Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 161 (2019) 509–515

Procedia Computer Science

www.elsevier.com/locate/procedia

The Fifth Information Systems International Conference 2019

SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm

Nilam Nur Amir Sjarif*, Nurulhuda Firdaus Mohd Azmi, Suriayati Chuprat, Haslina Md Sarkan, Yazriwati Yahya, Suriani Mohd Sam

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia Kuala Lumpur, Level 5, Menara Razak, 54100 Jalan Sultan Yahya Petra, Kuala Lumpur, Malaysia

Abstract

The daily traffic of Short Message Service (SMS) keeps increasing. As a result, it leads to dramatic increase in mobile attacks such as spammers who plague the service with spam messages sent to the groups of recipients. Mobile spams are a growing problem as the number of spams keep increasing day by day even with the filtering systems. Spams are defined as unsolicited bulk messages in various forms such as unwanted advertisements, credit opportunities or fake lottery winner notifications. Spam classification has become more challenging due to complexities of the messages imposed by spammers. Hence, various methods have been developed in order to filter spams. In this study, methods of term frequency-inverse document frequency (TF-IDF) and Random Forest Algorithm will be applied on SMS spam message data collection. Based on the experiment, Random Forest algorithm outperforms other algorithms with an accuracy of 97.50%.

Keywords: Short Message Service; Spam; TF-IDF; Random Forest

1. Introduction

The increasing mobile phones become one of the attached companions for many individuals. With the explosive penetration of mobile devices and millions of people sending messages every day, Short Message Service (SMS) has

* Corresponding author. Tel.: +60-32-203-1408.
E-mail address: nilamnur@utm.my

1877-0509
10.1016/j.procs.2019.11.150

become a multi-million-dollar commercial industry with a value between 11.3 to 24.7 percent of the developing countries' Gross National Income (GNI) in the early year of 2013 [1]. However, the downside of the increase mobile users and the cheap SMS text messages is that mobile phones are attracting more unsolicited bulk messages especially in the form of advertisements. Compared to these unsolicited messages in SMS, spams commonly plague e-mails as 90 percent of the e-mails are spams in 2010 [2]. Although SMS spams are not as common as electronic junk mails, they still manage to irritate mobile phone users while creating societal frictions to mobile phone devices [3]. Regardless, the number of mobile phone spams plaguing users can be different across regions. Based on Aski et al [4], SMS spam messages contribute to less than 1 percent out of text messages in North American region in the year of 2010. On the other hand, 30 percent out of the SMS messages are spams sent by spammers in Asian region. The popularity of SMS among many users leads to a drop in SMS charge, with below than US$0.001 in China's market while some telco providers do not charge the users [5].

SMS can be defined as text communication platform across mobile devices or fixed lines that permits their users to exchange short text messages using standardized communication protocols [2, 6]. While, spams can be defined as "unwanted electronic mail" [4]. Spams are undesirable but still exist in our messages. SMS spams or mobile phone spams are junk mails delivered across mobile devices in the form of text messages [6]. They are usually sent by spammers to intend a group of recipients by bulk. These spams usually sent by businesses taking advantages of receivers to advertise and promote their products or services. Besides promoting materials, spams also can threaten users' privacy with phishing, fraud and identity theft attacks through text messages [7]. Spams can originate from any country in the world, with China topping other countries as the top source of spams [8]. This shows that spammers do not refrain themselves from operating within their borders since some countries do little in preventing these spammers from spreading spams. Any individual can buy any mobile number from different area codes to spam mobile phone users; hence, they are hardly being identified and caught [9].

Recently in the research community, the trend of classification using machine learning become popular. As SMS message 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 Message spam characteristic [10]. Example of machine learning algorithm such as Random forest algorithm received more attention by the researcher due to the ability of the algorithm to boost performance and increase the accuracy [11]. In addition, for feature extraction approach, Term Frequency-Inverse document frequency (TFIDF) commonly used. This approach is based on the often-weighting, particularly in IR domain including text mining [10, 12].

Therefore, the objective of this paper is to solve the problem of SMS spam message detection using term frequency-inverse document frequency (TF-IDF) and Random Forest Algorithm. 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, while Section 4 presents the results and discussion on the finding. Finally, Section 5 concludes this paper.

2. Literature review

2.1. Related work

Since there is similarity between text documents in SMS spam and e-mail spam, the content-based technologies used in e-mail spam filtering can be adopted to combat the spreading of mobile phone spams [13,14]. Nonetheless, SMS spams still exhibit different properties compared to e-mail spams. Compared to e-mail spams, SMS spams are limited in characters since the standard text messages is limited to 140 bytes (characters) [15]. Hence, the number of features that can be utilized in SMS spam classification is much smaller compares to features in e-mail spams. Often mobile phone spams are written in less formal language or abbreviations or idioms due to restrictions. Nonetheless, SMS spams irritate users like e-mails, if not more, since the receiver might have to bear the cost of getting that SMS [2, 6, 7, 16]. As a result, the receivers of the text messages might experience financial lost due to spams. Regardless, spams in both SMS and e-mails are sent with the interest of either business, tricks or fraudulent activities. Yet, the format used in SMS spams are different such that the amount and format of information are restricted due to limited number of characters (e.g. alphabets, number and symbols). Compared to e-mail spams, SMS spams have no headers or headlines on top of these limited features shown in SMS text messages [2].

Some researchers have been done in the field of spam filtering methods and measures. Chan et al. [17], for instance, builds spam filtering methods based on good word attack strategy and feature reweighting method. The methods rely on the limited characters and short text messages on top of the weight values which are evaluated against to datasets (i.e. SMS and comment dataset). A good word attack strategy will mislead the classifier's output with the least number of inserted characters. The authors also introduce a feature reweighting method in which a novel rescaling function is proposed to minimize the significance of the feature characterizing a short word. This method is performed to rescale the weights and increase the linear classifier's robustness against a good word attack strategy.

Sethi et al. [18] compare different machine learning algorithms to filter and detect SMS spam messages using three information i.e. the raw text messages, the length of the messages and information gain matrix. The algorithms that have been used in the experiment are Naïve Bayes, Random Forest and Logistic Regression. Meanwhile, Mujtaba and Yasin [7] utilize four features derived from SMS text messages in order to be trained by machine learning algorithms. These features include the message size, frequently occurring monograms in the text messages, frequently occurring diagrams and class of messages (i.e. ham and spam as 0 and 1 respectively). The authors find that Naïve Bayes algorithm outperforms the other classifiers used in the study.

Choudhary and Jain [9] explore and analyze different features for SMS spam classification. They have extracted numerous features from SMS text messages including presence of mathematical symbols, special symbols, emotions among many others. In the study, they explore the characteristics and behaviors of SMS spam messages in depth for message classification with successful result in return.

Xu et al. [13] on the other hand propose a SMS spam filtering method using non-content features. Instead of using the content of SMS text messages as features, they use static features (i.e. number of messages and message size), temporal features (i.e. number of messages sent in one day, size of messages in one day and time of day) and network features (i.e. number of recipients and clustering coefficients) in the study. These features will then be experimented with Support Vector Machine algorithm and K-Nearest Neighbors algorithm. They find that by incorporating network and temporal features into conventional static features, the method gives better performance in detecting spammers.

Warade et al. [19] develop a spam detection system based on relationship between sender and receiver and the message contents. The study determines a text message as spam if the senders and receivers share no mutual relations while the SMS exhibits the content of the spams. The message will then be automatically transferred into the spam box. Meanwhile, legitimate SMS text messages will be sent to receiver's mobile inbox if there is a mutual relation between senders and receivers with no spamming contents visible. The relationship between senders and receivers will be examined through the inspection of SMS logs and direct relation between the two. Safie et al. [10] propose SMS spam classification using vector space and Artificial Neural Network (ANN) algorithm. The result shows a significant improvement based on the accuracy. Through this study, the author proved that SMS spam message classification can works well using sequence of features.

3. Propose method

In this paper, the proposed study involves process as shown in Fig.1 below. The main objective is to classify SMS text messages either as ham or spam.

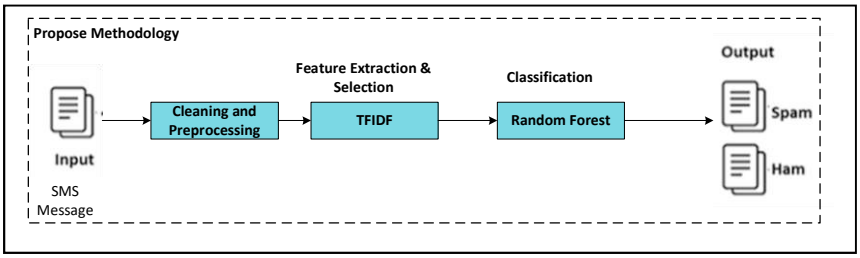

Fig. 1. Process of SMS Spam Detection.

3.1. Dataset

The public dataset of SMS labelled messages is obtained from UCI Machine Learning Repository. The data is originally collected by Almeida and Hidalgo [20]. It contains 5,574 English raw text messages with tag labels either as legitimate (ham) or spam. The text messages have been collected and derived from various sources such as UK forum from Grumbletext website, NUS SMS Corpus (NSC), Caroline Tag's PhD Thesis and SMS Spam Corpus v.0.1 Big. Table 1 shows the source of the dataset

Table 1. Source of dataset.

Spam Dataset	Total
Grumbletext website	425(spam)
NUS SMS Corpus (NSC)	3,375 (ham)
Caroline Tag's PhD Thesis	450 (ham)
Corpus v.0.1 Big	1,002 (ham) 322 (spam)

From Table 1 above, this study finds that there are only 5,574 labelled messages in the dataset, with 4827 of messages belong to ham messages while the other 747 messages belong to spam messages. Nonetheless, this dataset consists of two named columns starting with the message labels (ham or spam) followed by strings of text messages and three unnamed columns.

3.2. SMS spam detection phases

The phases of detection of spam involve preprocessing, feature extraction and selection and classification.

3.2.1. Preprocessing

Pre-processing is the first stage in which the unstructured data is converted into more structured data. Since keywords in SMS text messages are prone to be replaced by symbols. In this study, the stop word list remover for English language have been applied to eliminate the stop words in the SMS text messages. Fig. 2(a) shows the frequencies of words in SMS messages, while Fig. 2(b) shows the most frequent words used in spam text messages are from pronoun (e.g. to) and proposition (e.g. your) groups. Similarly, the top words in ham text messages are occupied by either pronoun, proposition among many other types of stop words.

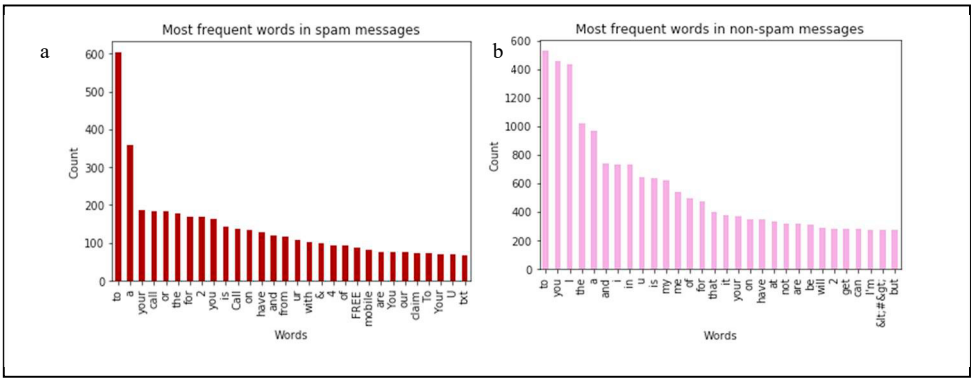

Fig. 2. (a) Most frequent words in SMS spam messages; (b) Most frequent words in SMS legitimate messages.

3.2.2. Feature extraction and selection

Feature extraction and selection is important for the discrimination of ham and spam in SMS text messages. For this phases TFIDF will be used. 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 occurrences a word appears in the document and usually normalized in positive quadrant between 0 and 1 to eliminate bias towards lengthy documents [10]. To construct the index of terms in TFIDF, punctuation is removed, and all text are lowercase during tokenization. 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 calculated on how infrequent a word or term is in the documents [10]. 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 stop words or common words that is generic. Hence only infrequent words in contrast of the entire dataset is relevant for that documents. TF-IDF does not only assess the importance of words in the documents but it also evaluates the importance of words in document database or corpus. In this sense, the word frequency in the document will increase the weight of words proportionally but will then be offset by corpus's word frequency [21]. This key characteristic of TF-IDF assumes that there are several words that appear more often compared to others in the document in general. Hence, the relevancy of a word to a document is shown in Eq. 1.

$$F - IDF = \frac{Ter \quad Frequency}{Document\ Frequency} \tag{1}$$

3.2.3. SMS message spam classification

Random Forest (RF) algorithm will used for classification of ham or spam during this phase. RF is averaging ensemble learning method that can be used for classification problem. This algorithm combines various decision tree models in order to eliminate the overfitting problem in decision trees [14]. In RF algorithm, each tree is capable in providing its own prediction results, different from each other. As a result, each tree gives different performances, in which the average of their performances will be generalized and calculated. During the training phase, a set of decision trees will be constructed before they can operate on randomly selected features [9]. Regardless, RF can work well with a large dataset with a variety of feature types, similar to binary, categorical and numerical. The algorithm works as follows (see Fig. 3): for each tree in the forest, a bootstrap sample is selected from S where $S(i)$ represents the ith bootstrap. A decision-tree is then learn using a modified decision-tree learning algorithm. The algorithm is modified as follows: at each node of the tree, instead of examining all possible feature-splits, some subset of the features text $f \subseteq F$ is selected randomly. where F is the set of Spam features. The node then splits on the best feature in f rather than F. In practice f is much, much smaller than F. Deciding on which feature to split is oftentimes the most computationally expensive aspect of decision tree learning. By narrowing the set of features, the speed up the learning of the tree is increase drastically.

```
Pseudocode: Random Forest
Precondition: A training set S := (x1, y1), ..., (xn, yn), feature F,
and number of trees in forest B
function RandomForest (S,F)
H ← ∅
for i ∈ 1, ..., B do
    S(i) ← A bootstrap sample from S
    hi ← RandomizeTreeLearn (S(i), F)
    h ← H ∪ {hi}
  end for
  return H
end function
function RandomizeTreeLearn (S,F)
  At each node:
    f ← very small subset of F
    Split on best feature in f
  return The learned tree
end function
```

Fig. 3. Random Forest Algorithm.

4. Result and Discussion

To compare the performance of algorithms in the experiment, this paper employs the performance evaluation measures such as accuracy, precision and f-measure. Table 2 shows the summary of performance of algorithms.

Table 2. Comparison Algorithm.

Algorithm	Accuracy	Precision	F-measure
TF-IDF+ Multinomial Naïve Bayes (MNB)	97.06	0.97	0.97
TF-IDF+ K-Nearest Neighbor (KNN)	91.19	0.92	0.89
TF-IDF+ Support Vector Machine (SVM)	87.49	0.77	0.82
TF-IDF+ Decision Tree (DT)	96.57	0.96	0.97
TF-IDF + Random Forest	97.50	0.98	0.97

It is notable that TF-IDF+RF achieves the best performance and outperforms the other evaluated algorithms in this experiment in terms of accuracy percentage. It accurately classifies 97.50% while it achieves 0.98 in precision while F-measure is 0.97 in both recall and f-measure. The second-best algorithm in terms of accuracy percentage is TF-IDF+MNB with 97.06%, and the result of precision is slightly different with only 0.1 with the propose method. Surprisingly, for the result for the F-measure is same with the propose method with the rate 0.97. The lowest performance is determined using the TF-IDF+SVM with the result accuracy is 87.49, precision 0.77 and F-measure is 0.82. The reason behind this result might be due to the fact that SVM cannot handle imbalanced dataset very well. The imbalanced data might cause the performance loss in several ways such as the imbalances ratio of positive and negative support vectors or the position of positive points being far from the ideal boundary. Nonetheless, SVM still manages to get more than 75% for overall performance.

5. Conclusion

The SMS spam message problem is plaguing almost every country and keeps increasing without a sign of slowing down as the number of mobile users increase in addition to cheap rates of SMS services. Therefore, this paper presents the spam filtering technique using various machine learning algorithms. Based on the experiment, TF-IDF with Random Forest classification algorithm outperforms good compare to other algorithms in terms of accuracy percentage. However, it is not enough to evaluate the performance based on the accuracy alone since the dataset is imbalanced; therefore, the precision, recall and f-measure of the algorithms must also be observed. After some examinations, RF algorithm still manages to provide good precision and f-measure with 0.98 of precision while 0.97 for f-measure. Different algorithms will provide different performances and results based on the features used. For future works, adding more features such as message lengths might help the classifiers to train data better and give better performance.

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 Razak Faculty of Technology and Informatics), under Cyberphysical Sytems Research Group (CPS RG) and funded by Universiti Teknologi Malaysia (GUP Tier 1: Q.K130000.2538.18H42).

References

[1] Modupe, A., O. O. Olugbara, and S. O. Ojo. (2014) "Filtering of Mobile Short Messaging Communication Using Latent Dirichlet Allocation with Social Network Analysis", *in Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2013, G.-C. Yang, S.-I. Ao, and L. Gelman, Eds. Springer Science & Business*. pp. 671–686.

[2] Shirani-Mehr, H. (2013) "SMS Spam Detection using Machine Learning Approach." p. 4.

[3] Abdulhamid, S. M. et al., (2017) "A Review on Mobile SMS Spam Filtering Techniques." *IEEE Access* **5**: 15650–15666.

[4] Aski, A. S., and N. K. Sourati. (2016) "Proposed Efficient Algorithm to Filter Spam Using Machine Learning Techniques." *Pac. Sci. Rev. Nat. Sci. Eng.* **18 (2)**:145–149.

[5] Narayan, A., and P. Saxena. (2013) "The Curse of 140 Characters: Evaluating The Efficacy of SMS Spam Detection on Android." p. 33–42.

[6] Almeida, T. A., J. M. Gómez, and A. Yamakami. (2011) "Contributions to the Study of SMS Spam Filtering: New Collection and Results." p. 4.

[7] Mujtaba, D. G., and M. Yasin. (2014) "SMS Spam Detection Using Simple Message Content Features." *J. Basic Appl. Sci. Res.* **4 (4)**: 5.

[8] Gudkova, D., M. Vergelis, T. Shcherbakova, and N. Demidova. (2017) "Spam and Phishing in Q3 2017." *Securelist - Kaspersky Lab's Cyberthreat Research and Reports.* Available from: https://securelist.com/spam-and-phishing-in-q3-2017/82901/. [Accessed: 10th April 2018].

[9] Choudhary, N., and A. K. Jain. (2017) "Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique", in *Advanced Informatics for Computing Research* **712**: 18-30.

[10] Safie, W., N.N.A. Sjarif, N.F.M. Azmi, S.S. Yuhaniz, R.C. Mohd, and S.Y. Yusof. (2018) "SMS Spam Classification using Vector Space Model and Artificial Neural Network." *International Journal of Advances in Soft Computing & Its Applications* **10 (3**): 129-141.

[11] Fawagreh, Khaled, Mohamed Medhat Gaber, and Eyad Elyan. (2014) "Random Forests: From Early Developments to Recent Advancements, Systems Science & Control Engineering." *An Open Access Journal* **2 (1)**: 602-609.

[12] Sajedi, H., G. Z. Parast, and F. Akbari. (2016) "SMS Spam Filtering Using Machine Learning Techniques: A Survey." *Machine Learning,* **1 (1)**: 14.

[13] Q. Xu, E., W. Xiang, Q. Yang, J. Du, and J. Zhong. (2012) "SMS Spam Detection Using Noncontent Features." *IEEE Intell. Syst*. **27(6)**: 44–51.

[14] Sethi, G., and V. Bhootna. (2014) *SMS Spam Filtering Application Using Android*.

[15] Nagwani, N. K. (2017) "A Bi-Level Text Classification Approach for SMS Spam Filtering and Identifying Priority Messages." **14 (4)**: 8.

[16] Delany, S. J., M. Buckley, and D. Greene. (2012) "SMS Spam Filtering: Methods and Data," *Expert Syst. Appl.* **39(10)**: 9899–9908.

[17] Chan, P. P. K., C. Yang, D. S. Yeung, and W. W. Y. Ng. (2015) "Spam Filtering for Short Messages in Adversarial Environment." *Neurocomputing* **155**: 167–176.

[18] Sethi, P., V. Bhandari, and B. Kohli. (2017) "SMS Spam Detection and Comparison of Various Machine Learning Algorithms", in *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN).* pp. 28–31.

[19] Warade, S. J., P. A. Tijare, and S. N. Sawalkar. (2014) "An Approach for SMS Spam Detection." *Int. J. Res. Advent Technol*. **2 (2)**: 4.

[20] Almeida. T. A., and J. M. G. Hidalgo. (2018) "SMS Spam Collection." Available from: http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/. [Accessed: 11st April 2018].

[21] Wang, Y., Z. Zhou, S. Jin, D. Liu, and M. Lu. (2017) "Comparisons and Selections of Features and Classifiers for Short Text Classification." *IOP Conf. Ser. Mater. Sci. Eng.* **261**: 012018.