International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 2
ISSN: 2088-5334

Support Vector Machine Algorithm for SMS Spam Classification in The Telecommunication Industry

Nilam Nur Amir Sjarif [a*], Yazriwati Yahya[a], Suriayati Chuprat[a], Nurul Huda Firdaus Mohd Azmi[a]

[a]*Advanced Technology Department, Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, 54100, Kuala Lumpur, Malaysia*
E-mail: [*] nilamnur@utm.my

***Abstract*— In recent years, we have withnessed a dramatic increment volume in the number of mobile users grows in telecommunication industry. However, this leads to drastic increase to the number of spam SMS messages. Short Message Service (SMS) is considered one of the widely used communication in telecommunication service. In reality, most of the users ignore the spam because of the lower rate of SMS and limited amount of spam classification tools. In this paper, we propose a Support Vector Machine (SVM) algorithm for SMS Spam Classification. Support Vector Machine is considered as the one of the most effective for data mining techniques. The propose algorithm have been evaluated using public dataset from UCI machine learning repository. The performance achieved is compared with other three data mining techniques such as Naïve Bayes, Multinominal Naïve Bayes and K-Nearest Neighbor with the different number of K= 1,3 and 5. Based on the measuring factors like higher accuracy, less processing time, highest kappa statistics, low error and the lowest false positive instance, it's been identified that Support Vector Machines (SVM) outperforms better than other classifiers and it is the most accurate classifier to detect and label the spam messages with an average an accuracy is 98.9%. Comparing both the error parameter overall, the highest error has been found on the algorithm KNN with K=3 and K=5. Whereas the model with less error is SVM followed by Multinominal Naïve Bayes. Therefore, this propose method can be used as a best baseline for further comparison based on SMS spam classification.**

***Keywords*— short message service; spam; classification; data mining; support vector machine.**

I. INTRODUCTION

Short Message Service (SMS) become one of the most popular services used for communication. The usage of SMS services by telecom companies is increasing along with the development of communication technology and the expansion of mobile devices [1], [2]. A study made by [3-6] has shown that SMS text messaging becomes the second functionality used by mobile phone users. It reveals that the readability rate for SMS is far superior compared with the chat room. This fact has driven more than 86 percent of enterprise business are currently using or planning to use SMS for their marketing strategies. Moreover, the user is preferring using SMS messages to communicate rather than emails because while sending SMS messages, there is no need of internet connection [2].

SMS is determined as a simple and efficient way to communicate with the global. This service also can be classified as the cheapest service. Due to its simple operation, there are also possible to become an easy way to perform phishing attacks as mobile devices contain sensitive and personal information like card details, username, password, etc. The Attackers would try to find a way to steal this information from mobile devices, and SMS is one of the easiest ways [2].

Regardless of legal laws, the SMS spam problem keeps increasing day by day. There are various security measures available to control the SMS Spam problem, but they are still not so mature. An increasing number of methods to solve the issue can be classified into blacklisting, statistical purposes that are based on the frequency of occurrence of words or characters, and data mining techniques have been proposed [7].

Data mining can be defined as the process of extracting useful information from large amounts of data. It is also known as knowledge discovery (KDD) from data [6]. Over the years, many data mining techniques have been identified for the classification of SMS spam detection [8]. Example of data mining techniques applied in SMS spam classification includes Bogofilter, Dynamic Markov Compression, Logistic Regression, Support Vector Machine and Open Shortest Path First (OSBF) [9], Support Vector Machine (SVM) [7, 10, 11, 12], Naïve Bayes [13-14], C4.5, PART, Association Rule, etc. [12]. In our proposed method, the main aim is to classify the spam and ham text SMS using data mining techniques. Therefore, the objective of this

paper is to propose Support Vector Machine (SVM) for SMS Spam Classification.

The topic of SMS spam has been widely discussed all over the world. SMS texts are generally shorter than e-mail messages. The standard length for the SMS is limited to 160 characters only [3]. Furthermore, SMS does not have any standard text format, and generally, abbreviations and a lot of symbols are used inside a message. For instance, instead of expressing "what are you doing," the users frequently type only "what r u doing." Therefore, this section discusses three subsections. The first subsection A discusses the previous related works for SMS classification using data mining techniques. The second subsection B discusses the data mining technique used by previous work. The last subsection C discusses detail of the proposed method.

Many data mining techniques have been proposed and successfully employed in SMS spam classification. The proposed SMS spam classification is analyzed by Gomex Hidalgo et. Al [15]. The author built two SMS spam data collection in Spanish corpus (1157 ham and 199 spam) and English corpus (1119 ham and 82 Spam). The experiment was done using different data mining techniques such as Naive Bayes (NB), C4.5, PART, and Support Vector Machine. The modeling was evaluated using 10-cross validation. The results conclude that Naïve Bayes techniques can be successfully employed to classify SMS spam. Gordon V.Cormack et al. [16] propose email filtering techniques that require the adaption of message feature representation to acquire a good performance on SMS spam. The author performed tests using an algorithm such as Bogofilter, Dynamic Markov Compression, Logistic Regression, Support Vector Machine, and Open Shortest Path First (OSBF) on mobile spam messages with suitable feature representation. Although several experiments were done, the author suggests that the difference among all the filters are not clear and further experiment with the larger dataset is required. Again, Gordon V. Cormark et al. [10] analyze the content-based spam filtering for short messages in three contexts, which include SMS, blog comments, and email summary information. The experiments show that SMS contains fewer words, and it does not support the bag of words spam classifier, whereas the compression model-based spam filters performed quite well on the dataset.

Sarah Jane Delany et al. [17] composed a new dataset of ham messages form called GrumbleText and WhoCallsMe websites and spam messages form called SMS Spam Collection. The researcher made n analysis of some types of spam using content-based clustering and identified ten (10) clearly defined clusters, which would reflect the extent of near repetition in data caused by the similarity between different spam attacks and the breadth of obfuscation used by spammers.

Next, Nikunj Chaudhari et al. [18] has reviewed different data mining technique. The survey was identified that Naïve Bayes, Bayesian Classifiers, and Support Vector Machine (SVM) techniques are more accurate for Spam SMS filtering compares to others. The author also suggests that hybrid spam filtering techniques using a combination of two or more different techniques have increased the efficiency and accuracy of the existing Spam SMS filtering techniques. Hassan Jadat et al. [10] was running a twelve various comparison SMS classifiers where the author concludes according to the result that the Discriminative Multinominal Naïve Bayes has given the highest accuracy along with the lowest time to build the model, followed by Stochastic Gradient Descent (SGD) which also gives high efficiency. Still, it takes a long time to make the model. For the SVM, the result also shows the accuracy rate was high and less time to build the model. Hassan also suggested that the balanced dataset could be very effective with classifiers.

Zainal et al. [19] reported the findings of spam management for Short Message Service (SMS) using classification and clustering. The experiment is done using two different tools, include Rapidminer and Weka. The public dataset downloaded from UCI, Machine Learning Repository. The resulting experiment shows that SVM is the best classifier for spam classification, and K- Means is the most suitable algorithms to cluster spam messages.

Polytechnic et al. [12] demonstrate the effectiveness of the proposed Apriori algorithm based on the association rule technique for SMS Spam detection. The proposed system used structural features only instead of textual features or tokens. As a result, the good accuracy is achieved with 97.65% using rules generated by the Apriori association rule mining algorithm with minimum support 0.2 and minimum confidence 0.8 based on SMS structural features only.

II. Materials and Method

A. Data Mining Technique

Data mining techniques used by the previous work is discussed in this subsection. The algorithms include Naïve Bayes and K-Nearest Neighbor (KNN).

1) Naive Bayes (NB): is based on the Bayes' theorem, which creates a probabilistic model. This algorithm usually has an effective result in the classification of SMS messages. NB algorithm assumes that the features are statistically independent of each other, although it contributes towards the overall probability of classification [16]. Even though this assumption is unrealistic since we want the variables to interact and be dependent, it makes the probabilities fast and easy to calculate, and it proves to be an effective algorithm. The posterior probability is calculated for each class, and the prediction is made for the class by the algorithm based on the highest chance [9]. The advantage of this algorithm is that it outperforms even on a small sample size of datasets. The NaïveBayes classifier model with basic decision rule can be assigned as a class label $\hat{y} = C_k$, for some k as follows:

$$\hat{y} = \underset{k \in \{1,\ldots,k\}}{\arg\max}\, p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \qquad (1)$$

Where i is the n th component of the vector $\hat{y}$, $p(C_k)$ and $p(x_i|C_k)$ are the probabilities estimated using the training data.

2) K-Nearest Neighbour (KNN): The next algorithm is K-Nearest Neighbour (KNN), which often achieves exceptional results in classifying the text. The algorithm tries to find the K-Nearest Neighbour of a test data point and uses a majority vote to determine its class label [16]. While

predicting classification issues, the algorithm would consider the mode that is the most common class of the K utmost alike instances in the training dataset. The size of the neighbor controls the performance of this algorithm.

B. Proposed Method

This section discusses the detail of the proposed method. The phases include preprocessing and classification. The process for SMS spam classification includes the pre-processing phase (feature extraction and feature selection) and classification phase.

The dataset used for this work is SMS SpamCollection v.1, whereby it is available at the UCI Machine Learning Repository. The dataset from this repository is composed of multiple data sources such as Grumbletest Website, NUS SMS corpus, Caroline Tag's Ph.D. Thesis, and SMS Corpus v.0.1 Big. All the sources are based on English text messages. Grumbletext website contains a UK forum whereby mobile phone users make a public claim about spam messages. It consists of a collection of 425 SMS spam messages extracted from a careful scanning of the web pages. NUS SMS Corpus contains 10,000 legitimate messages that originated from the students of the same university. A subset of 3,375 messages has been randomly chosen, which include ham SMS messages. A total number of 450 ham SMS messages were collected from Caroline Tag's Ph.D. Thesis, and 1,002 ham SMS messages and 322 spam SMS messages from SMS Spam Corpus v.0.1 Big has been composed together to produce the dataset of SMS Spam Collection v.1. Each row of data contains the correct class, which is either ham or spam and followed by the raw message, as shown in Fig.1. Therefore, the total number of messages that were evaluated are 4,827 instances label as ham and where 747 cases are labeled as spam.

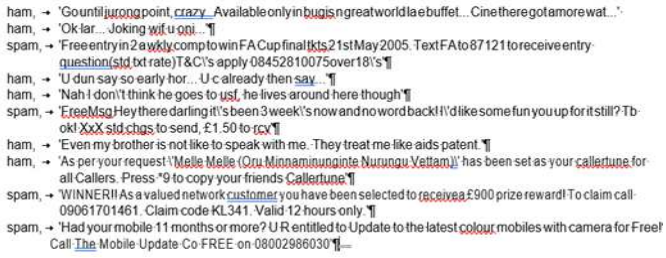
ham, → 'Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...'
ham, → 'Ok lar... Joking wif u oni...'¶
spam, → 'Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C\'s apply 08452810075over18\'s'¶
ham, → 'U dun say so early hor... U c already then say...'¶
ham, → 'Nah I don\'t think he goes to usf, he lives around here though'¶
spam, → 'FreeMsg Hey there darling it\'s been 3 week\'s now and no word back! I\'d like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv'¶
ham, → 'Even my brother is not like to speak with me. They treat me like aids patent.'¶
ham, → 'As per your request \'Melle Melle (Oru Minnaminunginte Nurungu Vettam)\' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune'¶
spam, → 'WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.'¶
spam, → 'Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030'¶

Fig. 1 Example of SMS Spam Message dataset

During the modeling phases, the dataset is divided into two different training set and testing set. For the experiment evaluation, the dataset is divided into two ratios. Dataset 1 contains ratio 80:20 for training and testing, while for Dataset 2 contains ratio 70:30 for training and testing data.

The pre-processing phase is important because it molds the data by cleaning, integrating, transforming, reducing, and discretizing. The attribute "text" contains the message strings. This message strings need to be converted into word vectors representation. The StringToWordVector technique is applied to convert the strings data, and the tokenization methods are used to remove the symbol such as.,;: '"()?!\/#&*+-_|@. For feature selection, the Information Gain technique is used to rank the feature based on the higher frequency order.

Once the preprocessing phases are done, the next phases are to use this features vector for the classification phase. Support Vector Machine (SVM) is a supervised learning algorithm that can help in analyzing data and recognizing patterns of data. SVM is developed for the numerical variables, but it also automatically converts nominal to numerical, and the input data would be normalized before being used. The procedure of the algorithm is to find an accurate line that would divide the data into a group that can be separated into classes, which are ham and spam in this dataset. The line could be straight, curved, or polygonal, and it would not be perfect in almost all cases. Therefore a margin is added around the line to relax the constraint, which would allow few instances to be misclassified but with a better result overall [10]. SVM can efficiently perform linear and non-linear classification, and the method is based on structural risk minimization [16]. The most accurate algorithm for classifying text as it focuses on separate classes.

The result for the performance measurement is evaluated based on accuracy, whereby the higher the accuracy is achieved, the more effective result would be. The accuracy is calculated based on the following formula.

$$A = TP + TN / TP+TN+FP+FN \quad (2)$$

A is defined as the accuracy in percentage value of messages being correctly classified. TP is True Positive, where spam messages are classified as spam, TN is True Negative where ham messages are classified as ham, FP is False Positive where the ham messages incorrectly classified as spam, and FN is False Negative where spam messages are wrongly classified as ham. Another performance measurement would be evaluated in this study is Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) and Kappa Statistic. The next section discusses the resulting experiment based on the proposed method.

III. Result and Discussion

The proposed method SVM is compared with the other three classifiers, such as Naïve Bayes, Multinominal Naïve Bayes, and K-Nearest Neighbour (KNN) with different K=1,3 and 5. The following tables show the resulting experiment of different classification algorithms on the SMS Spam Collection dataset.

TABLE I
Comparison Result Accuracy for Training Based on Ratio 80:20

Algorithm	Training Set (80%)		
	Accuracy	Correct Instance	Incorrect Instance
Support Vector Machine	97.47%	4346	113
Naïve Bayes	95.81%	4272	187
Multinomial Naïve Bayes	86.59%	3861	598
KNN (1)	95.56%	4261	198
KNN (3)	93.81%	4183	276
KNN (5)	92.71%	4134	325

Table I and Table II show the comparison result training and testing with the ratio 80:20. Table III and Table IV show the comparison result training and testing with the ratio 70:30. Tables V and IV present the comparison result based

on the MAE, RMSE, and Kappa Statistic, where each of the tables is being split into the training and testing ratio 80:20 and 70:30.

TABLE II
COMPARISON RESULT ACCURACY FOR TESTING BASED ON RATIO 80:20

Algorithm	Testing Set (20%)		
	Accuracy	Correct Instance	Incorrect Instance
Support Vector Machine	98.91%	1114	1
Naïve Bayes	96.95%	1081	34
Multinomial Naïve Bayes	86.64%	966	149
KNN (1)	98.61%	1113	2
KNN (3)	91.39%	1019	96
KNN (5)	89.33%	996	119

TABLE III
COMPARISON RESULT ACCURACY FOR TRAINING BASED ON RATIO 70:30

Algorithm	Training set (70%)		
	Accuracy	Correct Instance	Incorrect Instance
Support Vector Machine	98.26%	3833	68
Naïve Bayes	95.90%	3741	160
Multinomial Naïve Bayes	86.23%	3364	537
KNN (1)	95.80%	3737	164
KNN (3)	93.64%	3653	248
KNN (5)	92.13%	3594	307

TABLE IV
COMPARISON RESULT ACCURACY FOR TESTING BASED ON RATIO 70:30

Algorithm	Testing set (30%)		
	Accuracy	Correct Instance	Incorrect Instance
Support Vector Machine	98.76%	1669	4
Naïve Bayes	96.11%	1608	65
Multinomial Naïve Bayes	87.45%	1463	210
KNN (1)	97.92%	1667	6
KNN (3)	95.28%	1594	79
KNN (5)	91.51%	1531	142

Based on Table I and Table II, the resulting experiment shows that the SVM algorithm gives the highest accuracy of 98.91% for training, and for testing KNN with K=1 gives the highest accuracy 98.61%, respectively. The algorithm Multinomial Naïve Bayes gains the lowest accuracy compared to other algorithms, which is 86.63% for train and 86.64% for testing. Similarly, for the resulting experiment based on the ratio 70:30 for training and testing, as shown in Table III and Table IV. Support Vector Machine shows the highest result with each train, and the test result is 98.26% and 98.76%. Based on the result shows that the Multinomial Naïve Bayes algorithm shows the lowest accuracy rate for both ratio 80:20 and 70:30. Fig 2 and Fig 3 shows the graphical result comparison of the accuracy based on the training and testing dataset with a different ratio.

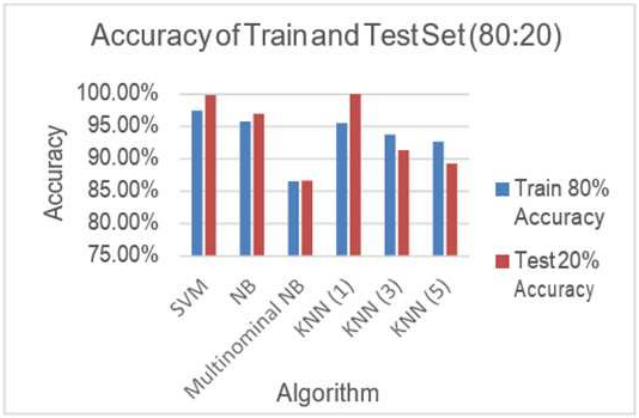

Fig 2. Comparison result training testing ratio 80:20

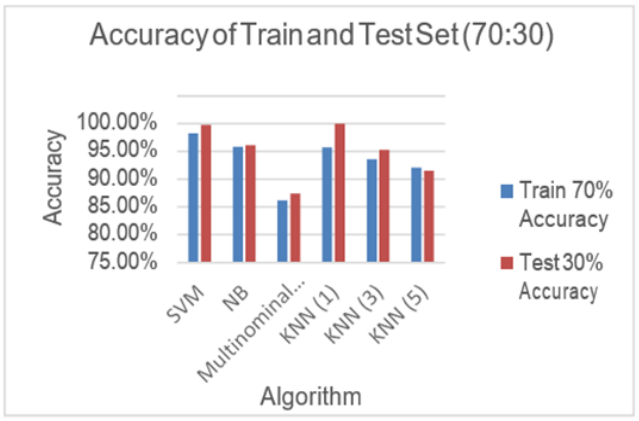

Fig 3. Comparison result training testing ratio 70:30

Next, to verify the reliability of the collected data and to check the validity of the data, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Kappa Statistics parameter is used. These metrics compare an observed accuracy with an expected accuracy and measures the agreement of prediction with the true class, for the Kappa score showed good results if the score near to 1.0, where the score signifies complete agreement. Meanwhile, for the MAE and RMSE, the lowest error is indicated as a good result. Tables V and VI show the result of the verification validity of the accuracy achievement.

TABLE V
COMPARISON RESULT ACCURACY FOR TRAINING BASED ON RATIO 70:30

Algorithm	80:20 Dataset		
	Kappa Statistic	Mean Absolute Error	Root Mean Square Error
Support Vector Machine	0.8829	0.0253	0.1592
Naïve Bayes	0.8276	0.0441	0.2018
Multinomial Naïve Bayes	0.8769	0.0251	0.1621
KNN (1)	0.7799	0.0462	0.2108
KNN (3)	0.6695	0.0661	0.2238
KNN (5)	0.5934	0.0842	0.2413

TABLE VI
COMPARISON RESULT ACCURACY FOR TRAINING BASED ON RATIO 70:30

Algorithm	80:20 Dataset		
	Kappa Statistic	Mean Absolute Error	Root Mean Square Error
Support Vector Machine	0.9245	0.0174	0.132
Naïve Bayes	0.8337	0.0428	0.1975
Multinomial Naïve Bayes	0.9031	0.024	0.1474
KNN (1)	0.7979	0.0445	0.2054
KNN (3)	0.6697	0.0712	0.2299

As shown in Table V and VI, the algorithms are compared based on the kappa score. The experiment result shows that the average Kappa score of the evaluated algorithms is 0.8829 and 0.9245 for the ratio 80:20 and 70:30 datasets, respectively. The SVM algorithm achieves the highest Kappa score. As also presented the result of MAE and RMSE for the comparison algorithm. The lowest the error is, the accurate the model will be. Comparing both the error parameter, the highest error has been found on the algorithm KNN with K=3 and K=5, whereas the model with less error is SVM, followed by Multinominal Naïve Bayes.

IV. CONCLUSION

Spam SMS messages are increasing, and it is one of the critical issues these days. Filtering the spam messages and identifying and labeling the spam instances is a challenge to resolve, hence this study used classification algorithms on the SMS Spam Collection Dataset. In this paper, we discussed several machine learning classifiers to classify the SMS Spam Collection Dataset. This study trained and tested the use of selected algorithms. The performance comparison of each algorithm suggested the best suitable algorithm. In measuring the performance of each classifier, the following features were considered: higher accuracy, less processing times, highest kappa statistics, low error, the lowest false positive instance.

Carefully considering all the factors, we identified that Support Vector Machines (SVM) could result in better performance than other classifiers, and it is the most accurate classifier to detect and label the spam messages with an average accuracy is 98.9%. Therefore, this proposed method can be used as the best baseline for further comparison based on SMS spam classification.

ACKNOWLEDGMENT

The authors are grateful to the Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia (UTM) for their educational and financial support. This work is conducted at Advances Informatics School (AIS) under Cyberphysical Systems Research Group (CPS RG) and funded by Universiti Teknologi Malaysia (GUP Tier 1: Q.K130000.2538.18H42).

REFERENCES

[1] T. a Almeida, J. María, G. Hidalgo, and T. P. Silva, “Towards SMS Spam Filtering: Results under a New Dataset,” *Int. J. Inf. Secur. Sci. T.*, vol. 2, no. 1, pp. 1–18, 2012.

[2] Choudhary, N., & Jain, A. K. "Towards Filtering of SMS Spam Messages Using Machine Learning Technique". *Advanced Informatics for Computing Research*, vol 712, pp. 18–30, 2017 https://doi.org/10.1007/978-981-10-5780-9.

[3] Pham, T.H., Le-Hong, P. "Content-based approach for Vietnamese spam SMS filtering", *in: 2016 International Conference on Asian Language Processing (IALP)*, pp. 41–44, 2016

[4] Bank Negara Malaysia. "Alert on SMS Scam and Fake Website Involving Bank Negara Malaysia Name". [Online]. Available: http://www.bnm.gov.my/index.php?ch=en_announcement&pg=en_announcement&ac=536. 2017

[5] Davenport, J.R.A., DeLine, R., "The Readability of Tweets and their Geographic Correlation with Education" https://arxiv.org/abs/1401.6058. 2014

[6] Dermawan, A., "Accountant loses RM510,000 to “Bank Negara” scammers". *News Straits Time*. October 19, 2017. 2017

[7] Kaya, Y., & Faruk, Ö. "A novel feature extraction approach in SMS spam filtering for mobile communication: one-dimensional ternary patterns". *Security and Communication Networks*, vol. 9 no.17, pp.4680-4690, 2016

[8] Abdulhamid, S.M., Latiff, M.S.A., Chiroma, H., Osho, O., Abdul-Salaam, G., Bakar, A.A., Herawan, T., "A Review on Mobile SMS Spam Filtering Techniques". *IEEE Access* pp. 1–1, 2017

[9] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., "Data Mining: Practical machine learning tools and techniques". Morgan Kaufmann. 2016.

[10] R. Article, H. Sajedi, G. Z. Parast, and F. Akbari, “SMS Spam Filtering Using Machine Learning Techniques: A Survey,” *Mach. Learn. Res.*, vol. 1, no. 1, pp. 1–14, 2016.

[11] P. Chhabra, R. Wadhvani, and S. Shukla, “Spam Filtering using Support Vector Machine,” vol. 1, no. 2, pp. 3–5, 2010.

[12] Polytechnic, S., & Region, K. "SMS Spam Detection Using Association Rule". *Journal of Theoretical and Applied Information Technology*, vol. 96, no.12, pp. 3962–3972, 2018.

[13] H. Najadat, N. Abdulla, R. Abooraig, and S. Nawasrah, “Mobile SMS Spam Filtering based on Mixing Classifiers,” *Int. J. Adv. Comput. Res.*, vol. 1, pp. 1–7, 2014.

[14] T. a Almeida, J. M. G. Hidalgo, and A. Yamakami, “Contributions to the study of SMS spam filtering: new collection and results,” *Proc. 11th ACM Symp. Doc. Eng.*, pp. 259–262, 2011.

[15] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García, “Content based SMS spam filtering,” *Proc. 2006 ACM Symp. Doc. Eng. - DocEng ’06*, no. January, p. 107, 2006.

[16] G. V. Cormack, J. M. G. Hidalgo, and E. P. Sánz, “Feature engineering for mobile (SMS) spam filtering,” *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’07*, pp. 871, 2007.

[17] S. J. Delany, M. Buckley, and D. Greene, “SMS spam filtering: Methods and data,” *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9899–9908, 2012.

[18] N. Chaudhari, P. Jayvala, and P. Vinitashah, “Survey on Spam SMS filtering using Data mining Techniques,” *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, vol. 5, no. 11, pp. 193–195, 2016.

[19] Zainal, K., Sulaiman, N. F., & Jali, M. Z. "An Analysis of Various Algorithms for Text Spam Classification and Clustering Using RapidMiner and Weka". *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 13, no 3, pp. 66–74, 2015