: design for a demonstration program. Health Policy 1996 Dec;38(3):135-153 [FREE Full text] [doi: 10.1016/0168-8510(96)00863-9] [Medline: 10162418]
10. Chase V. ProMED: a global early warning system for disease. Environ Health Perspect 1996 Jul;104(7):699 [FREE Full text] [doi: 10.1289/ehp.104-1469400] [Medline: 8841752]
11. Carrion M, Madoff LC. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. Int Health 2017 May 01;9(3):177-183 [FREE Full text] [doi: 10.1093/inthealth/ihx014] [Medline: 28582558]
12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
13. Cai O, Sousa-Pinto B. United States influenza search patterns since the emergence of COVID-19: infodemiology study. JMIR Public Health Surveill 2022 Mar 03;8(3):e32364 [FREE Full text] [doi: 10.2196/32364] [Medline: 34878996]
14. Husnayain A, Chuang T, Fuad A, Su EC. High variability in model performance of Google relative search volumes in spatially clustered COVID-19 areas of the USA. Int J Infect Dis 2021 Aug;109:269-278 [FREE Full text] [doi: 10.1016/j.ijid.2021.07.031] [Medline: 34273513]
15. Castillo-Salgado C. Trends and directions of global public health surveillance. Epidemiol Rev 2010;32(1):93-109. [doi: 10.1093/epirev/mxq008] [Medline: 20534776]
16. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis 2004 Jul 15;39(2):227-232. [doi: 10.1086/422003] [Medline: 15307032]
17. Lorthe TS, Pollack MP, Lassmann B, Brownstein JS, Cohn E, Divi N, et al. Evaluation of the EpiCore outbreak verification system. Bull World Health Organ 2018 May 01;96(5):327-334 [FREE Full text] [doi: 10.2471/BLT.17.207225] [Medline: 29875517]
18. Milne-Price S, Miazgowicz KL, Munster VJ. The emergence of the Middle East respiratory syndrome coronavirus. Pathog Dis 2014 Jul;71(2):121-136 [FREE Full text] [doi: 10.1111/2049-632X.12166] [Medline: 24585737]
19. Lessler J, Moore SM, Luquero FJ, McKay HS, Grais R, Henkens M, et al. Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. Lancet 2018 May 12;391(10133):1908-1915 [FREE Full text] [doi: 10.1016/S0140-6736(17)33050-7] [Medline: 29502905]
20. Tarnas MC, Desai AN, Lassmann B, Abbara A. Increase in vector-borne disease reporting affecting humans and animals in Syria and neighboring countries after the onset of conflict: a ProMED analysis 2003-2018. Int J Infect Dis 2021 Jan;102:103-109 [FREE Full text] [doi: 10.1016/j.ijid.2020.09.1453] [Medline: 33002614]
21. Chatziprodromidou IP, Arvanitidou M, Guitian J, Apostolou T, Vantarakis G, Vantarakis A. Global avian influenza outbreaks 2010-2016: a systematic review of their distribution, avian species and virus subtype. Syst Rev 2018 Jan 25;7(1):17 [FREE Full text] [doi: 10.1186/s13643-018-0691-z] [Medline: 29368637]
22. Izquierdo JL, Ancochea J, Savana COVID-19 Research Group, Soriano JB. Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: retrospective study using machine learning and natural language processing. J Med Internet Res 2020 Oct 28;22(10):e21801 [FREE Full text] [doi: 10.2196/21801] [Medline: 33090964]

23. Jiao Y, Sharma A, Ben Abdallah A, Maddox TM, Kannampallil T. Probabilistic forecasting of surgical case duration using machine learning: model development and validation. J Am Med Inform Assoc 2020 Dec 09;27(12):1885-1893 [FREE Full text] [doi: 10.1093/jamia/ocaa140] [Medline: 33031543]
24. Manning C, Schutze H. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts, United States: MIT press; 1999.
25. Kim M, Chae K, Lee S, Jang H, Kim S. Automated classification of online sources for infectious disease occurrences using machine-learning-based natural language processing approaches. Int J Environ Res Public Health 2020 Dec 17;17(24):9467 [FREE Full text] [doi: 10.3390/ijerph17249467] [Medline: 33348764]
26. Fernandes P, Allamanis M, Brockschmidt M. Structured neural summarization. In: Proceedings of the Structured neural summarization. 2019 Presented at: The International Conference on Learning Representations (ICLR); May 6 - 9, 2019; New Orleans URL: https://arxiv.org/abs/1811.01824
27. Liu G, Guo J. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing 2019 Apr;337:325-338. [doi: 10.1016/j.neucom.2019.01.078]
28. Murray NE, Quam MB, Wilder-Smith A. Epidemiology of dengue: past, present and future prospects. Clin Epidemiol 2013;5:299-309 [FREE Full text] [doi: 10.2147/CLEP.S34440] [Medline: 23990732]
29. Guzman MG, Halstead SB, Artsob H, Buchy P, Farrar J, Gubler DJ, et al. Dengue: a continuing global threat. Nat Rev Microbiol 2010 Dec;8(12 Suppl):S7-16 [FREE Full text] [doi: 10.1038/nrmicro2460] [Medline: 21079655]
30. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature 2013 Apr 25;496(7446):504-507 [FREE Full text] [doi: 10.1038/nature12060] [Medline: 23563266]
31. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS Negl Trop Dis 2012;6(8):e1760 [FREE Full text] [doi: 10.1371/journal.pntd.0001760] [Medline: 22880140]
32. Medical record. Wikipedia. URL: https://en.wikipedia.org/wiki/Medical_record [accessed 2022-07-04]
33. Zhai C, Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval. SIGIR Forum 2017 Aug 02;51(2):268-276. [doi: 10.1145/3130348.3130377]
34. Krallinger M, Rabal O, Lourenço A, Oyarzabal J, Valencia A. Information retrieval and text mining technologies for chemistry. Chem Rev 2017 Jun 28;117(12):7673-7761. [doi: 10.1021/acs.chemrev.6b00851] [Medline: 28475312]
35. Marcos-Pablos S, García-Peñalvo FJ. Information retrieval methodology for aiding scientific database search. Soft Comput 2018 Oct 12;24(8):5551-5560. [doi: 10.1007/s00500-018-3568-0]
36. Ricardo BY, Berthier RN. Modern Information Retrieval. London, United Kingdom: Pearson Education; 1999.
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310]
38. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276-282. [doi: 10.11613/bm.2012.031]
39. Guba E, Lincoln Y. Effective Evaluation: Improving the Usefulness of Evaluation Results Through Responsive and Naturalistic Approaches. San Francisco, California, United States: Jossey-Bass; Sep 03, 2016.
40. TMUDCC: Taipei Medical University Dengue Case Corpus. Taipei Medical University. URL: http://nlp.tmu.edu.tw/ProMED/TMUDCC/index.html [accessed 2022-07-04]
41. Natural Language Toolkit. NLTK. URL: https://www.nltk.org/ [accessed 2022-07-04]
42. Schütze H, Manning C, Raghavan P. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008.
43. Cheng Y, Chen Y, Yeh W, Chang Y. Valence and arousal-infused bi-directional LSTM for sentiment analysis of government social media management. Applied Sci 2021 Jan 19;11(2):880. [doi: 10.3390/app11020880]
44. Chen C, Warikoo N, Chang YC, Chen JH, Hsu WL. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. J Am Med Inform Assoc 2019 Nov 01;26(11):1227-1236 [FREE Full text] [doi: 10.1093/jamia/ocz128] [Medline: 31390470]
45. Wang H, Shi H, Lin K, Qin C, Zhao L, Huang Y, et al. A high-precision arrhythmia classification method based on dual fully connected neural network. Biomedical Signal Process Control 2020 Apr;58:101874. [doi: 10.1016/j.bspc.2020.101874]
46. Hussain S, Mokhtar M, Howe JM. Sensor failure detection, identification, and accommodation using fully connected cascade neural network. IEEE Trans Ind Electron 2015 Mar;62(3):1683-1692. [doi: 10.1109/tie.2014.2361600]
47. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct;323(6088):533-536. [doi: 10.1038/323533a0]
48. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929-1958. [doi: 10.5555/2627435.2670313]
49. Kingma D, Ba JL. Adam: a method for stochastic optimization. In: Proceedings of the ICLR 2015. 2015 Presented at: ICLR 2015; May 7-9, 2015; San Diego, CA, USA.
50. Shore J, Johnson R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans Inform Theory 1980 Jan;26(1):26-37. [doi: 10.1109/tit.1980.1056144]
51. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. Burlington, Massachusetts, United States: Morgan Kaufmann; 2012.

52. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13 - 17, 2016; San Francisco, California, USA URL: https://dl.acm.org/doi/10.1145/2939672.2939785 [doi: 10.1145/2939672.2939785]
53. Fan J, Wang X, Wu L, Zhou H, Zhang F, Yu X, et al. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Conversion Manag 2018 May;164:102-111. [doi: 10.1016/j.enconman.2018.02.087]
54. Rossi F, Conan-Guez B. Functional multi-layer perceptron: a non-linear tool for functional data analysis. Neural Netw 2005 Jan;18(1):45-60. [doi: 10.1016/j.neunet.2004.07.001] [Medline: 15649661]
55. Liao S, Wang J, Yu R, Sato K, Cheng Z. CNN for situations understanding based on sentiment analysis of twitter data. Procedia Comput Sci 2017;111:376-381. [doi: 10.1016/j.procs.2017.06.037]
56. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); Oct 25–29, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1181]
57. Rish I. An empirical study of the naive Bayes classifier. In: Proceedings of the IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001 Presented at: IJCAI 2001 workshop on empirical methods in artificial intelligence; Aug 4 - 10, 2001; Seattle, Washington, USA URL: https://dominoweb.draco.res.ibm.com/db24eb109a77428785256aff005d3df2.html
58. Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? Int Stat Rev 2001 Dec;69(3):385-398. [doi: 10.1111/j.1751-5823.2001.tb00465.x]
59. Webb GI, Boughton JR, Wang Z. Not so Naive Bayes: aggregating one-dependence estimators. Mach Learn 2005 Jan;58(1):5-24. [doi: 10.1007/s10994-005-4258-6]
60. Mitchell T. Machine Learning. New York, United States: McGraw-Hill Education; 1997.
61. Keras homepage. Keras. URL: https://keras.io/ [accessed 2022-07-04]
62. Hull D. Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. 1993 Presented at: SIGIR93: 16th International ACM/SIGIR '93 Conference on Research and Development in Information Retrieval; Jun 27 - Jul 1, 1993; Pittsburgh Pennsylvania USA. [doi: 10.1145/160688.160758]
63. Alphabetical list of part-of-speech tags used in the Penn Treebank Project. Penn Treebank. URL: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html [accessed 2022-07-04]
64. Nelson R. HealthMap: the future of infectious diseases surveillance? Lancet Infectious Diseases 2008 Oct 21;8(10):596. [doi: 10.1016/S1473-3099(08)70219-6]
65. Valentin S, Mercier A, Lancelot R, Roche M, Arsevska E. Monitoring online media reports for early detection of unknown diseases: insight from a retrospective study of COVID-19 emergence. Transbound Emerg Dis 2021 May;68(3):981-986 [FREE Full text] [doi: 10.1111/tbed.13738] [Medline: 32683774]
66. Rolland C, Lazarus C, Giese C, Monate B, Travert A, Salomon J. Early detection of public health emergencies of international concern through undiagnosed disease reports in ProMED-Mail. Emerg Infect Dis 2020 Feb;26(2):336-339 [FREE Full text] [doi: 10.3201/eid2602.191043] [Medline: 31961311]

Abbreviations

BiLSTM: bidirectional long short-term memory
CNN: convolutional neural network
DHF: dengue hemorrhagic fever
DT: decision tree
DuBiLSTM: dual-channel bidirectional long short-term memory
LLR: log-likelihood ratio
LSTM: long short-term memory
MLP: multilayer perceptron
NLP: natural language processing
POS: parts of speech
ProMED: Program for Monitoring Emerging Diseases
ReLU: rectified linear unit
RF: random forest
SVM: support vector machine
XGB: extreme gradient boosting

Edited by H Bradley; submitted 30.10.21; peer-reviewed by K Chih Hao, S Doan; comments to author 23.02.22; revised version received 15.04.22; accepted 27.05.22; published 13.07.22

Please cite as:
Chang YC, Chiu YW, Chuang TW
Linguistic Pattern–Infused Dual-Channel Bidirectional Long Short-term Memory With Attention for Dengue Case Summary Generation From the Program for Monitoring Emerging Diseases–Mail Database: Algorithm Development Study
JMIR Public Health Surveill 2022;8(7):e34583
URL: https://publichealth.jmir.org/2022/7/e34583
doi: 10.2196/34583
PMID: 35830225