A Comparative Study with RapidMiner and WEKA Tools over some Classification Techniques for SMS Spam

Cik Feresa Mohd Foozy¹, Rabiah Ahmad², M.A. Faizal Abdollah³ and Chuah Chai Wen⁴

¹,⁴Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
²,³Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Corresponding author: feresa@uthm.edu.my, cwchuah@uthm.edu.my, rabiah@utem.edu.my, faizalabdollah@utem.edu.my

Abstract. SMS Spamming is a serious attack that can manipulate the use of the SMS by spreading the advertisement in bulk. By sending the unwanted SMS that contain advertisement can make the users feeling disturb and this against the privacy of the mobile users. To overcome these issues, many studies have proposed to detect SMS Spam by using data mining tools. This paper will do a comparative study using five machine learning techniques such as Naïve Bayes, K-NN (K-Nearest Neighbour Algorithm), Decision Tree, Random Forest and Decision Stumps to observe the accuracy result between RapidMiner and WEKA for dataset SMS Spam UCI Machine Learning repository.

1. Introduction

There are several services on a mobile device that can be used by the spammer to launch these attacks such as Email, Mobile Browser, Short Messaging Service (SMS), Voice Call and other mobile applications. In recent years, social engineering attacks such as Spam affect the security and privacy of mobile phone users. These attacks have mistreated the use of mobile phone [1]. They will scam the mobile phone user to respond to the attacker's and get profit from this scam.

Among these services, SMS is widely used in all over the world and the cost to sent per message is considered cheaper than a call. Many SMS websites provide services to sent a message or in bulk for free to worldwide telephone number. This service not only can be used for personal, it also can be used for business marketing. By using SMS to advertise a business product, it can help sellers to improve their business profits. However, the message recipients might feel uncomfortable when receiving unwanted advertisement message from an unknown sender. This unwanted message is known as SMS Spam[2]. The mobile device is also at risk because of Spam attacks. These attacks keep growing [3] and has become a serious problem in many countries [1]. Several mechanisms have been proposed to detect these attacks in website, email and SMS. Despite this, attack frequency still increases.
Xu et al. [4] claims that SMS Spamming is, nowadays, a serious attack and manipulate the use of the SMS by spreading the advertisement in bulk. Sending unwanted SMS such as advertisement can make the user feeling disturb [5] and this against the privacy of the mobile user [6].

Several studies have been conducted in detecting SMS Spam using machine learning tool such as RapidMiner and WEKA. This paper will observe the SMS Spam detection results of dataset UCI Machine Learning repository [7] using several classification techniques in WEKA and RapidMiner.

2. Literature Review

RapidMiner and WEKA are one of the well known data mining tools and widely used by other researchers in SMS Spam detection. These tools were used to do tokenization, lemmatization, feature selection and classification.

2.1. Tokenization

Tokenization is a process to extract words in the message [8]. Tokenization is frequently applied in several attack detection frameworks. This process is one of the required processes in text messages [9]. Previously, several studies have already applied this process in SMS research including Shirani-Mehr [9], whom applied tokenization based on alphabets, the amount of dollar signs ($), the number of numeric strings and the total number of characters in the message. Najadat et al. [10] tokenized SMS using the WEKA tool to convert the text into individual number of words. Charninda et al. [11] applied the tokenization process to token the SMS text into words.

Tiago Almeida, José María Gómez Hidalgo, & Silva [12] used the same process to extract and amalgamate text into words in order to further calculate the word frequency in each of the Ham and Spam classes to concisely evaluate keywords for each class. In addition, Chen and Kan [13] tokenized English and Chinese SMS. Al-Talib and Hassan [14] developed a program to classify SMS through tokenization of its text in order to distinguish the individual verbs, nouns and adjectives by colour. Verbs were highlighted with green, nouns in blue and adjectives in an orchid colour.

Taufiq Nuruzzaman et al. [15] removed non-letter words during tokenization to reduce the amount of unnecessary words in SMS. Mahmoud and Mahfouz [16] also applied tokenization but implemented the stop word removal to remove unwanted words. Maaske Treurniet et al. [17] used tokenizer tools to perform basic tokenizing to SMS messages in Dutch. Delany, Buckley, and Greene [18] did not tokenize alphanumeric values and punctuations. However, domain names and email addresses were included.

Almeida et al. [19] resulted with two sets of tokens, of which are words that start with printable characters followed by any number of alphanumeric characters, excluding dots, commas and colons from the middle of the pattern and any sequence of characters extracted by blanks, tabs, returns, dots, commas, colons and dashes. Walkowska [20] tokenized all SMS text into words and transformed the words into root words and then proceeded to manually removed unwanted words. Hidalgo et al. [21] applied the Information Gain (IG) technique for the tokenization process to select suitable words above zero (0). Finally, Jue [22] utilised tokenization in an SMS efficiency study where the process was applied to all characters including punctuation marks, symbols and numbers.

2.2. Lemmatization

Lemmatization is a process to group and calculate the same words. This module is also applied in most attack detection frameworks. Karami & Zhou [23], in their study, applied this process and calculated word frequency as one feature. Charninda et al. [11] had grouped special characters, punctuations and uppercase letters. Ahmed, Guan and Chung [24], lemmatized single words and also combined two and three words as single features.

Tiago Almeida, José María Gómez Hidalgo and Silva [12] calculated the frequency for each word in each Ham and Spam class. Additionally, Chen and Kan [13] in their study, had grouped and sorted the words in a descending order to compare the common words between two SMS corpuses.
Meanwhile, Al-Talib and Hassan [14] grouped the words according to several classes such as occasions, greetings, friendships and sales.

Uysal et al. [25] lemmatized all words in SMS and applied the CHI2 technique to sort the words based on their occurrence frequency. Yadav et al. [26] also grouped each word in both English and Hindi SMSs based on the Ham and Spam classes. Nuruzzaman et al. [27] had also grouped individual letters from words and categorised them based on the Ham and Spam classes. Almeida et al. [19] identified two sets of token and calculated the occurrences of each token respectively.

According to the above studies, the lemmatization process aims to identify similar and unnecessary words. By identifying these reoccurrences, the removal of redundacy and the improvement of detection accuracy for SMS can be achieved. However, this process can also reduce the detection accuracy if the user selects the root word or basic words while removing other characters.

2.3. Feature Selection

Feature extraction in this study is a process conducted to extract words from the dataset. Some studies that are based on the SMS’s contents were previously conducted by Joe and Shim [28], Hidalgo et al. [21], Jie et al. [29], Yadav et al. [26], and Uysal et al. [30]. But, Eshmawi and Nair [31] proposed feature based for SMS Spam detection based on the domain knowledge.

According to Basnet et al. [32], most studies would utilise all features and eliminate those that do not positively contribute to the model. The benefit of the feature selection technique is that it helps eliminate unnecessary features in order to more concisely and accurately classify and detect SMS Phishing attacks.

2.4. Classification

A plethora of classification techniques have been implemented for Spam filtering and an equal amount if not more have been proposed to classify each corpus respectively.

Naive Bayes classifier is using estimator classes and the numeric estimator precision values are chosen based on analysis of the training data [33]. Several classification techniques are applied for SMS Spam detection (Table 2.4). Hidalgo et al. [34], Liu and Wang [35], Nuruzzaman et al. [27], K. Yadav et al. [26], Taufiq Nuruzzaman et al. [15], Uysal et al. [25], Ahmed et al. [24] and Raziki (2015) in their respective studies all utilised Naive Bayes’ technique. The Naive Bayes technique was proposed by Sahami, Dumais, Heckerman, and Horvitz [36]. It is one of the more effective techniques used to classify text documents [16]. The reason Naive Bayes is applied is because this technique is one of the simplest probability techniques that can predict the class that contains better results [24], [27]. Taufiq Nuruzzaman et al. [15] integrated Bayesian filtering to develop a simple tool to filter SMS Spam. Hidalgo et al. [34] used and extended version of Naïve Bayes filtering that filters emails and also adapts to filter SMS Spam. K. Yadav et al. [26] developed the mobile based SMS filtering system called SMSAssassin through Bayesian learning and a form of sender-blacklisting mechanism. Additionally, K. Yadav et al. [26] conducted experiments on Bayesian and SVM. Uysal et al. [25] also applied the Naive Bayes classifier in their framework.

K-nearest neighbour algorithm or known as K-NN is one of the classifier technique used to classify the similar object. K-NN depend to the "K" value to give the result. If the value K is small, then it will classify the nearest object. If using a big value of "K", the classification will classify larger. Additionally, K-NN choose a similar object or items which is similar with the attribute that has used.

In the SMS attack detection, several studies applied this techniques in their detection framework such as Liu and Wang [35], Najadat, et al. [10], Shirani Mehr [9], Ho et al. [37] and Duan et al. [38].

A decision tree is a model that can help to do decision making by identify the result by prediction. Najadat, et al. [10] has done several experiments to detect SMS Spam and one of the technique used by using the Decision tree techniques. Additionally, decision tree classification is a fast techniques but have fragmentation problem [39].

Random Forest is suitable to used with the larger dataset and works on the randomly attribute on the dataset [40]. This classification technique also can generate decision tree based on the different
sample. Several experiments has been done by Al Moubayed et al. [41] and Najadat, et al.[10] and one of the technique used is the random forest.

The Decision Stump classification techniques are one of the decision tree types and used to classify the data based on the single feature [42]. This techniques also called as one (1) rules because its detect the items by using single feature.

The experiments will be done to observe the result classification techniques of Naïve Bayes, KNN, Decision Tree, Random Forest and Decision Stump using RapidMiner and WEKA.

3. Methodology
According to Creswell [43], methodology is a strategy or plan of action that generates outcomes. The phases of the experiments are Phase 1: Data Collection, Phase 2: Experiment and Phase 3: Result Analysis as in Figure 1.

![Figure 1: Methodology Process]

3.1. Phase 1: Data Collection
Dataset that used in this experiments is UCI Machine Learning repository [7]. There are 5574 SMSs were download from the UCI Machine learning repository website. This dataset has two classes of SMS Ham (Legitimate) and Spam. The SMSs originate from four (4) sources:

a) 450 SMS from Caroline Tag’s PhD Thesis [44],

b) 425 SMS from the Grumbletext website,

c) 3375 SMS from the Ham NUS SMS corpus [13],

d) 1002 SMS Ham and 322 Spam from the SMS Spam Corpus v.0.1 Big

3.2. Phase 2: Experiment
The experiments will used UCI Machine Learning repository [7]. This datasets will be run into separate data mining tools of WEKA and RapidMiner. Each of tool will run five (5) classification techniques such as Naïve Bayes, KNN, Decision Tree, Random Forest and Decision Stump.

3.3. Phase 3: Result Analysis
The performance measurements for these experiments will be evaluated based on the result Accuracy(A) which is to measure the correct classify result. There are two (2) data mining tools involves in these experiments. This is to compare the Accuracy between these data mining tools. The reason is to investigate the result of classification accuracy by using several techniques by using the same dataset.

Several researchers such as Gonçalves and Graczyk et al.[45]Barros [46], and Zainal et al. [47] have examined the accuracy result between these two tools and the result accuracy is not same. But, the result is still acceptable when comparing both data mining tools; RapidMiner and WEKA on both SMS datasets. Table 1 shows the result from RapidMiner and WEKA. The RapidMiner shows higher Accuracy result compare to the WEKA. Average Accuracy results by using RapidMiner tool are
exceeded 90%. Eventhough, the Accuracy results by using the WEKA tool are lower than RapidMier but the Accuracy results’ range are still acceptable which are exceed 80%.

|        | RapidMiner | WEKA    |
|--------|------------|---------|
| 1. Naive Bayes                     | 90.33%    | 81.7595% |
| 2. K-NN                             | 95.53%    | 86.0411 %|
| 3. Decision Tree                    | 93.78%    | 81.2317 %|
| 4. Random Tree                      | 93.78%    | 86.0411 %|
| 5. Decision Stump                   | 95.53%    | 81.2903 %|

4. Conclusion

The SMS classification experiments are carried out using RapidMiner and WEKA tools. RapidMiner shows higher Accuracy results compare to the WEKA. However, WEKA Acuracy results are still acceptable which exceeds 80% in average. One major challenge in this experiment is that RapidMiner can execute large data meanwhile WEKA have limited memory which will suddenly stop to process the data. Some classification techniques exist for both tools but some classification techniques are different. RapidMinar and WEKA have several classification techniques. Thus, before applied the classification techniques, researchers are reminded to identify clearly the type of classification techniques which they are needed whether their desire classification techniques exists in these tools.

Acknowledgments

This research was supported by FRGS vot 1558, ORICC UTHM and Gates IT Solution Sdn. Bhd.

References

[1] M. Balduzzi, et al., "MobiPot: Understanding Mobile Telephony Threats with Honeycards," in Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security, 2016, pp. 723-734.
[2] N. F. Sulaiman and M. Z. Jali, "A New SMS Spam Detection Method Using Both Content-Based and Non Content-Based Features," in Advanced Computer and Communication Engineering Technology, ed: Springer, 2016, pp. 505-514.
[3] Darya Gudkova and N. Demidova, "Spam And Phishing in Q2 2014," August 12, 2014 2014.
[4] Q. Xu, et al., "SMS Spam Detection using Content-less Features," Intelligent Systems, IEEE, vol. PP, pp. 1-1, 2012.
[5] X. Zhang, et al., "A method of SMS spam filtering based on AdaBoost algorithm," in Intelligent Control and Automation (WCICA), 2016 12th World Congress on, 2016, pp. 2328-2332.
[6] W. Aslam, et al., "Attitudes and Behaviour of the Mobile Phones Users towards SMS Advertising: A Study in an Emerging Economy," Journal of Management Sciences, vol. 3, pp. 63-80, 2016.
[7] Tiago A. Almeida and J. M. G. Hidalgo. (2012, 23 September 2012). SMS Spam Collection Data Set. Available: http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
[8] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," Expert Systems with Applications, vol. 36, pp. 10206-10222, 2009.
[9] H. Shirani-Mehr, "SMS Spam Detection using Machine Learning Approach," 2014.
[10] H. Najadat, et al., "Mobile SMS Spam Filtering based on Mixing Classifiers," 2014.
[11] T. Charninda, et al., "Content based hybrid sms spam filtering system," 2014.s
[12] Tiago Almeida, et al., "Towards SMS Spam Filtering: Results under a New Dataset," INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE, vol. 2, p. 18, 2013.
[13] T. Chen and M.-Y. Kan, "Creating a live, public short message service corpus: the NUS SMS corpus," Language Resources and Evaluation, vol. 47, pp. 299-335, 2013/06/01 2013.
[14] G. A. Al-Talib and H. S. Hassan, "A Study on Analysis of SMS Classification Using TF-IDF Weighting," 2013.
[15] M. Taufiq Nuruzzaman, et al., "Simple SMS spam filtering on independent mobile phone," Security and Communication Networks, vol. 5, pp. 1209-1220, 2012.
[16] T. M. Mahmoud and A. M. Mahfouz, "SMS Spam Filtering Technique Based on Artificial Immune System," IJCSI International Journal of Computer Science Issues, vol. 9, 2012.
[17] O. D. C. U. Maaske Treurniet, Henk van den Heuvel, Nelleke Oostdijk, "Collection of a corpus of Dutch SMS," presented at the Proceedings of the Eight Conference of International Language Resources and Evaluation., 2012.
[18] S. J. Delany, et al., "SMS spam filtering: Methods and data," Expert Systems with Applications, vol. 39, pp. 9899-9908, 2012.
[19] T. A. Almeida, et al., "Contributions to the study of SMS spam filtering: new collection and results," presented at the Proceedings of the 11th ACM symposium on Document engineering, Mountain View, California, USA, 2011.
[20] J. Walkowska, "Gathering and Analysis of a Corpus of Polish SMS Dialogues," Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems, pp. 145-157, 2009.
[21] J. M. G. Hidalgo, et al., "Content based SMS spam filtering," presented at the Proceedings of the 2006 ACM symposium on Document engineering, Amsterdam, The Netherlands, 2006.
[22] H. Y. Jue, "Analysis of SMS Efficiency," ed: Department of Computer Science, School of Computing, National University of Singapore, 2003.
[23] A. Karami and L. Zhou, "Improving static SMS spam detection by using new content-based features," in 20th Americas Conference on Information Systems, AMCIS 2014, 2014.
[24] I. Ahmed, et al., "SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset," International Journal of Machine Learning & Computing, vol. 4, 2014.
[25] A. K. Uysal, et al., "A novel framework for SMS spam filtering," in Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on, 2012, pp. 1-4.
[26] K. Yadav, et al., "SMSAssassin: crowdsourcing driven mobile-based system for SMS spam filtering," presented at the Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, Phoenix, Arizona, 2011.
[27] M. T. Nuruzzaman, et al., "Independent and Personal SMS Spam Filtering," in Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on, 2011, pp. 429-435.
[28] I. Joe and H. Shim, "An SMS Spam Filtering System Using Support Vector Machine," in Future Generation Information Technology. vol. 6485, T.-h. Kim, et al., Eds., ed: Springer Berlin Heidelberg, 2010, pp. 577-584.
[29] C. Jie, et al., "Spam Filter for Short Messages Using Winnow," in Advanced Language Processing and Web Information Technology, 2008. ALPIT '08. International Conference on, 2008, pp. 454-459.
[30] A. K. Uysal, et al., "The Impact of Feature Extraction and Selection on SMS Spam Filtering," Elektronika ir Elektrotechnika (Electronics and Electrical Engineering), in press., 2012.
[31] A. Eshmawi and S. Nair, "Feature reduction for optimum sms spam filtering using domain knowledge," in Proceedings of the International Conference on Security and Management (SAM), 2013, p. 1.
[32] R. Basnet, et al., "Feature Selection for Improved Phishing Detection," in Advanced Research in Applied Artificial Intelligence. vol. 7345, H. Jiang, et al., Eds., ed: Springer Berlin Heidelberg, 2012, pp. 252-261.
[33] I. R. A Hamid, "Phishing Detection and Traceback Mechanism," Ph.D., School of Information Technology, Deakin University, 2015.
[34] G. V. Cormack, et al., "Content based SMS spam filtering," presented at the Proceedings of the 2006 ACM symposium on Document engineering, Amsterdam, The Netherlands, 2006.
[35] W. Liu and T. Wang, "Index-based online text classification for sms spam filtering," *Journal of Computers*, vol. 5, pp. 844-851, 2010.

[36] M. Sahami, *et al.*, "A Bayesian approach to filtering junk e-mail," presented at the AAAI'98 Workshop on Learning for Text Categorization, 1998.

[37] T. P. Ho, *et al.*, "Graph-based KNN algorithm for spam SMS detection," *Journal of Universal Computer Science*, vol. 19, pp. 2404-2419, 2013.

[38] L. Duan, *et al.*, "A New Spam Short Message Classification," in *Education Technology and Computer Science, 2009. ETCS '09. First International Workshop on*, 2009, pp. 168-171.

[39] R. Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid," in *KDD*, 1996, pp. 202-207.

[40] A. Yasin and A. Abuhasan, "An intelligent classification model for phishing email detection," *arXiv preprint arXiv:1608.02196*, 2016.

[41] N. Al Moubayed, *et al.*, "SMS Spam Filtering using Probabilistic Topic Modelling and Stacked Denoising Autoencoder," in *International Conference on Artificial Neural Networks*, 2016, pp. 423-430.

[42] K. O. Ogada, "N-grams for Text Classification Using Supervised Machine Learning," Jomo Kenyatta University Of Agriculture and Technology, 2016.

[43] J. W. Creswell, *Research Design Qualitative, Quantitative, and Mixed Methods Approaches*: Sage Publication, 2013.

[44] C. Tagg, "A Corpus Linguistics Study of SMS Text Messaging," Doctor of Philosophy, Department of English, The University of Birmingham, Birmingham, 2009.

[45] M. Graczyk, *et al.*, "Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA," in *International Conference on Computational Collective Intelligence, 2009*, pp. 800-812.

[46] P. M. Gonçalves and R. S. M. Barros, "Comparing approaches to prepare data in classification problems," in *2011 9th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, 2011, pp. 295-296.

[47] K. Zainal, *et al.*, "An Analysis of Various Algorithms For Text Spam Classification and Clustering Using RapidMiner and Weka," *International Journal of Computer Science and Information Security*, vol. 13, p. 66, 2015.