# A New SMS Spam Detection Method Using Both Content-Based and Non Content-Based Features

**Nurul Fadhilah Sulaiman and Mohd Zalisham Jali**

**Abstract** SMS spamming is an activity of sending 'unwanted messages' through text messaging or other communication services; normally using mobile phones. Nowadays there are many methods for SMS spam detection, ranging from the list-based, statistical algorithm, IP-based and using machine learning. However, an optimum method for SMS spam detection is difficult to find due to issues of SMS length, battery and memory performances. Hoping to minimize the aforementioned problems, this paper introduces another detection variance that is based on common characters used when sending SMS (i.e. numbers and symbols), SMS length and keywords. To verify our work, the proposed features were stipulated into five different algorithms and then, tested with three different datasets for their ability to detect spam. From the conduct of experiments, it can be suggested that these three features are reasonable to be used for detecting SMS spam as it produced positive results. In the future, it is anticipated that the proposed algorithm will perform better when combined with machine learning techniques.

## 1 Introduction

Spam messages are unsolicited messages sent by spammers to known and unknown users for various purposes. Among others are for fraud, advertisement and phishing. Spam messages can occur in many forms; emails, SMSs and social network platforms. This paper focuses on the activity of spamming using SMS or text messages as the increasing usage of SMS to communicate nowadays provides huge opportunities for spammers to do their job. SMS different with email in several aspects

N.F. Sulaiman · M.Z. Jali (✉)
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM),
Bandar Baru Nilai, 71800 Nilai, Negeri Sembilan, Malaysia
e-mail: zalisham@usim.edu.my

N.F. Sulaiman
e-mail: fadhilahsulaiman90@gmail.com



H.A. Sulaiman et al. (eds.), *Advanced Computer and Communication Engineering Technology*, Lecture Notes in Electrical Engineering 362,
DOI 10.1007/978-3-319-24584-3_43

i.e. it does not contain a mailing list of recipients and the maximum words for messages is up to 160 characters [1].

To the best of researchers' knowledge, the detection of spam messages can be done through content-based features [2], non-content features [3] and machine learning classifiers [4]. Many researchers study techniques and methods for filtering spam messages in email [5–11] but less publication reported on the phone.

The common way to detect SMS spam is using 'keywords of spam'. This approach however resulted in using more space and memory; as well as processing would take time due to the larger set of 'keywords of spam'. In this paper, we focus on detecting spam messages of SMS by combining both content-based and non-content based features. As such, three features are proposed; namely based on the length of messages that is more than 100 characters, messages contains special characters such as symbols and numbers as well as keywords of spam messages itself and our justification of using these are highlighted as following:

1. *Length of message* (i.e. greater than 100 characters).
   From our finding analysing SMS spam datasets, we found SMS spams tend to be longer in size. Besides, SMS spam normally use standard and formal language in order to attract users so that they can understand and put their interest to those particular messages. Therefore, in our work, we assume messages that are more than 100 in length could potentially be classified as spam.
2. *Special Characters* (i.e. numbers and symbols).
   From our study, we can reveal that the usage of special characters do exist and common in SMS spams. For instance, spammers prefer using numbers or digits such like phone number, code number to claim, service code and sum of money (i.e. 300 pounds). In addition to this, special characters or symbols such as ****, $, XXX are commonly found to be used in SMS spams.
3. *Keywords of spam and ham.*
   From our analysis, words that being used in spam messages are common and similar across platforms. Thereby, our approach also considers common keywords for detecting spam. Examples of keywords are like—'Free', 'Call' and 'Claim'. Besides, ham keywords are also investigated and use in our detection algorithms.

By detecting SMS spam using their lengths, used characters and keywords, we postulate it will make the detection rate higher; and at the same time minimizing device resources. To achieve what we have claimed for, three aforementioned detection features were stipulated into five different algorithms; with validation of the used datasets were done using four well-known machine-learning classifiers.

The rest of this paper is organised as follows. Section 2 provides an overview about existing methods and techniques for SMS spam detection. Section 3 then describes our methodology with Sect. 4 presents results and discussions from our conduct of testing. The paper concludes with an outlook for future works.

## 2 Existing Research Related to Spam Detection

Much research has been put to investigate spam messages especially on the email platform. For example, Pour et al. [11] discussed three techniques for email spam detection; namely *list-based*, *statistical algorithm* and *IP-based*. The *list-based* technique is classified into three categories, namely Blacklist, Whitelist and Greylist. Blacklist blocks the IP address based on complaints from recipient [12], with Greylist rejects mail from unknown sources on the theory that real mailers will retry the mail and spammers won't [13]. The *statistical algorithm* can be categorised into content-based method and rule-based method. Content-based method is commonly used and it filters the content of mail body and headers. It uses machine learning which need to be trained [9]. For example, Chakraborty and Mondal [10] applied different decision tree classifiers to filter spam mail while Amayri & Bouguila [14] used Support vector Machine (SVM) for spam filtering. Rule based method works through certain rules and these rules will decide to pass or block the email [9]. Reverse lookup is an example of method in the *IP-based* technique. It is a method of resolving an IP address into a domain name [15].

Development of technology in mobile phone with sophistication of new invention such as smartphones and tablets lead spammers to change their target to mobile users. This is due to the multifunction of mobile phone to access corporate network and data. The authorities and researchers have to take action to reduce and control SMS spams that are rising year by year. Generally, SMS spams can be detected by examining and reviewing message contents (i.e. *content features*) or the way messages are sent (i.e. *non-content features*). Sohn et al. [2] proposed a method of using stylistic information to the content-based mobile spam filtering. They focus on the way the SMS is written (i.e. stylistic aspect) and four features of stylistic were used—length of messages, function word frequencies, part-of-speech n grams and special character. Tan et al. [16] identified features of SMS spams based on word grams, character grams, alphanumeric and non-alphanumeric characters. They also tested a number of statistical features such as message length, proportion of upper-case letters and proportion of punctuation. Other study is by Mujtaba and Yasin [1], where they used four features—size of the message, the existence of frequently occurring monograms in the messages, existence of frequently occurring diagram in the message and messages class. These features are implemented and trained in machine learning algorithm for better accuracy.

Xu et al. [3] focus on the non-content features such as statistical features, temporal features and network features. Their study showed that the temporal features and network features were more effective as compared to the statistic features. Mosquera et al. [17] analyzed the effectiveness of machine learning filters based on linguistic and behavioral patterns. Shahi and Yadav [4] conducted experiments to compare the performance of machine learning classifiers to detect SMS spam in Nepali language and Bilal and Farooq [18] compared four types of evolutionary learning classifiers to filter SMS spams in order to obtain the best classifiers. Other study that conducted in Nuruzzaman et al. [19] used the same approach.

## 3 Algorithm and Methodology

From the review of existing works in the Sect. 2, it can be suggested that SMS spam research is still needed and provide wide opportunity for improvement. Having said, this paper proposes another variation for detecting SMS spams using the combination of content and non-content features. For testing purposes, the proposed features are expanded into five (5) different algorithms. The idea of expanding features into five (5) variations of algorithms is to investigate their performance, with the ultimate aim to find the best combination for optimum detection.

The first three algorithms use only two aforementioned features; with the remaining algorithms combine all features. Specifically, Algorithm 1 uses only keywords, Algorithm 2 uses the combination of message length and keywords, with Algorithm 3 uses the features of special characters and keywords respectively. Algorithm 4 and 5 use message length, special characters and keywords. Figure 1 presents pseudo code of the proposed method and Table 1 shows the different level of features used in each algorithm.

Two main phases involved in our experiments conduct. In the phase 1, validation of datasets was conducted. To do this, four (4) data mining classifiers in WEKA (i.e. Naïve Bayes (NB), Support vector machine (SVM), k-Nearest Neighbor (k-NN) and Decision Tree (DT)) were used. The used datasets named UCI Machine Learning (UCI) [20], British English SMS Corpora (BEC) [21] and

```
Input; M: SMS messages
Output; H: Ham SMS or S: Spam SMS

Begin (while TRUE)
     Read all SMS messages, M
     Detect S (i.e. using five algorithms as in Table 1)

     Copy to H or S folder
End
```

**Fig. 1** Generalized pseudo-code for five algorithms

**Table 1** Different features in each Algorithm

| Detection of ham/spam | | | | |
|---|---|---|---|---|
| Algorithm | Phase 1 | Phase 2 | Phase 3 | Phase 4 |
| 1 | Spam keywords | Ham keywords | (NA) | (NA) |
| 2 | Length + Spam keywords | Ham keywords | (NA) | (NA) |
| 3 | Number/digit + Spam keywords | Ham keywords | (NA) | (NA) |
| 4 | Length | Number/digit + Spam keywords | Ham keywords | (NA) |
| 5 | Length | Number/digit + Spam keywords | Ham keywords | Spam keywords |

**Table 2** Characteristics of used datasets

| | UCI | BEC | DIT |
|---|---|---|---|
| Ham | 4825 | 450 | – |
| Spam | 747 | 425 | 1353 |
| Total | 5572 | 875 | 1353 |

Dublin Institute of Technology (DIT) [22] respectively. Table 2 details out different numbers of spam and ham messages in each dataset.

In the phase 2, testing on the proposed algorithms was conducted. Pre-processing or cleaning process for the dataset is not done because characters (i.e. numbers and symbols) in these messages may help with the detection process. Results for experiments conduct for both phase 1 and 2 were analyzed, compared and reported in the next section.

The measurement of detection performance is based on Correctly Classified Messages, True Positive and True Negative.

- Correctly Classified (CC): SMS is correctly classified as spam or ham message. The rate for accuracy of a classifier or algorithm is determined by the formula as below.

$$Accuracy = (TP + TN)/(TP + FN + TN + FP)$$

- True Positive (TP): SMS messages are correctly classified as ham messages
- True Negative (TN): SMS messages are correctly classified as spam messages.
- False Positive (FP) & False Negative (FN).

## 4 Results and Discussions

Results are reported and discussed in three parts. Part 1 discusses the results for dataset validation; with part 2 and 3 reported the performance detection of the proposed algorithms.

### *4.1 Part 1: Dataset Validation*

Table 3 presents results of experiment conduct for validating datasets. From the Table 3, it can be found that the k-NN classifier produced the best detection rate where it managed to detect the highest number of spam and ham messages and is it also found that all classifiers managed to detect spam messages (TN) of the DIT dataset. As majority of chosen classifiers perform detection of ham and spam messages in a ‘good’ manner, it can be suggested that the chosen datasets are appropriate to be used for testing the proposed algorithms.

**Table 3** Results in WEKA using four classifiers

| Results in WEKA | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dataset | Classifier | Correctly classified (%) | Incorrectly classified (%) | TP | FN | TN | FP |
| UCI | NB | 5306 = 95.23 | 266 = 4.77 | 4646 | 87 | **660** | 179 |
| | SVM | 5537 = 99.37 | 35 = 0.63 | 4825 | 35 | **712** | 0 |
| | k-NN | 5564 = 99.86 | 8 = 0.14 | 4825 | 8 | **739** | 0 |
| | DT | 5416 = 97.20 | 156 = 2.80 | 4807 | 138 | **609** | 18 |
| BEC | NB | 804 = 91.89 | 71 = 8.11 | 413 | 34 | **391** | 37 |
| | SVM | 868 = 99.2 | 7 = 0.8 | 450 | 7 | **418** | 0 |
| | k-NN | 872 = 99.66 | 3 = 0.34 | 450 | 3 | **422** | 0 |
| | DT | 750 = 85.71 | 125 = 14.29 | 434 | 109 | **316** | 16 |
| DIT | NB | 1353 = 100 | 0 = 0 | 0 | 0 | **1353** | 0 |
| | SVM | 1353 = 100 | 0 = 0 | 0 | 0 | **1353** | 0 |
| | k-NN | 1353 = 100 | 0 = 0 | 0 | 0 | **1353** | 0 |
| | DT | 1353 = 100 | 0 = 0 | 0 | 0 | **1353** | 0 |

Bold represents the actual number of spam messages detected

## *4.2 Part 2: Algorithm 1–3 Results*

Results for Algorithm 1–3 towards three different datasets are shown in Table 4. From the Table 4, it can be stated that Algorithm 1 managed to detect more spam messages (TN) as compared to other algorithms. For correctly classified messages

**Table 4** Results for algorithm 1, 2 and 3

| Simulation results | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Algo | Detection feature(s) | Correctly classified (%) | Incorrectly classified (%) | TP | FN | TN | FP |
| UCI | 1 | Keywords | 4228 = 75.88 | 1344 = 24.12 | 3501 | 1324 | **727** | 20 |
| | 2 | Length and keywords | 4853 = 87.10 | 719 = 12.90 | 4178 | 647 | **675** | 72 |
| | 3 | Characters and keywords | 5155 = 92.52 | 417 = 7.48 | 4454 | 371 | **701** | 46 |
| BEC | 1 | Keywords | 766 = 87.54 | 109 = 12.46 | 357 | 93 | **409** | 16 |
| | 2 | Length and keywords | 788 = 90.06 | 87 = 9.94 | 409 | 41 | **379** | 46 |
| | 3 | Characters and keywords | 812 = 92.8 | 63 = 7.2 | 425 | 25 | **387** | 38 |
| DIT | 1 | Keywords | 1294 = 95.64 | 59 = 4.36 | 0 | 0 | **1294** | 59 |
| | 2 | Length and keywords | 1135 = 83.89 | 218 = 16.11 | 0 | 0 | **1135** | 218 |
| | 3 | Characters and keywords | 1221 = 90.24 | 132 = 9.76 | 0 | 0 | **1221** | 132 |

Bold represents the actual number of spam messages detected

into ham and spam, Algorithm 3 is more accurate for datasets of UCI and BEC while Algorithm 1 is more accurate for DIT dataset.

## 4.3 Part 3: Algorithm 4 and 5 Results

Algorithm 4 and 5 combine all three (3) proposed features and results are shown as in the Table 5. Table 5 indicates that Algorithm for 4 and 5 perform well in detecting spam messages (TN) for all three datasets, with Algorithm 5 performed better as compared to Algorithm 4. In term of accuracy classifying messages into ham and spam, Algorithm 5 produced more accurate detection for datasets of UCI and DIT.

## 4.4 Discussions

From the conduct of experiments, it was found that the detection feature that is based on keywords (i.e. spam and ham) is still producing good detection results. For all five (5) algorithms, Algorithm 1 (i.e. using our own keywords of spam and ham) performs well and better in detecting spam messages. Here, we argue that in order to have an optimum detection results, that method needs to have a 'sound' list of keywords. With respect to the clients' mobile environment, it is preferable to have a

**Table 5** Results for algorithm 4 and 5

| Simulation Results | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Algo | Detection feature(s) | Correctly classified (%) | Incorrectly classified (%) | TP | FN | TN | FP |
| UCI | 4 | Length, character and keywords | 5322 = 95.51 | 250 = 4.49 | 4684 | 141 | **638** | 109 |
| | 5 | Length, character and keywords (2x) | 5357 = 96.14 | 215 = 3.86 | 4670 | 155 | **687** | 60 |
| BEC | 4 | Length, character and keywords | 788 = 90.06 | 87 = 9.94 | 440 | 10 | **348** | 77 |
| | 5 | Length, character and keywords (2x) | 725 = 82.86 | 58 = 6.63 | 440 | 10 | **377** | 48 |
| DIT | 4 | Length, character and keywords | 1063 = 78.57 | 290 = 21.43 | 0 | 0 | **1063** | 290 |
| | 5 | Length, character and keywords (2x) | 1177 = 87.00 | 176 = 13.00 | 0 | 0 | **1177** | 176 |

Bold represents the actual number of spam messages detected

minimum list of keywords due to their limitation. On the other hand, if it meant to be implemented in servers' environment, it should be working perfectly. In addition, detection using keywords is sometimes 'unscrupulous' due to various language styles.

The number of messages in each dataset can also affect the detection rate as different datasets may have different number of spams and hams and also contains different message structures. In term of accuracy, results suggest that Algorithm 5 produced high accuracy for UCI dataset as compared to others but Algorithm 3 performed better for the BEC dataset albeit using only two of the proposed features. The similar condition is occurred for the DIT dataset. In this dataset, Algorithm 1 produced higher accuracy as compared to others although this algorithm used only one feature; which is keywords. We anticipate different detection results are due to the 'immature' of the algorithms and there is a need for improvement. When it combined with the machine learning classifier, we expect it to improve and perform better.

We reported that our model produced a 'fair' detection results and we expect this is caused by two conditions. First is due to the way we do our detection. Specifically, the algorithms work on a 'phase-by-phase' basis and they were non-iterative (i.e. static detection). Using static detection might limit the detected results, as messages that are less than 100 of length were not detected at the first stage (Algorithm 2, 4 and 5). Second is due to the nature of the messages itself, as not all messages contain special characters (Algorithm 3). However, from our second validation, we can confirm that our algorithms managed to detect all spam messages that contain 'spam' special characters, and greater than 100 in length.

## 5 Conclusions

In this paper, three features that based upon SMS length, special characters and keywords are discussed to detect SMS spams. These features are a combination of content and non-content of the messages. These features were tested using three well-known datasets and from our conduct of experiments, it can be suggested that our preliminary algorithms managed to detect spam messages that contains spam keywords, spam special characters and length. In summary we can suggest that, although different datasets could result in different detection rate, but we expect it will produce better detection rate if combined together into single algorithm.

For the future, we will cover our performance evaluations, which includes ROC curve, F-measure and effect towards battery and processing performances. In addition, we also plan to implement dynamic detection rather than static detection as practiced in the present paper. And finally, the proposed algorithms will be applied into machine learning in order to inject the element of 'learning' for optimum detection.

**Acknowledgments** Authors wish to thank the Ministry of Education (MOE), Malaysia for funding this research. The grant named RAGS with the grant code USIM/RAGS/FST/STH/36/50913.

## References

1. Mujtaba, G., Yasin, M.: SMS spam detection using simple message content features. J. Basic Appl. Sci. Res. **4**, 275–279 (2014)
2. Sohn, D-N., Lee, J-T., Rim, H-C.: The contribution of stylistic information to content-based mobile spam filtering. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 321–324 (2009)
3. Xu, Q., Xiang, E.W., Yang, Q.: SMS spam detection using non-content features. IEEE Intell. Syst. (2012)
4. Shahi, T.B., Yadav, A.: Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine. Int. J. Intell. Sci. **4**, 24–28 (2014)
5. Zhou, B., Yao Y., Luo J.: A three-way decision approach to email spam filtering. AI'10 Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence, Vol. 6085, 28–39 (2010)
6. Nazirova, S.: Survey on spam filtering techniques. Commun. Netw. J. **3**, 153–160 (2011)
7. Mohammad, N.T.: A fuzzy clustering approach to filter spam e-mail. In: Proceedings of the World Congress on Engineering, vol. 1, 945 (2011)
8. Dasgupta, A., Gurevich, M., Punera, K.: Enhanced email spam filtering through combining similarity graphs. In: WSDM'11 Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 785–794 (2011)
9. Nosrati, L., Pour, A.N.: Dynamic concept drift detection for spam email filtering. In: Proceedings of ACEEE 2nd International Conference on Advances Information and Communication Technologies (ICT 2011), vol. 2 (2011) 124–126
10. Chakraborty, S., Mondal, B.: Spam mail filtering using different decision tree classifiers through data mining approach—a comparative performance analysis. Int. J. Comput. Appl. **47**, 26–31 (2012)
11. Pour, A.N., Kholghi, R., Roudsari, B.: Minimizing the time of spam mail detection by relocating filtering system to the sender mail server. Int. J. Netw. Secur. Appl. **4**, 53–62 (2012)
12. Ramachandran, A., Dagon, D., Feamster, N.: Can DNS-based blacklists keep up with bots?. CEAS 2006-Third Conference on Email and Anti-Spam (2006)
13. Levine, J.R.: Experience with greylisting. In: Proceedings of second conference on Email and Anti-Spam (CEAS 2005), pp. 1–2 (2005)
14. Amayri, O., Bouguila, N.: A study of spam filtering using support vector machines. J. Artif. Intell. Rev. **34**, 73–108 (2010)
15. Rouse, M.: Reverse DNS (rDNS) definition. (2007). http://searchnetworking.techtarget.com/definition/reverse-DNS. Accessed 01 March 2015
16. Tan, H., Goharian, N., Sherr, M.: $100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam. SIGIR'12 (2012)
17. Mosquare, A., Aouad, L., Grzonkowski, S., Morss, D.: On detecting messaging abuse in short text messages using linguistic and behavioural patterns (2014)
18. Bilal, J.M., Farooq, M.: Using evolutionary learning classifiers to do mobile spam (SMS) filtering. In: Proceeding GECCO'11 Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, pp. 1795–1802 (2011)
19. Nuruzzaman, M.T., Lee, C., Choi, D.: Independent and personal SMS spam filtering. In Proceedings of IEEE Conference on Computer and Information Technology, pp. 429–435 (2011)

20. Almeida, T.A., GÃ[3]mez Hidalgo, J.M., Yamakami, A.: Contributions to the study of SMS spam filtering: new collection and results. In: Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11) (2011)
21. British English SMS Corpora. (2011). Downloaded from http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/. Accessed 07 Aug 2014
22. DIT SMS Spam Dataset. Dublin Institute of Technology (DIT). (2012). Downloaded from http://www.dit.ie/computing/research/resources/smsdata/. Accessed 05 Aug 2014