IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 8, No. 3, September 2019, pp. 299~306
ISSN: 2252-8938, DOI: 10.11591/ijai.v8.i3.pp299-306

Performance analysis of supervised learning models for product title classification

Norsyela Muhammad Noor Mathivanan[1], Nor Azura Md.Ghani[2], Roziah Mohd Janor[3]
[1,2,3]Center for Statistical and Decision Sciences Studies, Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]National Design Centre Universiti Teknologi MARA 40450 Shah Alam, Selangor, Malaysia

Article Info

Article history:

Received Jun 1, 2019
Revised Aug 12, 2019
Accepted Aug 29, 2019

Keywords:

Product title
Supervised learning model
Text classification

ABSTRACT

Online business development through e-commerce platforms is a phenomenon which change the world of promoting and selling products in this 21st century. Product title classification is an important task in assisting retailers and sellers to list a product in a suitable category. Product title classification is a part of text classification problem but the properties of product title are different from general document. This study aims to evaluate the performance of five different supervised learning models on data sets consist of e-commerce product titles with a very short description and they are incomplete sentences. The supervised learning models involve in the study are Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM) and Random Forest. The results show KNN model is the best model with the highest accuracy and fastest computation time to classify the data used in the study. Hence, KNN model is a good approach in classifying e-commerce products.

Copyright © 2019 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Nor Azura Md. Ghani,
Center for Statistical and Decision Sciences Studies,
Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor Malaysia.
Email: azura@tmsk.uitm.edu.my

1. INTRODUCTION

Currently, e-commerce platforms become a major way to promote products and services. It consists a wide range of interaction processes between various market users from ordering and delivering products until the sellers issue invoices and users make payments [1]. The fast development of e-commerce is contributed the most by the availability of different products and the easiness of transacting money over internet [2]. There are large inventories with various kind of products sold through online store websites such as Walmart, Amazon, Alipay and eBay. Users are able to view and buy many new products from the websites from time to time. Most of the websites are well-structured and they consist of product information such as the product name, description, price, and image. The often inclusion of new products in e-commerce websites leads to an important task of classifying a product title to assist sellers listing an item in a suitable category. Some system will perform classification directly after obtaining title of the product. Most of the text mining process involves a bundle of documents consist of lengthy words but title description normally is a short text. Titles are different from texts in various aspects such as the length of each sentence is very short, it consists of similar distribution of lengths, and the grammatical structure is mostly incompleted [3]. There is a study related to product title classification, but it focused on identifying general properties of product titles [3-4]. The most related but broader area is a short-text classification, which already consists of various literature. There are studies involve classifying short text such as question classification [5-6],

semantic annotation classification [7] and job title classification [8]. However, question and semantic annotation classification normally uses complete sentences and job title classification deals with shorter sentences compared to product title.

Besides that, there are several studies take into consideration other sources such as price of the product to be the additional feature with the aim to increase the accuracy rate of the classification model [9-10]. However, this study solely consider text from title of products. Previously, it was found that the properties of product title classification are different from those of text classification where stemming and stop-word removal are not needed in dealing with product titles [3]. This study is in the same direction, sharing the same spirit of keeping model simplicity and interpretability. The difference is the previous studies focused on properties of product titles, which involve transforming each pair of word tokens from the text sequences to useful features. In this paper, the authors main purpose is identifying the most suitable classification models to deal with short-text data especially involving e-commerce product titles. The classification of e-commerce products based on title of a product can be done using supervised learning model. A supervised learning model is able to solve classification problems because the goal is to make the computer learns a classification system that has been created. There are various kind of supervised learning models have been applied in many fields of studies such as pattern recognition [11], natural language processing [12], market segmentation [13] and bioinformatics [14]. Nevertheless, the comparison between well-known supervised learning models including Naïve Bayes, K-Nearest Neighbor (KNN), Decision Trees, Support Vector Machine (SVM) and Random Forest is not yet to be seen in a research related to product title classification. It is crucial to evaluate the performance of each model because the results provide useful information about the best model to classify this kind of data. Hence, this paper aims to compare the performance of supervised learning models which are Naïve Bayes, KNN, Decision Trees, SVM and Random Forest in classifying title of e-commerce products. The rest of this paper is organized as follows: Section 2 describes researches related to product title classification; Section 3 presents the research methodology includes description of data sets and research design used in the study; Section 4 shows the results obtain from the comparison of the supervised learning models towards e-commerce product classification; Section 5 concludes the findings from the research.

2. RESEARCH METHOD

2.1. Data description

Department of Statistics Malaysia (DOSM) has collected product information from one of the major online store website through STATSBDA project known as Price Intelligence (PI) using its prototype web scraper. A few leaf nodes were used to represent the chosen categories from the browse tree of the website. Table 1 presents the description of the two corpora selected for this research which are fresh food and household products data sets. The five categories under Fresh Food data set are fresh meat & poultry, fish & seafood, bakery, fresh fruits, and noodles. On the other hand, the five categories under Household data set are toilet cleaner, air freshener, floor cleaners, light bulbs, and household sundries.

Table 1. Summary description of data sets

Dataset	Category	Instance	Number of Feature	Number of Feature after Feature Selection
Fresh Food	5	447	88	78
Household	5	684	138	116

2.2. Research design

There were several steps need to be done in this research before classifying the data as presented in Figure 1. The steps were data extraction, data pre-processing, feature extraction and feature selection. These were the basic procedures in research related to text classification. There were three preprocessing steps involved after data extraction which were tokenization, stop word removal and stemming [15]. The data preprocessing is a crucial step to ensure the data is standardized and in a proper form. The standardized form was achieved after applying the three preprocessing steps where product descriptions were tokenized into words at first. Then, stop words were removed from the word list and the remaining words were stemmed to ensure the words followed the root word forms. The feature extraction and selection are important to make sure the data are well transformed into significance and good features before performing the classification process [16]. The selection of features may affect the accuracy of a classification model. Hence, the research had utilized the bag-of-word and correlation feature selection

technique to perform data extraction and selection respectively. Then, the chosen features were used as inputs to perform different classification models from supervised learning models.

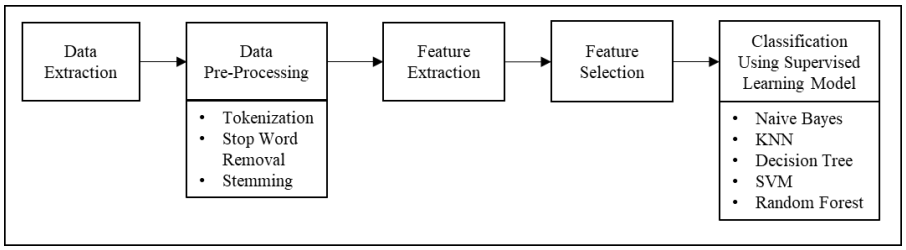

Figure 1. Flowchart of the Research

Supervised learning model is used to make predictions based on information about the targets and the features of data. It infers a function according to a given set of input-output data respectively. Normally, the input data provides a set of observations with which the computer is trained [17]. Each observation consists of an input vector and a desired output value. A supervised learning model trains the data and generates a general rule or function to be used for predicting or classifying new inputs. There were five supervised learning models that been used in the study which are Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM) and Random Forest. Naïve Bayes is a classification model based on Bayes theorem introduced by Thomas Bayes and it has been used as conventional paradigm since late 18th century [18]. It is one of probabilistic-based classifiers where it predicts the probability of the sample itself before choosing the class with highest probability given the observation. It is widely used in text categorization, sentiment analysis and spam filtering [15]. The algorithm for Naïve Bayes [19] as shown in Figure 2.

Dataset $D = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$

1st step:
Assign all the words in D as the feature vector, V

2nd step:
Train the model by,
Repeat
for each category $c_i \in C$,
let d_i be the subset of samples in D with category c_i
$P(c_i) = |d_i|/|D|$
let t_i be the concatenation of all the samples in d_i and n_i be the total number of occurrences in t_i
for each feature $x_j \epsilon V$
let n_{ij} be the number of occurrences of w_j in t_i
let $P(x_j|c_i) = (n_{ij} + 1)/(n_i + |V|)$

3rd step:
Given a test sample of X and let the number of feature occurrence in X,
Find the category by,

$$\arg\max_{c_i \in C} P(c_i) \prod_{i=1}^{n} P(a_i|c_i)$$

Where a_i is the feature occurring at the ith position in X

Figur 2. Naïve Bayes Algorithm

KNN is the fundamental classification model to classify observations according to the closest training examples in the feature space when there is little or no prior knowledge on the distribution of the data [20]. It is an instance-based learning where the function is close to local value and the computations are deferred before the classifying process occurs. Basically, the rule holds the training set during the learning process. Then, each observation is assigned to a class according to the majority label of its KNN in the training dataset. A sample should be grouped into its similar surrounding samples. Thus, the nearest neighbor

samples can be considered to classify or predict an unknown sample. The algorithm for KNN [21] as shown in Figure 3.

Dataset $D = \{x_1, \dots, x_n\}$ with training categories $c_1, \dots, c_j$ and sum of training sample is N

1st step:
Repeat
for each input vector $(x_1, \dots, x_n)$,
do
Calculate the similarities between training samples.
e.g. ith sample of $d_i(d_{i1}, \dots, d_{im})$ where the similarity SIM(x, d_i) can be represented by,

$$SIM(x, d_i) = \frac{\sum_{j=1}^{m} x_j . d_{ij}}{\sqrt{(\sum_{j=1}^{m} x_j)^2 \sqrt{\sum_{j=1}^{m} d_{ij}{}^2}}}$$

2nd step:
Choose k samples which are larger from N similarities of $SIM(x, d_i)$, $(i = 1,2, \dots, N)$ and serve them as KNN collection of x.

3rd step:
Calculate the probability of x belong to different categories accordingly using,

$$P(x, c_j) = \sum_{d} SIM(x, d_i).y(d_i.c_j)$$

Where $y(d_i.c_j)$ defined by category attribute function that satisfied,

$$y(d_i.c_j) = \begin{cases} 1, d_i \in c_j \\ 0, d_i \notin c_j \end{cases}$$

4th step:
Find the category for the sample that has the largest $P(x, c_j)$

Figure 3. K-Nearest Neighbor (KNN) Algorithm

Decision Tree is a model with flowchart-like structure. It is created by a tree and a set of rules representing each of the classes from a dataset. Decision Tree consists of three main elements which are internal node, branch and class label where each of them represents a test attribute, a test outcome and a leaf node respectively [22]. Figure 4 shows the algorithm for Decision Tree [15].

D = Training Dataset, M = Input Attributes and N = Target Attributes

1st step:
Create a tree, $T = TreeGrowing(D, M, N)$
If
one of the stopping criteria is fulfilled then mark the root node in T as a leaf with the most common value of N in D as the class.
else
Find a discrete function $f(M)$ of the input attributes values such that splitting D based on outcome of $f(M)$ that gains the best splitting metric.
If the best splitting metric ≥ threshold then label the root node of T as $f(M)$
For each outcome v_i of $f(M)$ do
$subtree_i = TreeGrowing\ (\sigma_{f(M)=v_i} D, M, N)$
Connect the root node of T to $subtree_i$ with an edge that is labelled v_i
end for
else
Mark the root node in T as a leaf with the most common value of M in D as the class.
end if
end if
The return output is the value of T

2nd step:
Prune the tree, $T = TreePruning(D, T, N)$
Select a node t in T and prune it maximally improve some evaluation criteria
If $t \neq \phi$ then
$T = pruned(T, t)$
end if
Repeat until $t = \phi$
The return output is the value of the current T

Figure 4. Decision Tree Algorithm

Support Vector Machine (SVM) is usually used for classification and was introduced [23]. It works based on the calculation of margins between the classes. The margins are drawn to minimize the classification error when the distance between the margin and the classes is a maximum. SVM had been applied into various fields of studies such as gene expression, text classification and image identification [24]. This model is considered to give good generalization accuracy, but it may cause a quadratic optimization problem with bound constraints and a lack of linear equality in the training process. The algorithm for SVM [25] as shown in Figure 5.

Dataset $D = \{(x_1 y_1, \ldots, x_n y_n)\}$ where x_i is a n-dimensional vector and y_i is denoted as class of 1 or -1 to each point belongs to x_i.

1st step:
Train the model using the function of

$$f(x) = v.x - b$$

Where v is the weight vector and b is the bias. The condition needs to be satisfied is,

$$y_i(v.x_i - b > 0, \forall\ (x_i, y_i) \in D$$

2nd step:
Maximize the margin which is the distance from the hyperplane to the closest data points. The distance is formulated as,

$$distance = \frac{|f(x_i)|}{||v||}$$

Hence, the margin can be written as,

$$margin = \frac{1}{||v||}$$

The training problem is presented by,

$$minimize: Q(v) = \frac{1}{2}||v||^2$$
$$subject\ to: y_i(v.x_i - b) \geq 1, \forall (x_i, y_i) \in D$$

Figure 5. Support Vector Machine (SVM) Algorithm

Random Forest is also known as the ensemble of decision tree algorithm. Figure 6 shows the algorithm for Random Forest [26]. It consists of a collection of tree-structured classifiers where each of the classifiers is an independent identically distributed random vector. This algorithm can maintain its performance even though the data consists of a large proportion of missing values [27]. All the steps were computed using R-Programming software.

N = Number of nodes
M = Number of features
D = Number of trees to be constructed

1st step:
Randomly draw a bootstrap sample A from the training data D

2nd step:
Construct tree T_i from the drawn bootstrapped sample A using,
i. Randomly select m features from M where $m < M$
ii. Calculate the best split point among the m features for node d
iii. Split the node into two daughter nodes using the best split
iv. Repeat i to iii until n number of nodes has been reached

The forest is built by repeating the steps i to iv for D number of times

3rd step:
Output all the constructed trees $\{T_i\}1D$ and apply a new sample to each of the constructed trees starting from the root node

4th step:
Assign the sample to the class according to the leaf node
Combine the decisions of all trees and Find the highest votes as the class for the sample

Figure 6. Random Forest Algorithm

3. RESULTS AND DISCUSSION

The evaluation was done by observing the classification results of five algorithms from supervised learning model. The analysis was made on two different datasets which were Freshfood and Household data sets as shown in Table 2. The highest accuracy for the data with five categories was performed by KNN model. On the other hand, the performance of Random Forest model was highly good as KNN model, but the performance of other classification models was less than 70%. Besides that, the result from Fresh Food data led to approximately similar conclusion as the result obtained from Household data. The highest accuracy rate to classify the data consisted of five categories is KNN model. Specifically, only KNN and Random Forest models showed good accuracy rates compared to other classification models. The Naïve Bayes model was the worst classifier among the five algorithms used in the study to classify both of the data. The execution time for each of the classification model is shown in Table 3. Despite the highest accuracy rates provided by KNN model for both data sets, it also the fastest classifier compared to other classification models used in the study. Even thought, the performance of Random Forest was fairly good compared to KNN but it was the slowest classifier to provide the classification results among the five algorithms.

Table 2. Accuracy of classification models

Method	Dataset	
	Household	Fresh Food
Naïve Bayes	16.99	14.07
KNN	94.66	82.96
Decision Tree	85.44	67.41
SVM	69.42	44.44
Random Forest	93.69	78.52

Table 3. Accuracy of classification models

Method	Dataset	
	Household	Fresh Food
Naïve Bayes	1.20	0.78
KNN	0.82	0.63
Decision Tree	0.94	0.66
SVM	0.99	0.65
Random Forest	1.51	1.17

From the results, it was obvious that KNN model outperformed other supervised learning models. The performance of Random Forest model was not far behind the KNN model but the weakneses of the model can be seen in term of computation time. This results inline with previous study by reference [28] wherein the performance of both models were preferable compared to other supervised learning models toward breast cancer data. Meanwhile, several studies had also found that KNN model is superior in classifying different kind of data [29-31]. Among the algorithms based on supervised learning models used in the study, Naïve Bayes performed not as good as the other algorithms. It is proved that the performance of Naïve Bayes model is affected by the distribution of the data [32]. Normally, it performed well on the real-world data where the nature of the data easily changes over the time. However, the data used in the study were independent and identically distributed data.

4. CONCLUSION

The paper presents comparative evaluation of five well-known algorithms from supervised learning model for the problem related to e-commerce product titles classification. On the whole, KNN model performed the best among the five supervised learning models. The simplicity of the model suits the requirement to classify a short text such as e-commerce product titles. The performance of KNN model can be enhanced by investigating the optimal number of neighbors (K) value.

ACKNOWLEDGEMENTS

The research is fully funded by the University Teknologi MARA and Ministry of Education Malaysia under the Grant Scheme (600-IRMI/FRGS 5/3 (120/ 2019)). The authors would like to express their deepest gratitude to the Department of Statistics Malaysia for providing knowledge and data supports.

REFERENCES

[1] D. Kim, S. Lee, and J. Chun, A semantic classification model for e-catalogs, in: Proceedings of th e IEEE Conference on E-Commerce, 2004.

[2] Sun, Chong, Narasimhan Rampalli, Frank Yang, and AnHai Doan. "Chimera: Large-scale cla ssification using machine learning, rules, and crowdsourcing. "Proceedings of the VLDB Endowment 7, no. 13 (2014).

[3] Hsiang-Fu Yu, Chia-Hua Ho, Prakash Arunachalam, Manas Somaiya, and Chih-Jen Lin. Product title classification versus text classification. Technical report, 2012.

[4] M. W. Libbrecht and W. S. Noble. "Machine learning applications in genetics and genomics". Nature Reviews Genetics.16:321–32. 2015.

[5] D. Zhang and W. S. Lee, "Question classification using support vector machines," in SIGIR, pp. 26–32, 2003.

[6] B. Qu, G. Cong, C. Li, A. Sun, and H. Chen, "An evaluation of classification models for question topic categorization," *Journal of the American Society for Information Science and Technology*, vol. 63, pp. 889-903, 2012.

[7] Lukas Galke, Florian Mai, Alan Schelten, Dennis Brunsch, and Ansgar Scherp. "Using Titles vs. Full-text as Source for Automated Semantic Document Annotation". In K-CAP. ACM, 20:1–20:4, 2017.

[8] R. Bekkerman and M. Gavish, "High-precision phrase-based document classification on a modern scale," in KDD, 2011.

[9] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in WWW, pp. 91–100, 2008.

[10] M. Chen, X. Jin, and D. Shen, "Short text classification improved by learning multi-granularity topics," in IJCAI, pp. 1776–178, 2011.

[11] A. E. Mohamed, "Comparative study of four supervised machine learning techniques for classification". *International Journal of Applied Science and Technology*, 7(2), 5-18, 2017.

[12] R. Balyan, K.S. McCarthy and D. S. McNamara. Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension. In A. Hershkovitz & L. Paquette (Eds.), In Proceedings of the 10th International Conference on Educational Data Mining (EDM), Wuhan, China: International Educational Data Mining Society, 2017.

[13] S. F. Sabbeh, "Machine-Learning Techniques for Customer Retention: A Comparative Study", *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 9, No. 2, 2018.

[14] M. W. Libbrecht and W. S. Noble. Machine learning applications in genetics and genomics. Nature Reviews Genetics.16:321–32. 2015.

[15] Ayon Dey. "Machine Learning Algorithms: A Review" *International Journal of Computer Science and Information Technologies*, Vol. 7 (3), 1174-1179, 2016.

[16] G. N. Ramadevi, K. U. Rani And D. Lavanya. Evaluation of Classifiers Performance using Resampling on Breast Cancer Data. *International Journal of Scientific & Engineering Research*, Vol. 6, No. 2, 2015.

[17] R. Sathya, and Annamma Abraham, "Comparison of Supervised and Unsupervised Learning Alg orithms for Pattern Classification", *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*, Vol. 2, No. 2, 2013.

[18] S. Michalski, Ryszard, G. Carbonell, Jamie and T. M. Mitchell, Machine learning: An Artificial Intelligence Approach. Morgan Kaufmann, 1985.

[19] Ajay S. Patil, and B.V. Pawar, Automated Classification of Web Sites using Naive Bayesian Algorithm, IMECS vol-1, 2012

[20] Devroye, L. "On the equality of Cover and Hart in nearest neighbor discrimination", *IEEE Trans. Pattern Anal. Mach. lntell*. 3: 75-78.1981

[21] N.Suguna, and K.Thanushkodi 2010 An improved K-Nearest neighbour classification using genetic algorithm IJCSI Issue 4, Vol-7 pp 18-21.

[22] J. Han, J. Pei and M. Kamber, Data Mining: Concepts and Techniques, 3rd ed. *Elsevier*, 2011.

[23] C. Cortes and V. Vapnik. "Support Vector Networks", *Machine Learning*, 20: 273-297, 1995.

[24] C. D. A. Vanitha, D. Devaraj, and M. Venkatesulu, "Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection," *Procedia Computer Science*, 2015.

[25] H. Yu and S. Kim, "SVM Tutorial: Classification, Regression, and Ranking," Handbook of Natural Computing, 2009.

[26] V. Y. Kulkarni and P. K. Sinha, "Effective Learning and Classification using Random Forest Algorithm," *International Journal of Engineering and Innovative Technolgy*, vol. 3, no. 11, pp. 267-273, 2014. Gopalakrishna Murthy at el., "Performance analysis and evaluation of different data mining algorithms used for cancer classification", *IJARAI*, Vol, No.5, 2013.

[27] G. Krishna, M. Nookala, N. Orsu, B. K. Pottumuthu, and S. B. Mudunuri, "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification," *(IJARAI) International Journal of Advanced Research in Artificial Intelligence*, 2013.

[28] N. M. N. Mathivanan, N. A. M. Ghani, and R. M. Janor, "Improving Classification Accuracy Using Clustering Technique," Bulletin of Electrical Engineering and Informatics, vol. 7, no. 3, pp. 465-470, 2018.

[29] G. N. Ramadevi, K. U. Rani, and D. Lavanya, "Evaluation of Classifiers Performance using Resampling on Breast Cancer Data," *International Journal of Scientific & Engineering Research*, vol. 6, no. 2, 2015.

[30] X. Shao, H. Li, N. Wang, and Q. Zhang, "Comparison of different classification methods for analyzing electronic nose data to characterize sesame oils and blends," Sensors (Switzerland), 2015.

[31] D. R. Amancio et al., "A systematic comparison of supervised classifiers," PLoS ONE, 2014.
[32] P. Horton and K. Nakai, "Better prediction of protein cellular localization sites with the k nearest neighbors classifier.," Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology, 1997.

BIOGRAPHIES OF AUTHORS

Norsyela Muhammad Noor Mathivanan is now a doctorate student in the Center for Statistical Studies and Decision Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia under the supervision of Nor Azura Md. Ghani and Roziah Mohd Janor. Her research interest related to big data, text mining and machine learning.
E-mail: syelamohdnoor@gmail.com

Nor Azura Md. Ghani is an Associate Professor in Center for Statistical Studies and Decision Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia. She is also Head of Data Research Unit, Research Management Center, Institute Research Management & Innovation, Universiti Teknologi MARA, Malaysia and Vice Chair IEEE Computer Society Malaysia Chapter. Her expertise is big data, statistical pattern recognition and forensic statistics.
E-mail: azura@tmsk.uitm.edu.my

Roziah Mohd Janor is a Professor of Statistics at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Malaysia. Currently she is serving as the Assistant Vice Chancellor at the Institute Quality & Knowledge Advancement, UiTM and she is now overseeing all the quality initiatives of the university, including institutional accreditation, programme accreditation, quality excellence model, quality management systems, Innovation @ Work and the University Ranking Project. Since 2018, she serves as the President of the MyQAN, a quality assurance network for all Malaysian higher education institutions.
E-mail: roziahmj@uitm.edu.my