Int. J. Advance Soft Compu. Appl, Vol. 10, No. 3, November 2018
ISSN 2074-8523

SMS Spam Classification using Vector Space Model and Artificial Neural Network

Wan Nazirul Hafeez Wan Safie, Nilam Nur Amir Sjarif [1,2]*, Nurulhuda Firdaus Mohd Azmi[1,2], Siti Sophiayati Yuhaniz[1,2], Rasimah Che Mohd Yusof[1], Suraya Yaacob[1]

[1]Advanced Informatics Department (UTM AIS)
Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jln Sultan Yahya Petra, 54100 Kuala Lumpur
e-mail: wannazirulhafeez@icloud.com, nilamnur@utm.my*, huda@utm.my, sophia@utm.my, rasimah.kl@utm.my, suraya.yaacob@utm.my
[2]Machine Learning for Data Science (MLDS) Interest Group, Jln Sultan Yahya Petra, 54100 Kuala Lumpur

Abstract

As there are increasing numbers of mobile subscriber and the market demands of reaching customer personally, Short Message Service (SMS) has become a target of unsolicited text message known as Spam that resulting waste in time, money, and privacy. Many text classification methods using traditional machine learning algorithm has been proposed to prevent spam. However, none of these solutions can guarantee 100% spam-proof solution as each filtering and modeling technique has their own weaknesses and strengths. The objective of this paper is to propose SMS spam classification using Vector Space Model (VSM) and Artificial Neural Network techniques on the publicly available SMS dataset. The result shows a significant improvement based on the accuracy which is 99.10%. This paper will contribute on practical applications and provide contribution to the body of knowledge.

Keywords: *SMS Spam Classification, Vector Space Model, Artificial Neural Network*

1 Introduction

Today's semiconductor industry follows closely the Moore's Law projection which stated that the density of transistor in integrated circuit doubles approximately every year since the invention [1]. This in turn made the electronic device such as mobile phones become cheaper, affordable, handy and provide many functionalities without requires one to have IT or mathematical literacy. There are estimated around 62.9% of the world population owned a mobile phone in 2016 and this numbers are expected to exceed more than five billion marks by 2019 [2]. 3G and 4G generation of mobile technology now day have provided users with many other functionalities that help they organized their daily live including clock, alarm, calendar, news and nevertheless connecting peoples effectively.

A study in Africa shows that text messaging or SMS (Short Message Service) are second functionalities widely used by the mobile phones owner [3]. Although in recent years there are various alternative option for messaging such as through email or social messaging, SMS has the advantage of reaching and readability to everyone with mobile phone personally. A survey by Frost & Sullivan in 2010 reveal that readability rate for text messages are far superior than emails; a whopping 98% compared to 26% for email. Followings study by Oregon University in 2014 states that 96% users use text message with the readability rate of 98% compared to tweets (29%) and Facebook post (12%). This fact has driven more 89% of enterprise businesses are currently using, or planning on using to use SMS for their marketing strategies [4]. SMS is cellular service that provide generous millions of dollars of income for mobile operator yearly as indication shows that billions of SMS are sent daily. According to [5], the compounded annual growth rate (CAGR) of SMS market is expected to surpass US$78 billion by 2022 which seen 4.9% increment throughout the period.

As there are growing numbers of marketers, the downside is that mobile phones users are becoming target of unwanted or junk text message; also known as SMS spam. As the cost of sending SMS are dropping and even free of charge for some countries, it has contributed to the affordability for SMS spamming. The main issue with SMS spam besides annoying the user, it also disturbs the privacy, waste of time and effort, and might incur additional cost to if user accidentally subscribed the service provided from the spam SMS. When SMS received on the mobile phone, the alert tone might disturb the privacy and add unnecessary frustration to the users as they must take out their phone and open the SMS before deleting it. There are no less 50 000 reports and complains received by Australian government each month in SMS related issues [6].

As SMS corpus have the tendency of growing bigger and complex along the time, proper machine learning algorithm might be helpful to classify or filter the SMS spam characteristic. Artificial neural network has gained many attentions as a recent trend in machine learning that models highly non-linear representations of data such in image and speech recognition [17], anonymous driving vehicle [18] and natural language processing [19]. Although the concepts have been established for decades, the advancement of computing power in recent years help to unleash the potential of artificial neural network. Categorizing spam and ham(legitimate) SMS is a challenging and tedious task as spam numbers keep growing and their characteristic are constantly changing. Spammers keep altering their techniques to avoid these spam SMS from being detected thus further create a barrier for the system to accurately detect the spam.

There is various studies and approach has been proposed to classify SMS spam with different performance metrics measurement such as accuracy, precision rate, and false positive rate. It shows that SVM and Random Forest algorithm are mostly used by researchers [20] as it provided highest accuracy among all. However, there are still other available algorithm that are not fully explored in SMS domain. Additionally, machine learning algorithm preferred a good features engineering to improve the classification. Manual feature engineering is difficult, time-consuming and requires domain expertise.

In this paper, we will propose Vector Space Model (VSM) and Artificial Neural Network as an alternative for SMS spam classification. Comparison will be made with another two-machine learning algorithm that is Random Forest (RF) and Support Vector Machine(SVM) while the performance will be measured base on three (4) metrics of accuracy, precision, false positive rate and time.

The rest of the paper is organized as follows. Section 2 covers the related works and the literature review. The methodology and propose algorithm on how the experiment will be carried out are presented in section 3 and 4, while section 5 presents the results and discussion on the finding. Finally, section 6 concludes this paper.

2 Related Work

In this section, we provide the related work of the proposed work and critical analyze them.

Joe and Shim [7] applied Support Vector Machine (SVM) algorithm and used experience based learning to detect spam SMS. The proposed system works by isolating words from sample SMS using pre-processing and thesaurus technique. In feature extraction, the relevant synonyms, antonyms, and hyponyms words are combined into single word based on frequency for chi-square statistics with feature vector was set to 100, 150, 200 and 300. SVM hyperplane will be modify base on learning data from generated vector values. A learning process is completed through SVM after all feature vectors is marked 0 or 1. In addition, the author used Gaussian Radial Basis Function (RBF) as a kernel function with constant value set at 10, 20, 40 while gamma value set as 0.01, 0.05 and 0.1. The performance is evaluated using Spam / Non-Spam Precision and Recall. Results shows that proposed system give optimal performance when feature value set to 150, constant value of 20 and gamma value as 0.01.

Mathew and Isaac [8] compared some of the well-known algorithm used for spam detection to find the methods that suite SMS text context better. The algorithms use in their study do not read strings therefore all data are converted to feature vector first. The results of the study distinguished that Bayesian methods able to give 98% of accuracy. Zainal, et al. [9] also study and compared various classification and clustering algorithm. The performance is measured through the accuracy and execution time taken to complete the task for each respective tool. The experiments show that SVM give the highest accuracy with average value of 99.93% respectively. Almeida, et al. [9] compare several machine learning algorithms on SMS corpus dataset. In the experiment, the algorithms will be combine with two types of different tokenizer; tokenizer 1 is tokens start with a printable character and followed by any number of alphanumeric characters while tokenizer 2 comprise of any sequence of characters separated by delimiter. The performance of SVM outperforms other algorithms by acquiring highest accuracy rate of 97.64% and able to correctly identify 83.10% spam SMS while misclassified slightly 0.18% of ham SMS.

Serrano, et al [10] proposed technique of preserving order sequence of the data in feature space to capture writing- style of SMS corpus. Instead of using the word in the corpus as the feature vector, this study used Sequential Labelling Extraction that assigning tags to term according to word grammatical function. The results show that Decision Tree with Naïve Bayes (DTNB) give highest Area under ROC Curve with value of 0.970 compared to SVM with value of 0.910 but SVM present highest ham misclassification only 1%. This study proved that SMS spam classification can works well using sequence of features.

3 Methodology

3.1 Vector Space Model

Vector Space Model is a statistical model used to vectorize the words in representing documents where each word become the features with term weight as the values [21]. It is one of the techniques in Information Retrieval (IR) domain that works effectively in text collection dataset. Early adoption of IR approach is using Boolean representation consist of AND, OR and NOT. However the major shortcoming of this technique is there is a lack of inherent notion of ranking the document. In 1992, various US Government agencies has sponsored the inception of Text Retrieval Conference (TREC) that aims to encourage studies and public contribution in IR domain using large text collection [16]. Three of the most used model in IR studies is vector space model, probabilistic model, and inference network model. Under TREC, the modified and enhanced techniques of IR have been developed with objective to effectively retrieve information from large text corpus. In vector space model, each word is selected as individual independent term thus the words in documents are likely to be transformed into a very high dimensionality vector space. As SMS has limited of 160 characters hence number of term per message are very sparse considered there a million of vocabulary term existed. In addition, vector space model is operated in a positive quadrant and there will be no negative value assigned. Term frequency-Inverse document frequency (TFIDF) are selected as the term weight method that reflect the importance of a word to a document in the whole corpus.

3.1.1 Term Frequency-Inverse Document Frequency (TFIDF)

TFIDF is the often-weighting method used to in the Vector Space Model, particularly in IR domain including text mining. It is a statistical method to measure the important of a word in the document to the whole corpus. The term frequency is simply calculated in proportion to the number of occurrence a word appears in the document and usually normalized in positive quadrant between 0 and 1 to eliminate bias towards lengthy documents [22]. To construct the index of terms in TFIDF, punctuation is removed, and all text are lowercase during tokenization. In addition, there is no stemming on the text and no stopwords are removed.

The first two letter TF or term frequency refers to how important if it occurs more frequently in a document. Therefore, the higher TF reflects to the more estimated that the term is significant in respective documents. Additionally, IDF or Inverse Document Frequency [23] calculated on how infrequent a word or term is in the documents. The weighted value is estimated using the whole training dataset. The idea of IDF is that a word is not considered to be good candidate to represent the document if it is occurring frequently in the whole dataset as it might be the stopwords or common words that is generic. Hence only infrequent words in contrast of the entire dataset is relevant for that documents.

3.2 Artificial Neural Network

Similar to biological neuron function in human brain, artificial neuron function sends the data to a connected neuron present within the hidden layers [24]. The level of deepness is determined according to the number of layers, where the deeper the layers the more non-linear function will be, and more complex functions can be learned. This kind of neural network structure also known as Multilayer perceptron (MLP). MLP is a feed forward neural network type that basically comprise of collection if artificial neurons organized in layers with the most used training algorithm for MLP is back-propagation algorithm [25]. Each layer comprises of nodes or neurons which input nodes represent the input variables while output layers represent the product classes. Hidden neurons inside the hidden layers help translate the relationship between inputs and outputs layer including all hidden layers.

MLP is a supervised learning algorithm that trained by exposing the network the desired output from each individual input. The first phase of training known as forward phase involved passing the generated network output starting from the input layer then go through hidden layers networks until it passed to the output layers. The second phase of training or known as backward phase involved calculating the network error which is the measure of difference between generated network output and the desired output. The goal of back-propagation is to minimize this network error by propagating backwards through the network to adjust its weight until acceptable threshold and optimum network performance is meet [26].

It is quite difficult to guessing the best training algorithm for any given network case. The gradient descending optimization algorithm plays a major role that determine the performance of a given network and most widely used approach is gradient descent algorithm [27]. Stochastic gradient descent (SGD) is a stochastic approximation algorithm that use iterative and recursive update rule that can be used in gradient descent optimization problem.

Fig.1 visualize the SGD in pseudocode form. In deep learning, setting the initial value of learning rate has been an issue that don't have a discrete solution as setting the rate too low may cause the learning slow to converge while setting it too high may cause it to diverge wildly. Adaptive Moment Estimation (Adam) is a variant and improvement of basic SGD method in a way that is works well on training time and validation score on relatively large dataset or high dimensional dataset [28].

- Choose an initial vector of parameters w and learning rate η.
- Repeat until an approximate minimum is obtained:
  - Randomly shuffle examples in the training set.
  - For $i = 1, 2, \ldots, n$, do:
    - $w := w - \eta \nabla Q_i(w)$.

Fig.1 Stochastic gradient descent pseudocode.

3.3 Dataset

The dataset used in this research are obtained from freely available public dataset SMS Spam Collection v.1 [11] that have been collected for spam studies. This dataset is the updated version from [12] has been used by various researchers as credible research dataset in their studies as described in previous the section. It is a corpus of compilation from free for research sources on the internet that contain 5574 un-encoded SMS message in English language with two attributes which hold the SMS and the label of 747 spam and 4827 hams as illustrated in Table 1. A subset of more than 61 % of the data set originated from students in National University of Singapore (NUS) [13] compared to data originated from UK [14], thus it can be concluded that the English slang used in the message will be more towards Singaporean English.

Table 1: Data Structure

Class	***Features***
Hams	He is there. You call and meet him
Spams	Want 2 get laid tonight? Want real….
Hams	Good stuff, will do.

4 Propose Model

In this paper, the new proposed SMS spam classification model scheme using vector space model with artificial neural network algorithm flow is summarized as in Fig.2.

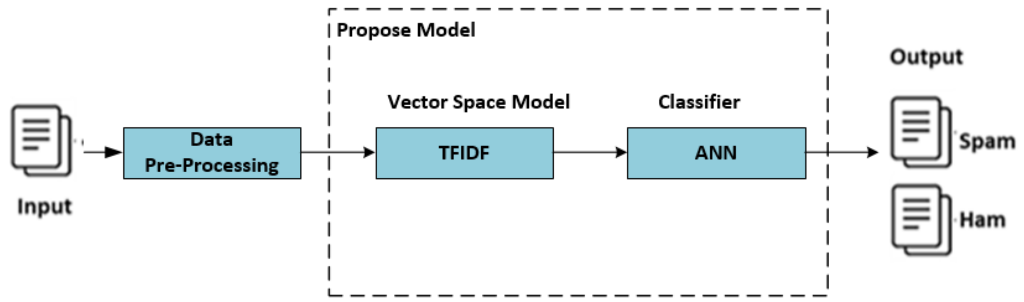

Fig.2 Propose SMS Spam Classification Scheme

Based on this scheme, the transformation or vectorizing the dataset produced a new total of 7800 attributes from originally 2 attributes in the dataset. Each word now having their own weight values relative to each row of data as shown in the snippet in Table 2.

Table 2: Vector Space Model Attributes with Tern Weight

	all	am	and	any
0	0	0	0	0.158032
1	0	0	0	0
2	0	0	0	0.177179
3	0	0.343564	0	0.251297

For better visualization, *t-distributed Stochastic Neighbor Embedding* or t-SNE was chosen as it able to visualize high dimension data space as shown in Fig.3. The results show how ham and spam vector grouped almost closely together and mostly overlap or redundant between each other. This mean that the same or similar words or attributes might belong on both classes. We also can’t dictate that this certain group of word belongs to spam or ham class.

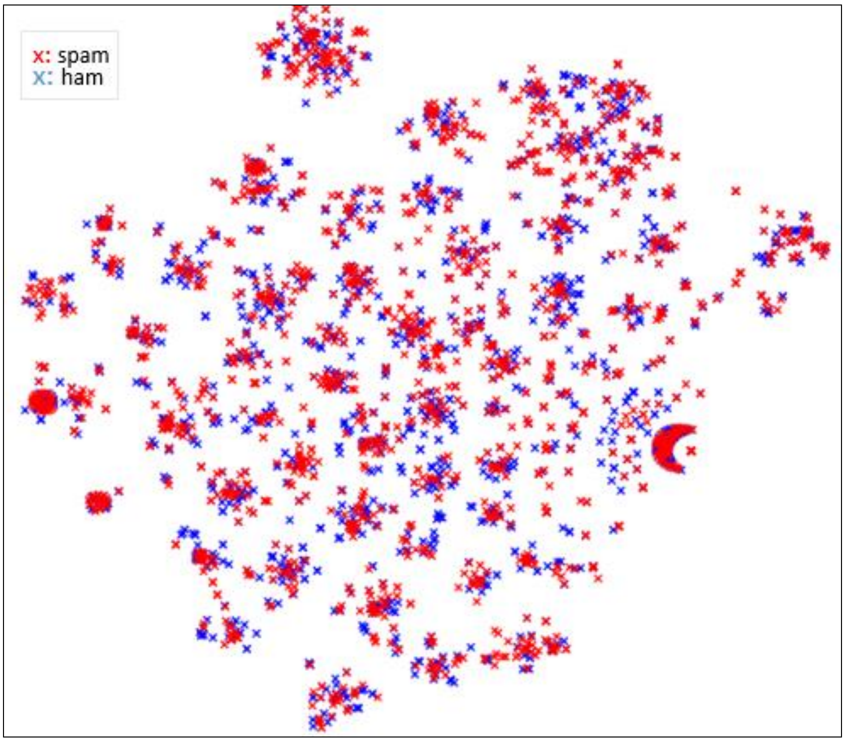

Fig.3 TSNE visualization of SMS vectorized data

For the classification, we used ANN algorithm. This ANN classification is a supervised learning as it is trained from a labeled dataset which the algorithm learns the relationship between the attributes and the label output. From the whole dataset, it was split into training and test set with 80:20 ratio.

During the learning phase using 80% train data, the best learning rate value and gradient descending optimization scheme will be determined through initial experiment. The learning rate selected for comparison starting from 0.0001, 0.01, and 0.1 while gradient descending optimization is narrowed to Stochastic Gradient Descent (SGD) and Adam. In addition, effect from number of neuron per hidden layer and number of hidden layers also will be studied.

4.2 Performance Measurement

There are two (2) classifiers algorithm that has been selected to be used for comparison in this paper which are Random Forest (RF), and Support Vector Machine (SVM). In random forest, it has the underlying data structure as a decision tree, but with random selection of features to split on. SVM works by producing linear vector or hyper-plane that try to segregate the target class and give the one with the maximum margin.

To compare the results, four (4) performance measures will be use that is: Accuracy, Precision and False Positive Rate and time. To compute these evaluation metrics, four identifiers in confusion matric will be describe that is True Positive (TP) refers to SMS that correctly predicted as spam, True Negative (TN) refers to SMS that are correctly predicted as ham, False Positive (FP) refers to SMS that are predicted as spam while they are not, and False Negative (FN) that refers to misclassified SMS spam as a ham.

Accuracy measurement refers to the SMS that are correctly classified overall while precision on the other hand refers to how many predicted spam SMS is being spam SMS. The false positive rate or ham blocked reflects the probability of misclassifying a legitimate SMS as a spam. Ham blocked refers to fraction of legitimate SMS that wrongly classified as spam. This paper intends to look for a lower ham blocked as it more expensive or offensive for user not receiving legitimate SMS that marked as spam. The impact if user’s banking Transaction Authorization Code (TAC) number SMS is wrongly classified as spam will be highly devastating for the user. In this section, the results of applying three different machine learning algorithms are presented.

5 Results, Analysis and Discussions

5.1 Learning Rate & Gradient Descending Optimization

The results for both learning rate and gradient descending optimization parameters are being compared together and visualized as the training loss curves in Fig.4. The highest learning rate experiment with SGD that is 0.1 shows a deep dive in initial epoch and continue with good learning curve towards the end of epoch. However, all three SGD method exhibit slow converging rate and not yet achieve minimal cost error after meeting the maximal number of 200 epoch. In overall, Adam optimization performed better than basic SGD in all three different learning rates. Early stopping is triggered in all three Adam optimizations as there is no significant improvement for training loss in two consecutive iteration. Adam with 0.01 is the best learning rate value as it converges quickly, highest training set score and give the lowest training set loss.

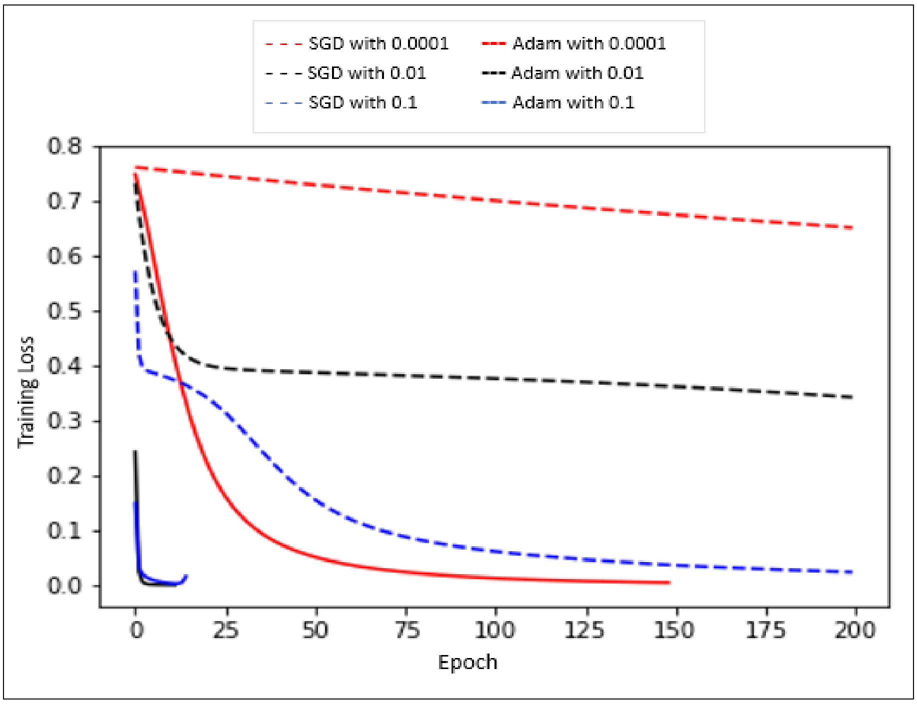

Fig.4 Training loss curve

5.2 Hidden Layers & Neurons

The network consists of input layer, hidden layer and output layer. The number of hidden layers and neurons per layer may affect the performance of our modelling. Therefore, the experiment tries to compare the extreme value of 800 neurons using one layer in initially while the rest of the test using 200 neurons for each hidden layer. This is shown in Table 3.

Table 3: Results evaluation against performance metrics

No. Hidden Layers	Neurons per Layer	Training Epoch	Training Loss	Spam Accuracy (%)	Spam Precision	False Positive Rate (%)	Learning Time (seconds)
1	800	94	0.0041677	99	0.99	0.1	170
1	200	148	0.0057227	99	0.99	0.1	34
2	200, 200	61	0.0020272	99	0.99	0.1	25.7
3	200, 200, 200	39	0.0009285	99	0.99	0.1	27.2

While spam accuracy, precision and false positive rate remain constant; same values for all experiment, three hidden layers with 200 neurons per layer proved to the best and acceptable due to the result of learning time of 27.2 seconds and lowest training epoch and loss.

5.3 Benchmark against Comparison Algorithm

As shown in Table 4, the accuracy of modelling using VSM model is 99% for ANN, 97.13% for Random Forest, and 87.09% for SVM. Based on our performance metrics, SVM are unable to classify any spam message at all with precision value of 0% thus could not provide any value for False Positive Rate calculation. In term of classifying spam correctly, ANN and Random Forest give precision value of 0.99 and 0.98 respectively.

On the other hand, ANN give a false positive rate of 0.103 for mistakenly misclassified the ham SMS compared with Random Forest with value of 0.206. Based on these evaluation, the highest performance algorithm in overall is classification using Vector Space Model with ANN (VSM+ANN) algorithm. The tradeoff of this method is execution time needed to train the model is slightly higher with additional hidden layer and neuron added.

Table 4: Classification Comparison

	Spam Accuracy (%)	Spam Precision (%)	False Positive Rate (%)	Learning Time (seconds)
Support Vector Machine	87.09	0	NA	0.941
Random Forest	97.13	0.98	0.206	0.176
Artificial Neural Network	99	0.99	0.103	0.272

As for poor performance of SVM, although it gives 87% accuracy it is meaningless as it did not give clear picture as it only correctly classified ham message with none spam are classified; with True Positive and False Positive value of 0 as shown in Table 5. SVM are based on the concept of decision planes or hyperplane that define decision boundaries thus it is not surprising as VSM is having high dimensionality attributes and the word is very closely inseparable between ham and spam.

Table 5: Confusion Table

Algorithm	**Confusion Metrics**		
		Predicted Ham	Predicted Spam
ANN	Actual Ham	970	1
	Actual Spam	8	136
		Predicted Ham	Predicted Spam
SVM	Actual Ham	971	0
	Actual Spam	144	0
		Predicted Ham	Predicted Spam
RF	Actual Ham	969	2
	Actual Spam	30	114

6 Conclusion

In this paper, we present techniques for classifying SMS spam using Vector Space Model with Artificial Neural Network (VSM+ANN). Based on the proposed methods, it shows a promising result with scores for accuracy 99%, precision of value 0.99 and low false positive rate of 0.103. ANN algorithm fused well with VSM and better result can be achieved by adding additional hidden layer and neuron while tuning the hyperparameter.

ACKNOWLEDGEMENTS

The authors would like to thank 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 Sytems Research Group (CPS RG) and funded by Universiti Teknologi Malaysia (GUP Tier 1: Q.K130000.2538.18H4).

References

[1] R. R. Schaller (1997). Moore's law: past, present and future. *IEEE Spectr*., 34(6), pp. 52–59

[2] "Mobile phone users worldwide 2013-2019," Statista, 2015. [Online]. Available: https://www.statista.com/statistics/274774/forecast-of- mobile-phone-users-worldwide/. [Accessed: 02-Jun-2017].

[3] R. Bauer (2016). Mobile Phone Penetration," World Bank Blogs, 23-Mar-2016. [Online]. Available: https://blogs.worldbank.org/category/tags/mobile-phone-penetration. [Accessed: 02-Jun-2017].

[4] "86% of Enterprise Businesses Using MMS Marketing," Tatango - SMS Marketing Software, 23-Feb-2016. [Online]. Available: https://www.tatango.com/blog/86-percent-of-enterprise-businesses-using-mms-marketing/. [Accessed: 03-Jun-2017].

[5] Credence Research (2016). SMS Market Trend And Forecast To 2022, 57791-04–16,

[6] Spam Statistic: March 2017, Australia Communication and Media Authority

[7] I. Joe and H. Shim (2010). An SMS Spam Filtering System Using Support Vector Machine. *in Future Generation Information Technology,* 577–584.

[8] K. Mathew and B. Issac, (2011). Intelligent spam classification for mobile text message. *in Proceedings of 2011 International Conference on Computer Science and Network Technology,* 1, 101–105.

[9] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami (2011). Contributions to the study of SMS spam filtering: new collection and results, *in Proceedings of the 11th ACM symposium on Document engineering*, 259–262.

[10] J. M. B. Serrano, J. H. Palancar, and R. Cumplido (2014). The Evaluation of Ordered Features for SMS Spam Filtering. *in Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 383–390.

[11] T. A. Almeida and J. Gomez (2017). SMS Spam Collection. [Online]. Available: http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/. [Accessed: 24-May-2017].

[12] G. V. Cormack, J. M. G. Hidalgo, and E. P. Sánz (2007). Feature engineering for mobile (SMS) spam filtering. p. 871.

[13] T. Chen and M.-Y. Kan, (2013). Creating a live, public short message service corpus: the NUS SMS corpus. *Language. Resource Evaluation*., 47(2), 299–335

[14] C. Tagg (2009). A corpus linguistics study of SMS text messaging, University of Birmingham, 2009.

[15]G. V. Cormack, J. M. Gómez Hidalgo, and E. P. Sánz (2007). Spam filtering for short messages, 313.

[16] D.K. Harman (1993). The first text retrieval conference (TREC-1). In: US Department of Commerce, *National Institute of Standards and Technology*, 1–20.

[17]Graves, A., Mohamed, A., Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. arXiv:1303.5778 [cs]

[18]Chen, C., Seff, A., Kornhauser, A., Xiao, J. (2015). Deep Driving: Learning Affordance for Direct Perception in Autonomous Driving. IEEE, pp. 2722–2730

[19] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., (2013). A Neural Probabilistic Language Model 19.

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

[21]Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620

[22] He, B., Ounis, I., (2005). Term Frequency Normalisation Tuning for BM25 and DFR Models, in: Losada, D.E., Fernández-Luna, J.M. (Eds.), Advances in Information Retrieval. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 200–214.

[23] Jones, K.S., (1973). Index term weighting. Information Storage and Retrieval 9, 619–633

[24] Jesan, J.P., Lauro, D.M., (2003). Human Brain and Neural Network Behavior: A Comparison. Ubiquity 2003, 2–2.

[25] Haykin, S.S., (2009). Neural networks and learning machines. Pearson, Upper Saddle River, NJ, USA

[26] Samarasinghe, S., (2007). Neural networks for applied sciences and engineering. From fundamentals to complex pattern recognition.

[27] Ruder, S., 2016. An overview of gradient descent optimization algorithms. arXiv:1609.04747 [cs].

[28]Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs].