IJMLNCE
JOURNAL

International Journal of Machine Learning and Networked Collaborative Engineering

Journal Homepage: http://www.mlnce.net/home/index.html

DOI : https://doi.org/10.30991/IJMLNCE.2020v04i04.001

Assessment of Recruitment Records using Machine LearningApproach

Bui Thanh Hung

Faculty of Information Technology

Ton Duc Thang University

19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam

buithanhhung@tdtu.edu.vn

Abstract

In the era of the fourth industrial revolution (Industry 4.0), the applications of Information Technology (IT) have been widely used in various aspects of life. As the result, analysing and predicting the result for the application of candidates as well as employers are also growing significantly. Jobseekers and employers want to have accurate information and prediction results in order to have suitable job proposals for themselves and candidates. The primary approach to solve this problem is based on machine learning. This research uses Machine Learning and Deep Learning approaches in the recruitment evaluation process. We propose to use 3 machine learning methods – Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF) and Deep Learning- Recurrent Neural Network (RNN) to predict job applications. The dataset is collected from the Job Center of Binh Duong province. On the basis of the best results method, we integrate it in a job application.

Keywords

Assessment of Recruitment Records,

Machine Learning,

SVM,

Decision Tree,

Random Forest,

Recurrent Neural Network

1. Introduction

With the significant development of Information Technology (IT), it has had its applications in various aspects of life. As a benefit from that, people nowadays have more tools and methods to process and capture more information. Information technology becomes an inevitable trend when it is applied in all industries, all fields of production, business, and tourism as well. The result of applying the IT in management is the establishment of management information systems to serve the needs of processing data and providing information to those system owners.

Recruitment is undeniably one of the important functions of the organization - not only small-sized but also medium-sized enterprises (SME) and global corporations. That said, the recruiting process also holds as much critical role as its function. Recruiting occurs to fill out the gap of personnel to meet the day-to-day business activity as well as organization’s operation. Recruitment process is typically a multiple-step process, in which candidate evaluation and selection shall be deemed as the crucial step.

Evaluation and selection of promising candidates is the activity of reviewing and comparing all the candidate applications following specific hiring criteria or standards to determine the mostly suited one. It has never been easy to evaluate candidates, especially when there might be a huge number of applications. Therefore, prior to conducting the assessment of candidates, it is essential for the employer to identify the process as well as hiring criteria, along with an appropriate assessment method to identify the finalist.

In practice, it is popular known of two methods to evaluate and assess candidates: rating and scoring

- Rating method: The candidate will be rated based on the hiring criteria. However, the challenges of this method are to define clearly the priority as well as the importance of each criteria in the overall assessment. The rating process could not be proceeded until all candidates have been rated. It could be challenging to remember every single details of each candidate when there are too many applications.

- Scoring method: each candidate will be assessed and scored based on the hiring criteria. It needs to be specified clearly the correlative scores for each criterion. Frankly speaking, there are multiple ratings and evaluating methods as well; however, it is not yet seen a perfect method. There will be circumstances that the members of the interview panel only score the candidates following their personal preferences or gut feelings. That said, it appears to be challenging to obtain a neutral scoring assessment. Therefore, in practice, the employers, to some certain extent, need to be flexible when assessing and evaluating the candidates or selecting the appropriate recruitment method.

On the other hand, from candidate perspective, it would be beneficial if they could be recommended on what type of existing jobs or positions might be a good match for them. In order to do that, it is also a complex process that requires multiple steps. The number of steps might be varied depending on the candidate profile and the job requirements as well as availability. Therefore, it is necessary to have an application that evaluates the candidate profile automatically to propose the appropriate position for the person.

It could be said that there are multiple approaches to this matter. However, the primary approaches are the manual selection of the recruiter which is still human work. Thus, it has caught the attention of many researchers on how to use machine learning and apply the natural language to solve the problem [1-6]. Nikolaos et al. [1] used machine learning and text processing to build a recommender system. Data Vishnu et al. [2] used estimated emotion to rank a candidate. Combining TF-IDF feature and machine learning for personality classification is proposed by Manasi et al. [3]. Vivian et al. [4] built the system to evaluate resume based on machine learning approach. In [5-6], they proposed machine learning and TF-IDF features to analysis CV and rank a candidate. Based on previous researches, we propose machine learning approach to solve the above recruitment challenge.

This paper consists of 4 sections, Section 1 introduces about the problem. Section 2 describes the proposed model. Section 3 presents the experiment settings and discusses the results of the experiments. And Section 5 concludes our work and future work

2. Proposed Method

The overall model is given below in Figure 1 with 2 main parts: The training model by the application (on website) for users (candidates) and employers.

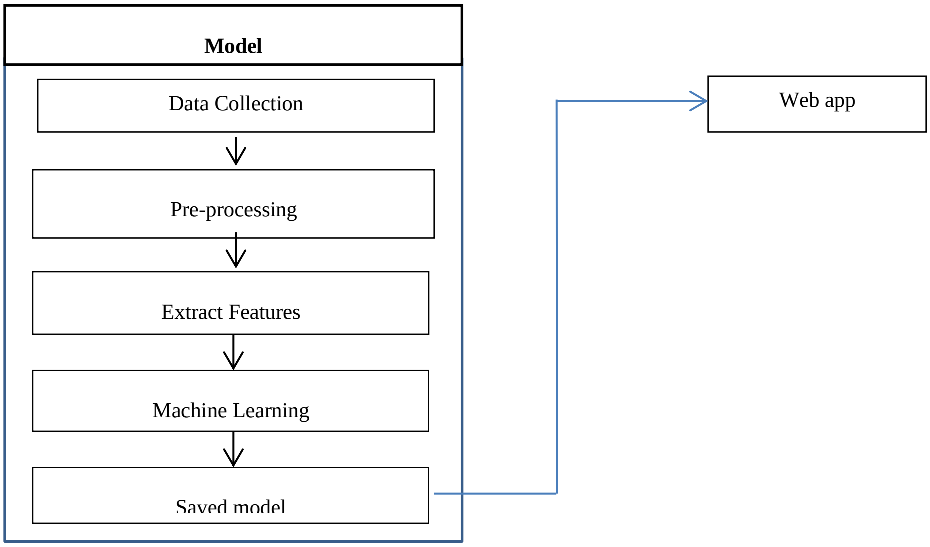

Figure 1: The proposed method

After collecting data from the job site in Binh Duong province and labeling the job categories for candidates, features will be represented numerically and built into machine learning models based on machine learning algorithms like the Recurrent Neural Network (RNN) the Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT). The processed data will be converted into digital vectors and put into machine learning models for training to serve the prediction process. The application is built on pre-trained machine learning models. The prediction process is based on the user's input and returns the outcomes on the user interface.There are numerous inspirations for utilizing highlights as opposed to the pixels straightforwardly. The element based framework works for identifying facial milestones from nonpartisan and posture variety pictures was planned. Prior to figuring the likeness between faces, the face pictures should be adjusted. To do this, first create the milestone purposes of the eyes, mouth and nose. The three milestone focuses are produced for each picture accessible in the preparation dataset. The fourth milestone point is known as stero. It makes a 3*3 channel over the picture and concentrates the facial highlights and computes the separation between the test picture and preparing pictures.

2.1 Features

Training data includes 2 types: numeric data and string data.

- The numerical data consists of: Age, Gender, Year of Experience.
- String data includes: Education, Previous job, Foreign Language, Computing.

Each sort of data has been pre-processed, extracted different features to convert into numeric data and put into a training model. Besides, the processing and converting data into featured-vectors are carried out as follows:

Numeric data: these values have different values that affect the efficiency of many algorithms that are relevant to the problems such as: implementation time, convergence process, and the accuracy of the algorithm. Therefore, we need a further step to normalize data into standardized-data. In this research, we use the following formula to standardize data into the form of [0,1]:

$$z_i = \frac{x_i - min\ (x)}{max(x) - min\ (x)} \qquad (1)$$

Text data: Before converting, we convert text data into vectors, we preprocess it with further steps:

Step 1: Removing commas, dots, spaces.

Step 2: Extracting Vietnamese words using Pyvi library

Step 3: Convert all words into lowercase.

After preprocessing, we transform text data into a vector using TF-IDF (Term Frequency - Inverse Document Frequency) method [7-8].

TF-IDF is a technique used in text mining to evaluate the importance of a word in the text. A high value indicates the significance of the word and it depends on the number of times that the word appears in the overall document and is offset by how often the term presents in the data set. The formula TF-IDF is presented as follows:

TF (Term Frequency): the number of times a word/ term occurs in a document.

$$TF(t,d) = \frac{F(t,d)}{max(\{F(w,d): w \in d\})} \qquad (2)$$

where,

TF(t, d): the frequency of the word t in the document d

F(t, d): the number of occurrences of word t in d

max({F(w, d) : w ∈ d}): the number of occurrences of word with the most occurrences in the document d

IDF: aims to define the importance of a word/term within a document. In TF calculation, the importance of all words is considered equal.

$$IDF(t, D) = log \frac{|D|}{|\{d\in D:t \in d\}|} \qquad (3)$$

where,

IDF(t, D): the idf value of the word t in a text corpus

|D|: the total number of texts in corpus D

{d ∈ D : t ∈ d}|: number of document in D containing word t

The formula of TF-IDF based on TF and IDF is calculated as below:

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D) \qquad (4)$$

We use the TF-IDF technique to represent the textual data columns in the candidate information. All information in each column will be collected to create a dictionary of words at that column. The data of each candidate are represented by vectors based on the dictionary, then the TF-IDF formula is computed per vector and gives a vector which represents each information of the candidate.

2.2 Training

In this research, we use Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and Recurrent Neural Network (RNN) to train the model. The methods are detailed below:

SVM

Support Vector Machine (SVM) method was developed from the theory of statistics of Vapnik and Chervonenkis in 1995 [9-10], and has a lot of potential for developing it in theory as well as in practice. The SVM method has the ability to classify the classification problem as well as in many practical applications. Support Vector Machines (SVM) method is a new technique for data classification, which is a learning method using the hypothetical space of linear functions over multi-dimensional feature space, based in optimization theory and statistical theory. In the SVM technique, the initial input data space will be mapped into the feature space. In this high dimensional space, the optimal separating hyperplane will be determined. Figure 2 shows the separating hyperplane (w, b) in 2 dimensional space.

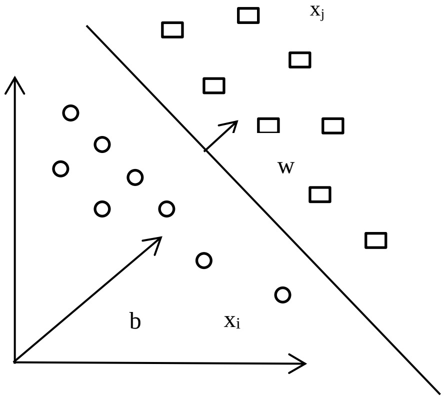

Figure 2: Divided according to hyperplane (w, b) in 2-dimensional space

Decision Tree

Decision Tree (DT) is a model, expressly mapping from observations of a data to make conclusions about the target value of thing [11-12]. Each internal node corresponds to a variable; the line between it and the child nodes represents a specific value for that variable. Each leaf node represents the predicted value of the target variable, given the values of the variables represented by the path from the root node to that leaf node. Figure 3 shows Decision Tree model.

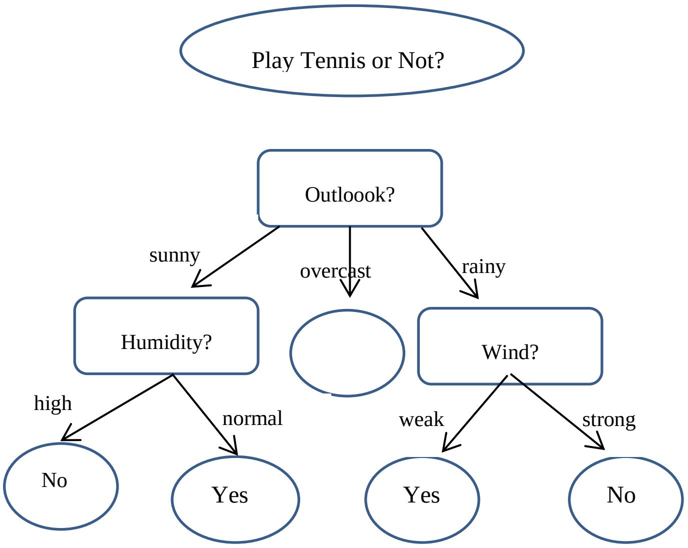

Figure 3: Decision Tree Model

The Decision Tree is a supervised machine learning model that can be applied to both classification and regression problems. Building a decision tree on top of a given training data is the identification of questions and their order. One notable feature of Decision Tree is that it can work with categorical features, often discrete and out of order.

Random Forest

Random Forests (RF) is a supervised learning algorithm. “Ensemble” means gathering all the "weak learners" and helping it work together to produce a highly reliable prediction [13]. In this case, the "weak learners" are all Decision Trees randomly combined to form highly reliable predictions - Random Forest is one of the most popular machine learning algorithms. Figure 4 illustrates the random forest model.

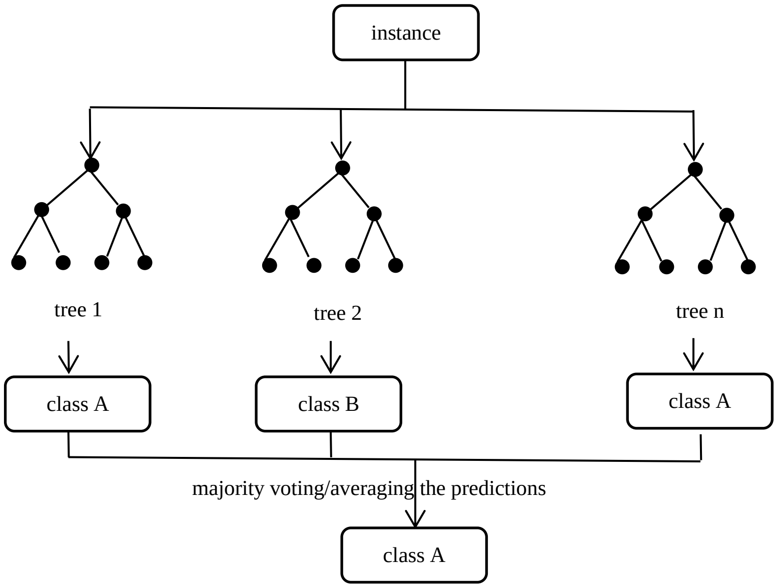

Figure 4: Random Forest Model

Recurrent Neural Network

Recurrent neural networks (RNNs) [14-18], are a class of neural networks. This model uses previous outputs as inputs while having hidden states. RNNs have many advantages such as:

- Any length of input can be processed effectively
- The size of input does not affect to the model size
- Account historical information contains the computation
- By the time, all weights are shared

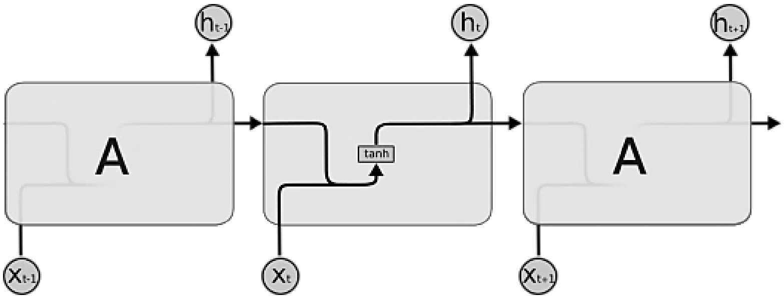

Figure 5: A diagram of a simple RNN cell

Fig. 5 shows a diagram of a simple RNN cell. The cell unit represents the memory. Individual cells are combined together to form a large network thereby befitting the term deep neural networks.

3. Experiments

3.1 Dataset

The dataset is collected directly from Binh Duong career site of Binh Duong Job Center. The raw data set included 1967 profile samples that the user uploaded on the website. The data set is saved in an Excel format, including 13 columns of different information such as: Name, Date of Birth, Gender, ID Number, Phone Number, Address, Registration Time, Job Position, Workplace, Years of Experience, Education Level, Major/Industry, Foreign Language and Computing.

3.2 Processing data

We implement the data pre-processing process by removing personal information such as: Name, ID Number, Phone Number, Address, Registration Time and other information in the remaining columns. The Date of Birth column will transform into the current age. However, there is still a lot of repetitive information among the candidates in this data set. In order to avoid noise in the data making process, we remove the duplicates and retain only one based on the ID Number column. Since this is a defined value (key-value) between the candidates, the results after filtering the overlap in the data set, we obtained 1516 samples labeled data. After eliminating the irrelevant information, we assign the appropriate job category to each profile sample. The job description table is referenced directly from the website of the Job Center of Binh Duong province. Next, we proceed to label data that candidates entered into the data set. Based on the columns of information that the candidates enter: Occupation, Job Position that they apply; it is classified into the above occupations. There are also some spelling errors, incorrect syntax, and invalid information being preprocessed to match the column information during the labeling process. Details of the data column information that candidates provide jobs include: Age, Gender, Years of Experience, Education, Foreign Language and Computing. Data is divided into 2 parts Train and Test in proportional of 8 and 2.

3.3 Model Training

After preprocessing the raw data, we use Min-Max scaling technique to transform the numeric data into vector form [0,1]. For textual data, we use TF-IDF technique to convert text values to representational vectors. Then join the vectors of these columns together to make the vector representing each data row, the job label is also converted to the corresponding numeric form, the operator puts these two values into the algorithmic machine learning model. SVM, RF, DT, RNN to train and evaluate the model.

3.4 Experiment Results

The datasets were converted into vectors and put into a training model using four machine learning methods: Support Vector Machine (SVM), Random Forest, Decision Tree and RNN. We use the Python programming language, the pyvi library Tran Viet Trung (2016) for tokenize words, the Sklearn, keras machine learning library with Numpy and Scipy. In addition, the application interface is designed by HTML, Javascript, CSS and Bootstrap. The results are evaluated on three measurements: accuracy, coverage and F1 score. These measurements are calculated by using the formulas below.

$$\text{Precision} = TP/(TP + FP) \qquad (5)$$

Precision is the ability of the categorical algorithm to not assign negative values to the positive pattern. For each class, it is defined as the ratio of True Positive to the sum of the True Positive and the False Positive.

$$\text{Recall} = TP/(TP+FN) \qquad (6)$$

Recall is the ability of a classification algorithm to find the positive patterns. For each class, it is defined as the True Positive ratio to the True Positive to False Negative ratio.

And, F1 score is the neutralization of the Precision and Recall values.

$$F1 - score \;= 2 * \frac{Precision*Recall}{Precision+Recall} \qquad (7)$$

The results are shown in Table 1. Fig. 6 shows the comparison of results of the 4 methods SVM, RF, DT and RNN.

Table 1: Results of 4 methods: SVM, RF, DT and RNN.

Method	Precision	Recall	F1-score
SVM	**73.64**	70.29	**71.92**
RF	69.87	48.51	57.27
DT	64.21	62.05	63.11
RNN	71.58	**70.72**	71.15

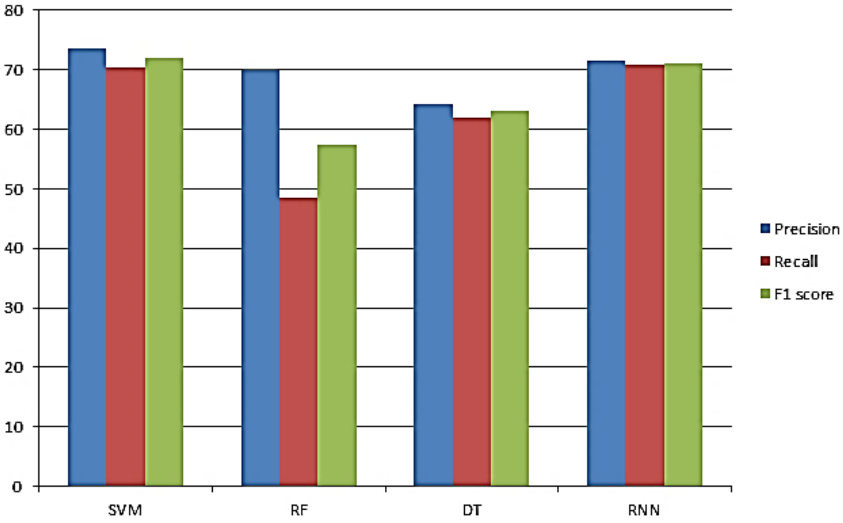

Figure 6: Comparison of results of the 4 methods SVM, RF, DT and RNN.

According to the results in Table 1, we see that the Support Vector Machine (SVM) method got the best performances with a precision of 73.64%, recall of 79.29%, and F1-score of 71.92% respectively. The results are much higher than the other three methods, Random Forest, Decision Tree and RNNs. RNNs got the second place. Because the dataset is small, that's why RNNs didn't work well. The SVM model worked well on the dataset by the effective algorithm. By the results, the SVM method will be saved for building Web application.

On the Web application, users will directly input candidates' information: Age, Degree, Years of Experience, Foreign Language, Computing and Gender. The application will take this information pre-processed and pass through the SVM model to predict and give the most appropriate occupations. Four other suitable jobs are arranged in order from high to low suggestion.

4. Conclusion

This study presents a method of assessing applications using machine learning and deep learning approach. Basing on standardized input dataset which are converted to TF-IDF feature vectors and trained by four machine learning models: SVM, Decision Tree, Random Forest and RNN. The experiment shows that the SVM machine learning method gives the best results in the F-measure metric. We have also built an online application to assess the recruitment records and initially surveyed users' feedback. In the future, we will research and process the collected data and test on other models to find the best solution for the recruitment assessment.

References

[1] FoDRA – Nikolaos D. Almalis George A. Tsihrintzis, Aggeliki D Strati: A New Content-Based Job Recommendation Algorithm for Job Seeking and Recruiting. 6th International Conference on Information, Intelligence, Systems and Applications (IISA), 2016.

[2] Data,Vishnu M Menon Computer Rahul Nath H A.: A Novel Approach to Evaluate and Rank Candidates in a Recruitment Process by Estimating Emotional Intelligence through Social Media. 2016 International Conference on Next Generation Intelligent Systems (ICNGIS)

[3] Manasi Ombhase, Prajakta Gogate, Tejas Patil: Automated Personality Classification Using Data Mining Techniques. 10.13140/RG.2.2.35949.59363. 2017

[4] Vivian Lai, Kyong Jin Shim, Richard J. Oentaryo, Philips K. Prasetyo, Casey Vu Ee-Peng Lim, David Lo: Career Mapper: An Automated Resume Evaluation Tool. arXiv:1611.05339, 2016.

[5] Jayashree Rout, Sudhir Bagade, Pooja Yede, Nirmiti Patil: Personality Evaluation and CV Analysis using Machine Learning Algorithm. International Journal of Computer Sciences and Engineering. Vol.-7, Issue-5, May 2019.

[6] Agnes van Belle, Eike Dehling, and Daniel Foster: Improving Candidate to Job Matching with Machine Learning. 2018

[7] Stephen Robertson: Understanding Inverse Document Frequency: on Theoretical Arguments for IDF. Journal of Documentation, Vol. 60 Issue: 5, pp.503-520, 2004.

[8] Shahzad Qaiser,Ramsha Ali. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications (0975 – 8887), volume 181 –No.1, July 2018.

[9] Tom M. Mitchell: Machine Learning. McGraw Hill, Inc. 1997.

[10] Jiawei Han, Micheline Kamber: Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, 2006.

[11] Leo Breiman, Jerome Friedman, Charles J. Stone & R.A. Olshen: Classification and Regression Trees – Taylor & Francis, 1984.

[12] Mihaela van der Schaar: Classification and Regression Trees. Department of Engineering Science University of Oxford. 2017

[13] Breiman, L.: Random Forests. Machine Learning vol.45(1):5-32, 2001.

[14] Samuel: A Thorough Review on the Current Advance of Neural Network Structures. Annual Reviews in Control. 14: 200–230, 2019.

[15] Bui Thanh Hung: Vietnamese Question Classification based on Deep Learning for Educational Support System. The 19th International Symposium on Communications and Information Technologies, ISCIT 9.2019

[16] Bui Thanh Hung: Document Classification by Using Hybrid Deep Learning Approach. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering - LNICST, volume 298, pp 167-177, Springer, 2019.

[17] Bui Thanh Hung, Le Minh Tien: Facial Expression Recognition with CNN-LSTM. Research in Intelligent and Computing in Engineering. Springer Series in Advances in Intelligent Systems and Computing, 2021.

[18] Bui Thanh Hung: Combining Syntax Features and Word Embeddings in Bidirectional LSTM for Vietnamese Named Entity Recognition. Further Advances in Internet of Things in Biomedical and Cyber Physical Systems, 2021.

Author’s Biography

Dr. Bui Thanh Hung received his M.S. degree and Ph.D. degree from Japan Advanced Institute of Science and Technology (JAIST) in 2010 and in 2013. He is currently a lecturer of Faculty of Information Technology, Ton Duc Thang University. He has completed 2 projects, published 10 journals, 11 book chapters, 32 International conference papers and 15 domestic conference papers. He had three best paper awards of FICTA 2018, RICE 2020 and ICAIAA 2021 and is selected as Excellent Scientific Researcher of Thu Dau Mot University in 2019, 2020, 2021. He is a reviewer of many reputed journals and conferences. His main research interests are Natural Language Processing, Machine Learning, Machine Translation, Text Processing, Data Analytics, Computer Vision and Artificial Intelligence.