Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 6, December 2020, pp. 2518~2525
ISSN: 2302-9285, DOI: 10.11591/eei.v9i6.2630

Performance evaluation of decision tree classification algorithms using fraud datasets

Eddie Bouy B. Palad, Mary Jane F. Burden, Christian Ray dela Torre, Rachelle Bea C. Uy
Department of Information Technology, MSU-Iligan Institute of Technology, Philippines

Article Info

Article history:

Received Mar 21, 2020
Revised May 17, 2020
Accepted Jun 14, 2020

Keywords:

Classification
Cybercrime
Decision trees
Online scam
Text mining

ABSTRACT

Text mining is one way of extracting knowledge and finding out hidden relationships among data using artificial intelligence methods. Surely, taking advantage of different techniques has been highlighted in previous researches however, the lack of literature focusing on cybercrimes implies the lack of utilization of data mining in facilitating cybercrime investigations in the Philippines. This study therefore classifies computer fraud or online scam data coming from Police incident reports as well as narratives of scam victims as a continuation of a prior study. The dataset consists mainly of unstructured data of 49,822 mainly Filipino words. Further, 5 decision tree algorithms namely, J48, Hoeffding Tree, Decision Stump, REPTree, and Random Forest were employed and compared in terms of their performance and prediction accuracy. The results show that J48 achieves the highest accuracy and the lowest error rate among other classifiers. Results were validated by Police investigators where J48 was likewise preferred as a potential tool to apply in cybercrime investigations. This indicates the importance of text mining in the field of cybercrime investigation domains in the country. Further work can be carried out in the future using different and more inclusive cybercrime datasets and other classification techniques in Weka or any other data mining tool.

Corresponding Author:

Eddie Bouy B. Palad,
Department of Information Technology,
MSU-Iligan Institute of Technology,
A. Bonifacio Ave., Tibanga, Iligan City, Philippines, 9200.
Email: eddiebouy.palad@g.msuiit.edu.ph

1. INTRODUCTION

In this study, the concept of data classification is utilized in order to evaluate computer fraud or online scam data coming from incident record forms and written narratives of online scam victims in the Philippines. Classification is said to be the most popular technique [1] of finding a set of models that describes and distinguishes data classes and concepts [2, 3] for the purpose of being able to use the model to predict the class whose label is unknown [4]. According to E. B. B. Palad *et al.* [1], the main objective of the classification method is to correctly predict the target class for each situation in the information. Factually, there are several classification methods that are used to analyze information, most of which are implemented in Weka text mining tool namely, Bayesian network, fuzzy logic, decision trees, K-nearest neighbor (KNN), neural networks, and support vector machines (SVM). Despite the fact that there are many existing studies which impressed the use of data mining in the field of cybercrime investigations, there is absence or lack of studies that has been conducted in Philippine setting that applies text mining specifically on classification for unstructured texts.

Nevertheless, [1] ventures into data mining on a cybercrime data in the Philippines. It is considered as an initial attempt to exploit data mining techniques in the field of cybercrime investigations. In their study, the performance and prediction accuracy of three (3) different classifiers namely J48 decision tree, Naïve Bayes, and sequential minimal optimization (SMO) were evaluated using a relatively small online scam dataset consisting only of 14,098 attributes. Their results show that J48 classifier outperformed the other two classifiers using a set of evaluation metric. Such results were also validated through a Focus Group Discussion (FGD) wherein J48 decision tree classifier was said to be highly favored by the participants from the Philippine National Police-Anti Cybercrime Group (PNP-ACG). Participants also agreed that the analysis performed using the J48 classifier provides an opportunity for police investigators in possibly employing data mining in the field of criminal investigations in the country. Truly, based on such initial results, decision tree algorithm performs better than other methods. Even [5] argued that the advantage of using decision tree classifier as compared to other classification algorithms is that the tree can be visualized, understood and interpreted easily [6] as it has a tree-like structure. Other researchers also concluded that decision tree classifier is one of the efficient as well as popular classifiers [5, 7, 8].

Thus, the primary objective of this present study is to provide a significant improvement on the previous study of [1] based on the following identified research problems, as follows: the Filipino dataset used was relatively small; and the classifiers being compared namely Naïve Bayes, J48, and SMO were under different categories. Hence, the researchers aimed to evaluate and compare the performance of other decision tree classification algorithms as against J48 using a larger fraud dataset. Suffice to say that the the focus is shifted to decision trees. Primarily, the end goal is to assess if such algorithms can be effectively used in the country's data mining endeavors that may positively impact or at least contribute to criminal investigations. It is also hoped that the results of this study may be appended as a significant contribution to the country's dearth of scholarly works on data mining using cybercrime data. In passing, the paper is organized as follows: Section 2 provides the methods employed in conducting this research work including the dataset used; Section 3 presents the discussion of the results; and Section 4 concludes the work.

2. RESEARCH METHOD

Data mining has various techniques that include classification, clustering, regression, among others and are implemented in several data mining tools. This present paper however still exploited Weka as a tool and mainly focused on the classification technique. As a significant continuation of a prior study that used Weka, this research still utilized the pipe-lined system model presented in [9] and further exploited in [1]. This so-called pipe-lined system model is used to exploit the online scam incident reports, complaints, and statements as input, and the results of the classification process performed through utilizing several algorithmic models as output. Reseachers used various decision tree algorithms that are available in the Weka text mining tool which aid in classifying the online scam records. The online scam dataset contains instances or the classes and the attributes which help in the classification phase. Figure 1 illustrates this pipe-lined model and is further discussed in the succeeding sub-sections.

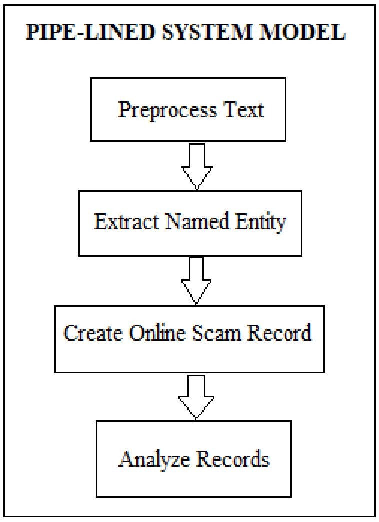

Figure 1. Pipe-lined system

2.1. Dataset used

In this research work, researchers managed to gather online scam data with a total of 49,822 mainly Filipino words coming primarily from Police incident reports as well as narrative of online scam victims. This is one of the main problems motivating the researchers in pursuing this research as the previous study only evaluated a total of 14,098 words. Hence, researchers want to investigate the performance of the classification algorithms using a larger dataset. Further, according to a PNP-ACG report, online scams consistently topped the list of most common cybercrimes since the year 2013 when the PNP-ACG was established. In a Philippine Cybercrime Report, online scam topped as the most number of cybercrimes complaints since 2015 up to 2019, and which number doubled every year. Hence, online scam is still preferred as data to the exclusion of other cybercrimes.

2.2. Pre-processing phase

Prior to converting the text files in a Weka-suitable format, the researchers manually pre-processed the data by removing sensitive personal information such as, but not limited to, names of victims, suspects, and reporting persons. Some English words were also converted to Filipino. In Weka, StringToWordVector filter was used to transform the string attribute to a set of attributes that represent the occurrence of words on a dataset [10]. StringToWordVector can be found under unsupervised attribute filter which was used in filtering text data for improved classification metrics. Using its filters, TFTransform when set to true will have the transformation term-frequency (TF) score representing textual data in a vector space be executed. However, some Filipino words like “ang”, “mga”, “sa”, and “at” are so common which occurs in almost each document. Applying the TF parameter, the documents which use the term ‘‘ang” more frequently will incorrectly get more weight without giving enough weight to more meaningful but less common terms like “bayad”, “pera” and “balanse”, and others. Therefore, an Inverse Document Frequency (IDF) factor was set to true, to combine with TF to moderate the weight of terms that occur frequently in the document set and to increase the weight of terms that occur rarely [11].

Additionally, since there are attributes that appear more frequently in the dataset which actually do not provide significant or relevant information about a class, hence stopwords handler (indicated as Filipino.txt in Table 1) in Weka was used in order to determine whether a sub string in the text is an empty word. Contrary to the findings of [12] where it was concluded that there is no need of removing stopwords as StringToWordVector makes satisfactory choices with respect to the words, the authors in this present study maintain that it is important to avoid having these stopwords while using the classifier model, because stopwords if not being eliminated during the pre-processing phase can lead to a less accurate classification of Filipino fraud data. Further, in order to break a stream of textual content up into tokens, Alphabetic Tokenizer was also employed. Table 1 summarizes all the parameters that were applied on the fraud dataset using the StringToWordVector filter with their corresponding values patterned from the study of [1]:

Table 1. StringtoWordVector parameter and values

Parameter	Values
TFTransform	False
IDFTransform	True
attributeIndices	First-last
attributeNamePrefix	“ “
doNotOperateOnPerClassBa sis	False
invertSelection	False
lowerCaseTokens	True
minTermFreq	1
normalizeDocLength	No normalization
outputWordCounts	True
periodicPruning	-1
Stemmer	NullStemmer
stopWordsHandler	WordsFormFile:Filipino.txt
Tokenizer	Alphabetic tokenizer
wordsToKeep	2,000

After the pre-processing stage, the named entities were extracted and represented as an online scam record. Similar to the study of [1], the output after extracting the named entities is a vector space model also called as document x term matrix [13]. In such a model, each row represents a document, each column a term, and each element is the frequency (TF) or the term influence in the respective document (e.g., TF-IDF). The researchers further converted the data to numeric values into a binary form for the experiment, since there are instances that attribute will take a value 1 if the word or character is present in the dataset and 0 if it is not present in the dataset. Therefore, the numeric to binary filter was selected in order to convert all the numeric values to binary. The next phase is creating a crime record in which each named entity (word or phrase) of the crime story is assigned into a corresponding online scam category.

2.3. Classification

After cleaning the data, and in order to analyze the records (the last phase reflected in Figure 1), the classification process is performed to classify the online scam into the following: banking, boiler room, buy and sell, employment, imposter, investment, lottery, online game, and online romance scams. In this study, 5 different decision tree algorithms were utilized namely Decision Stump, Hoeffding Tree, Random Forest, REPTree, and J48, each having been briefly discussed in Table 2.

Table 2. Brief description of decision tree algorithms used

Decision Tree	Description
Decision stump	Reference [2] defined a decision stump as a model consisting of a one-level decision tree. It is characteristically used by combining it with a booster algorithm [14, 15]. Such algorithm does regression (based on mean squared error) or classification (based on entropy) [16]. Furthermore, decision stumps have a robust nature that allows them to work well with large datasets and helps algorithms to make better decisions about the variables [17].
Hoeffding tree	An incremental, all-time decision tree induction algorithm that is said to learn from large information streams, assuming that the distribution generating examples does not differ over time [14].
Random Forest	A mixture of tree predictors [18] where each tree depends on the arbitrary vector values sampled individually and with the same allocation for all trees in the forest [19]. Random Forest can classify large datasets with high accuracy, a fast algorithm, and is said to not suffer from over-fitting problems which also has the ability to estimate missing data [20].
REPTree	A quick decision tree learner [14]. Constructs a decision or regression tree utilizing data gain or variance and prunes it adopting reduced error pruning (with back fitting) [21].
J48	An algorithm that produces a decision tree using C4.5 algorithms to add an ID3 algorithm and is used for classification. It is the improved version of the C4.5 decision tree classifier and has become a famous decision tree classifier [22].

Classification is a two-step process [23] involving building the classification model using training data and, the model usage [24]. Hence, in this research, the algorithms are tested using the training set and the k-fold cross validation methods. For using the training set means building up a model wherein the method is trained using all available data and then applies the results on the same input data collection [23]. On the one hand, argued that it is through cross-validation that the model built by a classification algorithm is validated as utilizing it means that the data will be divided into folds wherein some folds will be used as a testing set whereas the remaining folds for training. In this paper, a 10-fold cross-validation was utilized wherein the data were divided into 10 groups [24]. Hence, the process of testing and calculating accuracy is repeated 10 times such that each of the folds is given a chance to be trained and tested [22]. In order to compute the overall classification accuracy of an algorithm, Weka gets the average of the testing accuracy obtained from all the 10 rounds [25].

In order to evaluate the results and compare the performance of the decision tree algorithms in classifying the fraud dataset, the researchers are guided by these evaluation metrics namely: build time (in seconds), prediction accuracy, Kappa statistic, error rates, true positive (TP) rate/recall, and false positive (FP) rate. A classifier yields a good performance when low error, namely mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and root relative squared error (RRSE), is indicated in the result [5, 26, 27]. In addition, a classifier that yields Kappa Statistics closest to 1 will be evaluated as the most efficient classifier. Kappa Statistic measures the relationship between classified instances and true classes whose value usually lies between 0 and 1. When a classifier returns a value of 1, it means there is a perfect relationship between instances and classes while the value of 0 means random guessing. Moreover, a high TP rate and Recall means the model has returned most of the relevant results as these metric show the correctly classified instances [8, 28]. FP rate, on the one hand, reports instances incorrectly labelled as correct instances.

Lastly, to rationally confirm the viability of the 5 classification algorithms utilized in this research, the results were presented to police investigators and a trial court judge through FGD which was a four-hour exercise. A total of 5 participants were carefully selected by the researchers as only those police officers investigating cybercrimes and have extensive background in computing were allowed to participate. Further, it is also important to note that prior to the conduct of the validation activity, participants were briefed by the researchers that they had to sign a disclosure agreement stating that they had agreed to allow the use of whatever information they provide all throughout the validation exercise. Participants were entrusted to assess or evaluate the performance of the chosen decision tree algorithms as well as the Weka tool software.

3. RESULTS AND DISCUSSION

This section presents the data classification as performed on the gathered online scam dataset through different decision tree classifiers implemented in Weka. Results were further validated through FGD being participated by police investigators.

3.1. Classification results using the training set

The training set was first implemented to build the model. The results as reflected in Tables 3 and 4 show that among the five decision tree algorithms, Random Forest got the highest prediction accuracy with 99.51% followed by J48, REPTree, decision stump, and Hoeffding tree with 92.65%, 78.92%, 66.18%, and 58.82% respectively. Although, it took Random Forest 2.27 seconds to build the model as compared to the other classifiers which took lesser building time. Looking at the error rates shown in Table 3, Random Forest and J48 record the least errors which could translate that both classifiers have almost the same average prediction error.

Table 3. Performance parameters and their values using the training set

Parameters	Algorithms				
	Decision Stump	Hoeffding Tree	J48	REPTree	Random Forest
Time to Build the Model (in secs)	0.07	1.45	1.29	0.57	2.27
Kappa Statistic	0.2632	0	0.8742	0.6037	0.9919
Mean Absolute Error	0.1196	0.1381	0.0277	0.0777	0.0406
Root Mean Squared Error	0.2445	0.2604	0.1177	0.197	0.0865
Relative Absolute Error	86.58%	100%	20.071%	56.23 %	29.42%
Root Relative Squared Error	93.90%	100%	45.21 %	75.68 %	33.22 %
Prediction Accuracy	66.18%	58.82 %	92.65 %	78.92 %	99.51 %

For Kappa Statistic, however, it was Random Forest which obtained the best results having a value of 0.995, that is, nearest to the value of 1; as researchers translate 1 as the perfect Kappa Statistic value which translates to a perfect classifier. In addition, Table 4 shows that Random Forest returns most of the relevant data as it garnered a TP Rate and Recall of 0.995 followed by J48's 0.926. This could translate that both algorithms correctly classified instances the most as compared to the other three classifiers. Suffice to say that based on the training set results, Random Forest outperformed the other classifiers as J48 a close second.

Table 4. Summarized result on the training set

Algorithms	TP Rate	FP Rate	Recall
Decision Stump	0.662	0.436	0.662
HoeffdingTree	0.588	0.588	0.588
J48	0.926	0.078	0.926
REPTree	0.789	0.212	0.789
RandomForest	0.995	0.007	0.995

3.2. Classification results using cross-validation

A. K. Mishra *et al.* [29] suggests that k-fold cross-validation is recommended for estimating accuracy since its concept is to give an opportunity to make a prediction for each instance of the dataset. Through cross-validation, the unseen instances in training set were able to be seen and classified. Hence, in this present study, results obtained from using cross-validation are to be considered in assessing or evaluating the 5 decision tree algorithms as it gives an accurate estimate of the performance than the training set. Therefore, using the 10-fold cross-validation, the classifier which yields the highest values for prediction accuracy and recall with the least error rates is to be considered the most efficient algorithm in classifying the available online scam dataset [1, 11]. As per the results obtained with 10 different training folds, it is revealed that J48 got the highest prediction accuracy with 67.16% and a Kappa statistic of 0.4143 leaving other classifiers behind as shown in Table 5. It also recorded the least mean absolute error and relative absolute error. REPTree and decision stump both recorded prediction accuracy of 62.74%, while Random Forest and Hoeffding Tree resulted 62.74% and 60.29% accuracy rates, respectively. Furthermore, as shown in Table 6, all the classifiers yields almost the same TP Rate and Recall values; but with J48 having the best values; a Recall of 0.659 and TP rate of 0.672. At this point, it is evident that J48 algorithm outperforms all the other classifiers.

Table 5. Performance Parameters and their values using cross-validation (10 folds)

Parameters	Algorithms				
	Decision Stump	Hoeffding Tree	J48	REPTree	Random Forest
Time to Build the Model (in secs)	0.08	1.38	1.28	0.58	2.25
Kappa Statistic	0.1633	0.0826	0.4143	0.2351	0.0799
Mean Absolute Error	0.1252	0.0923	0.081	0.1186	0.1147
Root Mean Squared Error	0.2548	0.2907	0.2553	0.2549	0.2334
Relative Absolute Error	90.45%	66.69%	58.48%	85.65%	82.83%
Root Relative Squared Error	97.79%	111.58%	98.01 %	97.86 %	89.59 %
Prediction Accuracy	62.74 %	60.29 %	67.16 %	62.74 %	60.78 %

Table 6. Summarized result on cross-validation (10 folds)

Algorithms	TP Rate	FP Rate	Recall
Decision Stump	0.627	0.536	0.603
HoeffdingTree	0.603	0.590	0.590
J48	0.672	0.279	0.659
REPTree	0.627	0.414	0.627
RandomForest	0.608	0.548	0.608

3.3. Validation results

During the exercise, participants were prompted to peruse the instructions and explanations and indicate the degree to which they agree or disagree with each the provided statement using the survey scale, such as: SA–strongly agree; A–agree; N–neutral; D–disagree; and SD–strongly disagree. They were given a validation instrument prepared by the reseachers which inquired the participants to agree on (a) whether it is easy to understand the predictive model created by the classification algorithms; (b) whether the analysis performed and visualizations provided offer an opportunity for police investigators to appreciate the importance of employing data mining tools in the legal and criminal investigation domains; (c) whether the analysis and results provided by the classification are accurate in terms of patterns and insights about online scam; and finally, (d) whether the generated results produced can be applied and useful in cybercrime investigations. Such activity was conducted at the Hall of Justice in Iligan City, Philippines. The score of 5 is being set to SA, 4 for A, 3 for N, 2 for D, and lastly, 1 for SD.

Overall, in statement (a), J48 and Hoeffding tree got tied up in ratings that these 2 algorithms were easy to understand which translates that the results as well as the decision trees generated and visualized by these algorithms perked the interests of the investigators as no further explanations were needed from the researchers while participants were evaluating the results. Corrolarily, J48 scored the highest in statement (b) as participants strongly agreed with such statement, while other algorithms were just rated scores of either a Disagree or Neutral. In statement (c), only J48 was rated with a score of 5 or Strongly Agree unlike other the algorithms which were rated as Agree as well as a Disagree score. Lastly, participants strongly agreed that J48 generates results that may be applied and useful in cybercrime investigations.

4. CONCLUSION

This work was motivated by the initial study of which had somehow opened the opportunity of exploring data mining techniques in the cybercrime investigation field in the Philippines. Such study evaluated the performance and prediction accuracy of three classifiers under different categories namely J48 decision tree, Naïve Bayes and SMO using only 14,098 mainly Filipino words. Their results show that J48 classifier outperformed the other two classifiers as it resulted with the highest accuracy rate and the least error rates. J48 classifier was likewise favored by police investigators during their validation exercise. Prior to this work and as to the authors' personal knowledge, no studies had been conducted or concluded exploiting cybercrime data and data mining in the country.

In this present study, the researchers veer towards evaluating the performance of different decision tree classification algorithms namely Decision Stump, Hoeffding Tree, REPTree, and Random Forest to see which decision tree algorithm would give the best result as compared to the previously used J48 classifier; using a larger online scam dataset with 49,822 Filipino words collected from police incident reports and narrative reports from fraud victims. Based on the prediction accuracy results as well as the classification errors, the researchers conclude that even with a larger online scam dataset, J48 classifier still generates the best performance and is the most efficient in learning and classification as against other decision tree algorithms implemented in Weka. The reason for its best performance is that it correctly classifies the instances with 67.16% using the 10-fold cross validation leading to correct classification of the data as opposed to decision stump and REPTree both with 62.74% accuracy and random forest (60.78%) and Hoeffding tree (60.29%).

Another reason for its best performance is that J48 also obtains the highest values in terms of TP rate and Recall, as it generates the least amount of errors as well. Having the highest precision rate, it can be concluded that the model generated by the J48 classifier returns more relevant data than irrelevant data as against the other classifiers. Since J48 also obtains the highest recall, it can be concluded that the model it generated has returned most of the relevant results. Further, results of the validation exercise conducted still reveal that J48 is most favored by police investigators even when it is being compared to other decision tree algorithms. As future works, the researchers recommend comparing the results obtained from these decision tree algorithms by applying appropriate weight allocation and subsequent ranking approaches in order to rank those classifiers being evaluated. Classification of cybercrime data may also be further conducted using other available data mining tools or techniques.

ACKNOWLEDGEMENTS

The authors are indebted and thoroughly grateful to Municipal Trial Court in Cities Branch 01-Iligan City Presiding Judge Fe Jimeno-Dajao for the support in conducting and facilitating the focus group discussion with Police investigators in order to validate the results of the study. A special mention is also proper for Mark Harold H. Calderon of the De La Salle University–Manila, Philippines for the assistance and his commitment to continue this research work.

REFERENCES

[1] E. B. B. Palad, M. S. Tangkeko, L. A. K. Magpantay and G. L. Sipin, "Document Classification of Filipino Online Scam Incident Text using Data Mining Techniques," *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*, Ho Chi Minh City, Vietnam, pp. 232-237, 2019.

[2] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Comput. Sci.*, vol. 85, pp. 662-668, 2016.

[3] D. R. Ibrahim and A. H. Hadi, "Phishing Websites Prediction Using Classification Techniques," *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, Amman, pp. 133-137, 2017,.

[4] P. H. Patil, S. Thube, B. Ratnaparkhi and K. Rajeswari, "Analysis of Different Data Mining Tools using Classification , Clustering and Association Rule Mining," *Int. J. Comput. Appl.*, vol. 93, no. 8, pp. 35-39, 2014.

[5] S. Hussain, N. A. Dahan, F. M. Ba-Alwib and N. Ribata, "Educational Data Mining and Analysis of Students ' Academic Performance Educational Data Mining and Analysis of Students ' Academic Performance Using WEKA," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447-459, 2018.

[6] C. Anuradha and T. Velmurugan, "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance," *Indian J. Sci. Technol.*, vol. 8, no. 15, pp. 1-12, 2015.

[7] A. H. Aliwy and E. H. A. Ameer, "Comparative Study of Five Text Classification Algorithms with their Improvements," *Int. J. Appl. Eng. Res.*, vol. 12, no. 14, pp. 4309-4319, 2017.

[8] R. Panigrahi and S. Borah, "Rank Allocation to J48 Group of Decision Tree Classifiers using Binary and Multiclass Intrusion Detection Datasets," *Procedia Computer Science*, vol. 132, pp. 323-332. 2018.

[9] R. Nokhbeh Zaeem, M. Manoharan, Y. Yang and K. S. Barber, "Modeling and analysis of identity threat behaviors through text mining of identity theft stories," *Comput. Secur.*, vol. 65, pp. 50-63, 2017.

[10] C. Tsochataridou, A. Arampatzis and V. Katos, "Improving Digital Forensics through Data Mining," in *The 4th International Conference on Advances in Information Mining and Management, IMMM 2014*, Paris, pp. 1-8, 2014.

[11] M. Bilal, H. Israr, M. Shahid and A. Khan, "Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 3, pp. 330-344, 2016.

[12] A. J. Khdr and C. Varol, "Age and Gender Identification by SMS Text Messages," *2018 Int. Conf. Artif. Intell. Data Process*, Malatya, pp. 1-5, 2018.

[13] D. M. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia and J. Teixeira, "Analysis of Document Pre-processing Effects in Text and Opinion Mining," *Inf.*, vol. 9, no. 4, pp. 1-13, 2018.

[14] A. Hodžić, J. Kevrić and A. Karadag, "Comparison of Machine Learning Techniques in Phishing Website Classification," in *International Conference on Economic and Social Studies*, Sarajevo, pp. 249-256, 2016.

[15] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," *Mach. Learn. Appl. An Int. J.*, vol. 2, no. 1, pp. 1-2, 2015.

[16] P. Rajesh and M. Karthikeyan, "A Comparative Study of Data Mining Algorithms for Decision Tree Approaches using WEKA Tool," *Adv. Nat. Appl. Sci.*, vol. 11, no. 9, pp. 230-241, 2017.

[17] M. A. Shahri, M. D. Jazi, G. Borchardt and M. Dadkhah, "Detecting Hijacked Journals by Using Classification Algorithms," *Sci. Eng. Ethics*, vol. 24, pp. 655-668, 2018.

[18] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science*, vol. 127, pp. 511-520, 2018.

[19] F. Alam and S. Pachauri, "Comparative Study of J48 , Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA," *Adv. Comput. Sci. Technol.*, vol. 10, no. 6, pp. 1731-1743, 2017.

[20] E. Ahishakiye, D. Taremwa, E. O. Omulo, and I. Niyonzima, "Crime Prediction using Decision Tree (J48) Classification Algorithm," *Int. J. Comput. Inf. Technol.*, vol. 06, no. 03, pp. 188-195, 2017.

[21] W. N. H. W. Mohamed, M. N. M. Salleh and A. H. Omar, "A comparative study of Reduced Error Pruning method in decision tree algorithms," *2012 IEEE International Conference on Control System, Computing and Engineering*, Penang, pp. 392-397, 2012.

[22] G. Obuandike, A. Isah and J. Alhasan, "Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data," *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 12, pp. 44-48, 2015.

[23] G. D. K. Kishore and M. B. Reddy, "Comparative Analysis between Classification Algorithms and Data Sets (1:N & N:1) through WEKA," *Open Access Int. J. Sci. Eng.*, vol. 2, no. 5, pp. 23-28, 2017.

[24] S. G. Cho and S. B. Kim, "A Data-driven Text Similarity Measure based on Classification Algorithms," *Int. J. Ind. Eng.*, vol. 24, no. 3, pp. 328-339, 2017.

[25] A. S. Suguitan and L. N. Dacaymat, "Vehicle Image Classification Using Data Mining Techniques," in *2nd International Conference on Computer Science and Software*, Xi'an, pp. 13-17, 2019.

[26] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Comput. Sci.*, vol. 116, pp. 107-112, 2017.

[27] G. Saltos and E. Haig, "An Exploration of Crime Prediction Using Data Mining on Open Data," *Int. J. Inf. Technol. Decis. Mak.*, vol. 15, no. 9, 2017.

[28] K. S. Digamberao and R. Prasad, "Author Identification using Sequential Minimal Optimization with rule-based Decision Tree on Indian Literature in Marathi," *Procedia Comput. Sci. J. elsevier*, vol. 132, pp. 1086-1101, 2018.

[29] A. K. Mishra and B. K. Ratha, "Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis," *Int. J. Adv. Electr. Comput. Eng.*, vol. 3, no. 4, pp. 5-7, 2016.

BIOGRAPHIES OF AUTHORS

Eddie Bouy B. Palad is an assistant professor at the Department of Information Technology, College of Computer Studies of the MSU - Iligan Institute of Technology. He earned his bachelor of laws degree at the Mindanao State University in 2013; his Master in information technology at the De La Salle University - Manila in 2018; and was conferred the degree of juris doctor by the Mindanao State University in 2019. He is a member of the Integrated Bar of the Philippines (IBP) and is a pracitising lawyer in the Philippines. His research interests are cybercrime analytics, data mining, information systems, and law.

Mary Jane F. Burden is a graduate of Bachelor of Science in information systems from the College of Computer Studies at the MSU-Iligan Institute of Technology.

Christian Ray Dela Torre is a graduate of bachelor of science in information systems from the College of Computer Studies at the MSU-Iligan Institute of Technology.

Rachelle Bea C. Uy is a graduate of bachelor of science in information systems from the College of Computer Studies at the MSU-Iligan Institute of Technology.