Open Access

Open Journal of Internet of Things (OJIOT)
Volume 6, Issue 1, 2020
http://www.ronpub.com/ojiot
ISSN 2364-7108

Assuring Privacy-Preservation in Mining Medical Text Materials for COVID-19 Cases - A Natural Language Processing Perspective

Bo Ma [A], Jinsong Wu [B,C], Shuang Song [A], William Liu [A]

[A]Department of Information Technology and Software Engineering,
Auckland University of Technology, Auckland, New Zealand
[B]School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China
[C]Department of Electrical Engineering, Universidad de Chile, Santiago, Chile
bo.ma@aut.ac.nz, wujs@ieee.org, sue.song@autuni.ac.nz,william.liu@aut.ac.nz

Abstract

Currently, there is a very large volume of Covid-19 related medical data that have been stored in cloud based systems and made available for studing the disease dynamics. without any privacy-preservation. In order to reduce possible privacy leakage and also accommodate massive medical reports with high efficiencies, we proposed a privacy-preserving word embody-based text classification method for mining COVID-19 medical documents. It uses the recurrent neural network deep learning algorithm according to the identified internal hiding centralization pattern. In addition, a new model-fusion method is proposed for the continuous improvement of the system performance.The extensive numerical studies have demonstrated that the classifier of the proposed system has superior performance via integrating with the keywords extraction approach. Moreover, the advanced new model does not only accurately capture the keyword patterns but also effectively capture the analogical hierarchy structure of the pathology related datasets with lower computational complexity.

Type of Paper and Keywords

Regular research paper: *Big data, Internet of Thing, Infectious disease surveillance, Natural Language Processing (NLP), Word2vec, Vector Space Model (VSM), Privacy-Preserving (PP), Gaussian distributed independent frequently subsequence extraction algorithm (GDIFSEA)*

1 Introduction

In traditional infectious disease text processing, public health data has been extensively studied, which including clinical, hospital, interview and census data, to create a robust model to understand [4]. In addition, a crowdsourcing natural language processing approach offers a holistic and nearly real-time information processing from diagnosed texts or records [1].

However, crowdsourcing word mining systems present their own challenges and there are two most significant concerns. Firstly, the total volume of the corwdsourced texts are usually very large. Moreover, there are many sources of original documents, such as medical papers, medical records and pathological statistics, that manually extracting the required databases

This paper is accepted at the *International Workshop on Very Large Internet of Things (VLIoT 2020)* in conjunction with the VLDB 2020 conference in Tokyo, Japan. The proceedings of VLIoT@VLDB 2020 are published in the Open Journal of Internet of Things (OJIOT) as special issue.

to classified the infections disease records is a very time-consuming and expensive process. Secondly, protecting those involved privacy in the records, while still informing the public. Many crowdsourced word mining systems contain not only a description of the words, but also the individuals, personal sensitive information, such as contact number, home address and disease of the person involved. Therefore, there are more and more concerns about the possible privacy leakages which critically barrier the wide adoptions of any Internet based system.

In order to process massive crowdsourced data with high efficiency, some research works have proposed word embedding techniques associated with natural language processing rules for text classification [7]. The main purpose of text classification is to categorize documents according to the subjects or themes of the documents. According to a certain classification principle the machine can classify texts automatically [2]. Text categorization is widely used by industries. One of the most important applications is to filter the news and figure out the information interesting readers.

2 Proposed of Gaussian Distributedindependent Frequently Subsequence Extraction Algorithm (GDIFSEA)

However, here still have two challenge, one is most medicine documents are write by paper, second is collection time need long time. To address the above two technical challenges, we propose to use the Optical Character Recognition (OCR) technology in the IoT to collect and organize medical data from different sources, for participatory COVID-19 clinical report, as shown below:

The main theoretical findings in this paper include: (1) Based on the NGRAM correlation graph, The paper proposes the Privacy Preserving Gaussian distributed independent frequently subsequence extraction algorithm (ppGDIFSEA), which can be used to optimize the results of the multi-word unit (MWU) and use the interactive information to optimize the result; (2) Under the positive documents condition, we integrate the keyword extraction techniques to solve the problem of semi-supervised learning under ppGDIFSEA, and ppGDIFSEA using Recurrent neural network to classified the sensitive words in the documents. Last, we compare with other supervised learning model Support Vector Machine and in the algorithm performance,as accuracy rate, has been verified.

3 Privacy Preserving with Classification Approach

3.1 Related Definitions

If any character string $\gamma \supset \alpha$ has arbitrary string support $(\gamma) <$ support (α), it is called a frequent α closed sub-string S. Both are called frequent closed strings. According to this assumption, the definition is shows as follow:

Definition 1. *Define the input string as* $\gamma =< \gamma_1, \gamma_2, \cdots, \gamma_N >\in \mathbb{N}$ *, and comparison string* $\alpha =< \alpha_1, \alpha_2, \cdots, \alpha_M >\in \mathbb{M}$. *If there is an parameter as integer* t*, for any integer* $i \in [1, N], \gamma_{t+i} = \alpha_i$ *and here is an other parameter as integer* $x \in [0, t], y \in [1, M - N - t]$*, the string have relationship with parameter integer* t*, result as* $S =< \gamma_0, \gamma_1, \cdots, \gamma_t, \gamma_{t+1}, \gamma_{t+2}, \cdots, \gamma_{t+N}, \gamma_{t+N+1}, \cdots, \gamma_{t+N+y} >$*, the string is satisfied with condition* $S \notin C$*,it is called* α *is* γ *an independent substring under the limit* C*. The number of parameter integers* t *also satisfies the condition* γ *, this is called the number of* α *independent matches under the limit* C*.*

We combine the examples to explain the above concepts.

Assumption 1. *S1="An enzyme-linked immunosorbent assay using canine coronavirus-infected CRFK cells as antigen for detection of anti-coronavirus antibody in cat"*
S2 = "Pathogenic porcine respiratory coronavirus"
S3 = "Canine corona virus vaccine"
S4= "Cellular immune status of coronavirus"
x= "coronavirus"

The concept of substrings is reflection of the concept of strings, such as $l_1 \supset x$ and $l_1 \notin l_2$ so on. The number of matches is the number of substrings in the string, such as $\mathrm{match}\,(l_1, x) = 1$ and $\mathrm{match}\,(l_4, l) = 2$.

For the set $l = \{l_1, l_2, l_3, l_4\}$, set the threshold $\xi = 3$, then frequent substrings and their coordination are "COVID-19":5,"Coronavirus":3, "Coronavirus":3.

However, in fact, when looking at various strings, it can be found that, wherever there is a "Coronavirus", the substring of "Virus" actually appears, since the "Coronavirus" is a substring of "Virus", it leads the degree of coordination to "COVID-19 "must not be less than that to "Coronavirus". It is to solve this problem that the concept of frequent closed substrings is proposed as coordination(Virus)= coordination(Coronavirus) , which can filter out the "Virus "character combination. Therefore, frequent closed substrings and their support are "COVID-19": 5, "Coronavirus ": 3. It can be seen that, for relatively frequent sequences, frequently closed

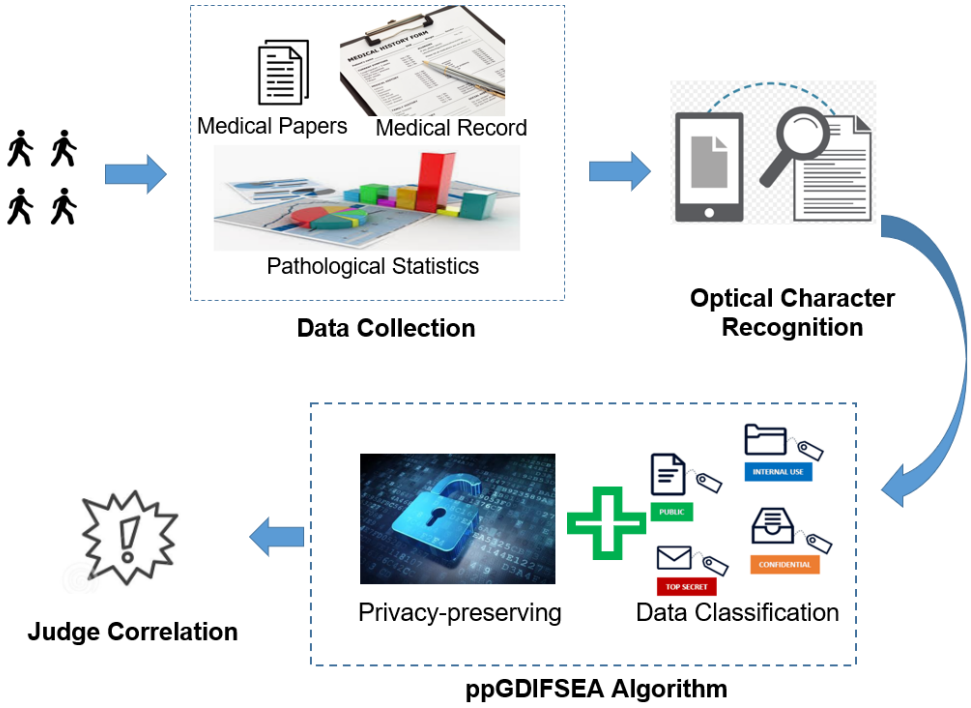

Figure 1: Word Embodying based Crowdsourcing ppGDIFSEA Architecture

substrings are more refined and accurate without loss of information.

3.2 Privacy-Preserving Gaussian Distributed Independent Frequent Subsequence Extraction Algorithm to Eliminate Limitations

Frequently independent substring mining requires independent calculations of string support, which can help to eliminate limitation of word frequency vector models. This section f_{wi} will discuss the reason and improvement to calculate independent support.

We record the frequency of occurrence F_{w_i} of the vocabulary (support) v_i , and f_{v_i} record the frequency of independent vocabulary v_i(independent support). Obviously, for any vocabulary v_i , we have

$$\sum_{\mathrm{v}_i} f_v = S_{v_i}. \tag{1}$$

Note that

$$\zeta(i,j) = \begin{cases} 1, & v_j \supseteq v_i \\ 0, & v_j \supsetneq v_i \end{cases} \tag{2}$$

Among them, the support degree of frequent substrings can be obtained by via counting the character strings generated in the N-GRAM process. For example,

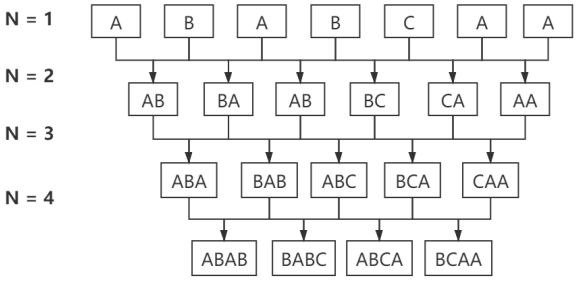

Figure 2: N-GRAM generation process

for a $s_1 = ABABCAA$, we take the maximum length $L = 4$, that is, the N -GRAM is $N = \{1, 2, 3, 4\}$ taken separately. The generated substring is shown in the following figure:

Utility three properties of N-gram associate with privacy-preserving approach, we can generate the new algorithm and the name as Privacy-Preserving Gaussian distributed independent frequently subsequence extraction algorithm(GDIFSEA) (α).The pseudo-code of the algorithm is shown in Table1:

For engineering implementation, the construction of N-GRAM correlation graphs, the screening of Gaussian distributed independent frequently subsequence extraction algorithm(GDIFSEA) graphs based on

Algorithm 1 Privacy-Preserving Gaussian distributed independent frequently subsequence extraction algorithm(ppGDIFSEA)

Input string set $S = \{s_1, s_2, \cdots, s_N\}$,
$s_i =< s_{i,1}, s_{i,2}, \cdots, s_{i,m} >$
support threshold ξ
Output Independent frequent substring set F

procedure INIT EMPTY DAG G
PROCESS DOCUMENTS VIA WORD2VEC

    for Each s_i in S **do**
        for $l = 1$ to L **do**
            for $j = 1$ to $\|s_i\| - l$ **do**
                Set
                $x =< S_{i,j}, S_{i,j+1}, \cdots, S_{i,j+l-1} >$
                Set
                $p1 =< s_{i,j}, s_{i,j+1}, \cdots, s_{i,j+l-1} >$
                Set
                $p2 =< s_{i,j+1}, s_{i,j+2}, \cdots, s_{i,j+l} >$
                G.Add_Vertex_If_Not_Exist (x)
                G.Update_Edge (x, 1)
                G.Add_Noise:$Gau(b)$
            end for
        end for
    end for

    Set $X = G.Get - Vertices0$
    Set $X = \{x_1, x_2, \cdots, x_K\}$
    Sort X By $\|x_i\|$
    for Each x_i IN X **do**
        if $G.Get\_Edgex_i < \xi$ **then**
            $G.Remove\_All\_Child\_Node\,(x_i)$
            G.Remove Vertex (x_i)
        end if
    end for
    for $l = L$ to 1 **do**
        for Each x_i In X On $\|x_i\| = l$ **do**
            if Difsea $(x_i) \geq \xi$ **then**
                $F.Add(x_i)$
            end if
        end for
    end for
    F $\Rightarrow$ RNN(F)
return F
end procedure

GDIFSEA thresholds, and the calculation of the independent GDIFSEA degree can all be performed concurrently, thereby significantly increasing the computational speed.

4 PERFORMANCE EVALUATION

4.1 Testing Privacy of the Reconstruction of Mean Squared Error (rMSE) Distribution with ppGDIFSEA

We test the privacy logic using datasets in two steps:

1. Privacy similar words(PSW): This is a standard for measuring privacy level. Its purpose is to identify privacy information after classification. Most of the sets also appear to be unchanged in the classification, which appear to be a good introduction. A typical example is to calculate the similarity of privacy words between the COVID-19 datasets (processed by ppGDIFSEA) and the Medical Transcription corpus [6] given to the model (processed via Auto Encoder);

2. Data Utility: The main purpose of privacy-preserving GDIFSEA programs is to protect data privacy capabilities to simplify their analysis while preserving confidentiality and facilitating the distribution of samples for other scholars. Therefore, we review the data strengths of the compliance types we learned via applying them into the steps below.

The adoption of mark and privacy relevant mark can be used to update the training set of the supervised learning classifier. After retraining on a privacy processed sample set containing more samples, the obtained classifier tends to have better performance and makes the class label more accurate. At the same time, we can check privacy data inside of control set, find out how many privacy information has been found out inside the control dataset.

Second, the annotated sample can be used to update the keyword list and fine-tune the boundary of the privacy processed classification approach, thereby further reducing the privacy relevant information.

- Hyper-parameter settings
- Hidden layer: (400, 200, 100)
- Activation function: Tanh

With the Privacy similar words(PSW) as privacy test function, the training results are: training set PSW is 0.00578 and Control set PSW is 0.000012, separately.

Using the GDIFSEA, the covid-19 validation set is reconstructed to obtain the privacy similar words of each sample of the validation set. And Accuracy rate is also other important index for measure the training result between train and validated corpus. As the Figure 3 shows, in order to measure the classification ability

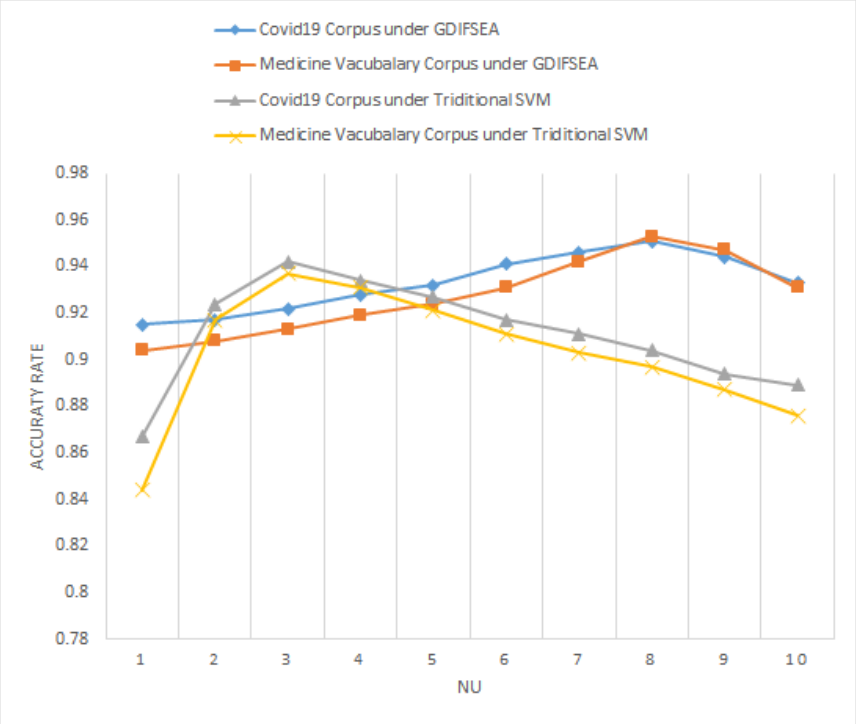

Figure 3: Comparison of accuracy rate and nu threshold under different classification algorithms

of GDIFSEA, we also introduce medical vocabulary corpus [6], compare with Covid19 corpus [3]. The acceptance rate for Covid19 Corpus is better than medical vocabulary corpus if Nu is below to 7. And the cooperation approaches is Support Vector Machine, in this comparison sets, we also use Covid19 datasets and medical vocabulary corpus, the result shows that when threshold is approach to 3, accuracy rate increase to its peak point. And all to expirements show that Covid19 tend better performance of classification accuracy , the reason I analyzed that the Covid-19 corpus has less type of medical words and much easier to training.In conclusion, when training under different threshold learning level(NU), the difference between two corpus are similar, and when NU increase to 7, the accuracy rate reach to maximum point, which means here is no overfit or underfit issue. Compare with other classification algorithm (SVM), the GDIFSEA can connect with different algorithm such as Recurrent Neural Network(RNN) or other deep learning algorithm, when increase the hardware computing ability, which means GDIFSEA is more flexible compare with SVM.

In next experiment, we compare the statistical characteristics. And the PSW parameter sets for characteristics are:

- total is 92055,
- minimize is 0.000037,
- q1 is 0.000264,
- q2 is 0.000435,
- q3 is 0.000663,
- maximize is 0.056947.

The distribution histogram of the covid-19 verification set PSW is as follows:

Since the Medical Vocabulary Transcription corpus[6] (short-named MV corpus)does not involve privacy-related content, the Medicine Vocabulary corpus (MV corpus) is used as a control sample set. We will check whether ppGDIFSEA model processes COVID19 Corpus appear the same as or similar to Medicine Vocabulary corpus. The former corpus contain patient privacy information and MV corpus without privacy information so the PSW of Control set is 0.000012, we can recognized the number as privacy deviation(mistakenly recognized some words as privacy).

The ppGDIFSEA trained by the covid-19 corpus is used to reconstruct Mean Square Error and calculate the PSW. The result are: total is 9799 minimize is 0.000091, q1 is 0.000386, q2 is 0.000681, q3 is 0.0010600 and maximize is 0.056947.

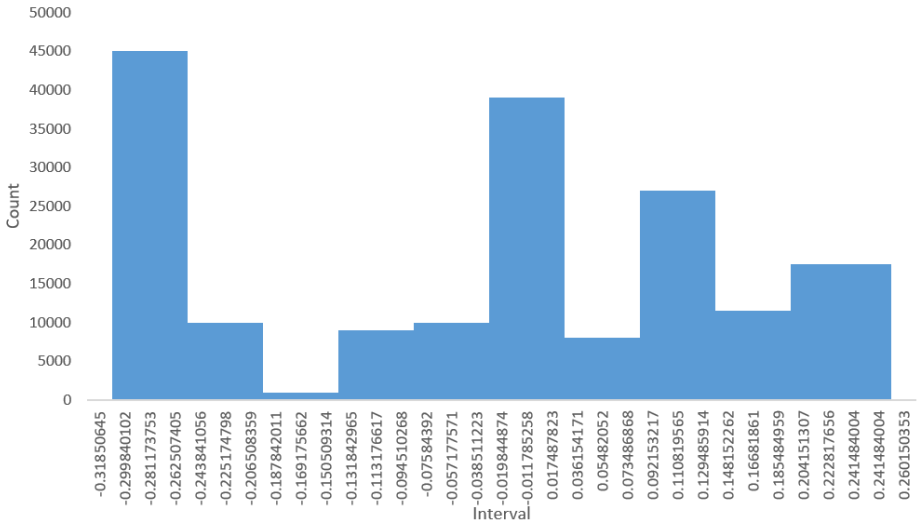

Figure 4: The PSW distribution on the covid-19 validation set for the ppGDIFSEA method

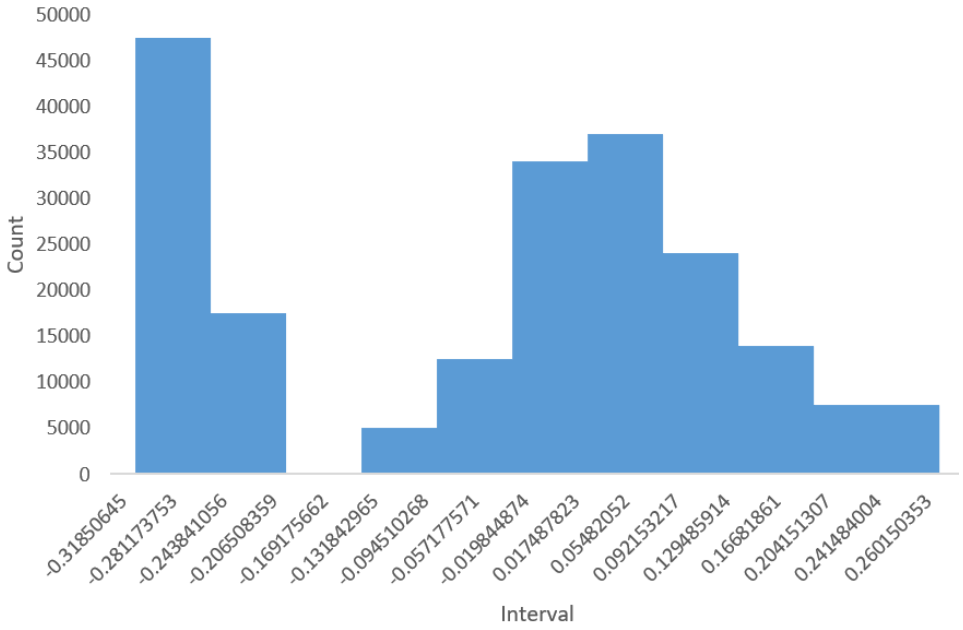

Figure 5: The PSW distribution of ppGDIFSEA method on the control set

The PSW distribution histogram of the control sample set(MV Corpus) is as follows:

In Figure 4, the privacy information in same parts only occupies the small amount of percentage within total words. As it shows the patient or authors' location information is the lowest histogram(3rd) in Figure 4. And patient or authors' name is the 4th histogram, but name and locations are classified separately , and both of group contain lots of fake information(as added noise). So even if the result is received, privacy information can hardly be linked together.

In Figure 5, we can see the lowest histogram(3rd) is locations and without any data inside. The 4th histogram as patient or authors' name decreases a lot compared with that in Figure 4, and the reason may be that MV corpus do not have any location information and most of name is from disease name or practitioner name or medicine name.

In general, the PSW experiment shows the ppGDIFSEA can help to protect privacy after processing target documents. Due to the fake or noise words, even find out privacy word, each of privacy words can hardly be linked together. So the confidential level is acceptable.

5 Conclusion and Future Works

In this work, we have proposed new algorithm named ppGDIFSEA for learning Covid-19 text representation (i.e., word embeddings) under Word2vec. We have empirically evaluated the algorithms on a realistic COVID-19 dataset and demonstrated that the proposed embedding model benefits from medicine data while the performance efficiency and accuracy for extra word with genetic algorithm has been compared with that for non-genetic algorithm approach. The novel ppGDIFSEA allows information classification with high accuracy when the quantity of samples is large enough. We have trained COVID19 dataset [5] with ppGDIFSEA and other popular approaches such as CNN, Auto Encoder(AE). Our trained model has been applied for data privacy-preserved with performance-guarantee. As the very fast and confidentially working on publicly embedding models, this paper highlights the new direction of privacy-preserving classification embedding models on sensitive text data. Many relevant future works remain. For example, one promising direction would be to explore strategies to detect what are exactly Coronavirus related contents to certain users (e.g., building a knowledge base) to apply our proposed high-performance and time-saving guarantee.

Acknowledgment

Thanks for Professor Edmund Lai, who provides the professional advices to this paper, supports and sponsors our work. Professor Edmund Lai is the Professor in School of Engineering, Computer and Mathematical Sciences, at the Auckland University of Technology, Auckland, New Zealand.

References

[1] K. A. Al Mamun, M. Alhussein, K. Sailunaz, and M. S. Islam, "Cloud based framework for parkinson's disease diagnosis and monitoring system for remote healthcare applications," *Future Generation Computer Systems*, vol. 66, pp. 36–47, 2017.

[2] P. Gantar, L. Colman, C. Parra Escartín, and H. Martínez Alonso, "Multiword expressions: between lexicography and nlp," *International Journal of Lexicography*, vol. 32, no. 2, pp. 138–162, 2019.

[3] kaggle, "Covid-19 open research dataset challenge from www.kaggle.com," kaggle, 2020.

[4] O. Krylova and D. J. Earn, "Effects of the infectious period distribution on predicted transitions in childhood disease dynamics," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130098, 2013.

[5] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday *et al.*, "Early dynamics of transmission and control of covid-19: a mathematical modelling study," *The Lancet Infectious Diseases*, 2020.

[6] mtsamples, "Medical transcription data scraped from mtsamples.com," mtsamples, 2019.

[7] I. J. B. Young, S. Luz, and N. Lone, "A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis," *International journal of medical informatics*, p. 103971, 2019.

AUTHOR BIOGRAPHIES

Bo (Bob) Ma is working toward the PhD degree in the Department of Information Technology and Software Engineering, School of Engineering, Computer and Mathematical Sciences at the Auckland University of Technology, Auckland, New Zealand. His research interest is Cybersecurity for various networks and systems in general, specially focus on privacy-preserving federated machine learning methods, privacy issues of Big Data analytics and Internet of Things, and information theory of deep learning. He obtained his MSc in Computer Software and Theory, and Bachelor in Electronic Commerce from Sichuan Normal University, China.

Jinsong Wu received PhD from Department of Electrical and Computer Engineering at Queen's University, Canada. He has been elected Vice-Chair, Technical Activities, IEEE Environmental Engineering Initiative, a pan-IEEE effort under IEEE Technical Activities Board (TAB), since 2017. He was the Founder (2011) and Founding Chair (2011-2017) of IEEE Technical Committee on Green Communications and Computing (TCGCC). He is also the co-founder (2014) and founding Vice-Chair (2015-present) of IEEE Technical Committee on Big Data (TCBD). His received both 2017 and 2019 IEEE System Journal Best Paper Awards. His co-authored paper won 2018 IEEE TCGCC Best Magazine Paper Award. He received IEEE Green Communications and Computing Technical Committee 2017 Excellent Services Award for Excellent Technical Leadership and Services in the Green Communications and Computing Community. He was the leading Editor and co-author of the comprehensive book, entitled "Green Communications: Theoretical Fundamentals, Algorithms, and Applications", published by CRC Press in September 2012. He has been IEEE Senior Member since 2011.

Shuang (Sue) Song is working toward the Master's degree in the Department of Information Technology and Software Engineering, School of Engineering, Computer and Mathematical Sciences, at the Auckland University of Technology, Auckland, New Zealand. Her research interests include sustainable computing, communications and networking, cloud and edge computing, greening Big Data analysis and Internet of Things.

William Liu is currently a Senior Lecturer in the Department of Information Technology and Software Engineering, School of Engineering, Computer and Mathematical Sciences at the Auckland University of Technology, New Zealand. He holds a Masters degree and a PhD degree in Electrical and Computer Engineering, both obtained at the University of Canterbury, New Zealand. He had been working as a network planner and designer in Beijing Telecom for 5 years before he immigrated to New Zealand. He has co-authored more than 90 papers published in international journals and conferences, and he participates in the Program Committees of several premier IEEE conferences including GLOBECOM, INFOCOM, ICC, GreenCom, CloudNet, DRCN, RNDM and ATNAC. His main research interests are in the design and performance evaluation of the infrastructure and protocols for packet-oriented networks. He is working especially in the areas of network survivability, sustainability and security, sustainable and trustworthy computing.