Improving Static SMS Spam Detection by Using New Content-based Features

Completed Research Paper

Amir Karami Lina Zhou

University of Maryland Baltimore County

amir3@umbc.edu, zhoul@umbc.edu

Abstract

As the use of mobile phones grows, spams are becoming increasingly common in mobile communication such as SMS, calling for the research on SMS spam detection. Most of the existing detection techniques for SMS spams are have been mostly adapted from other contexts such as emails and the web but ignored some unique characteristics of SMS in spam detection. Moreover, spammers are constantly developing more sophisticated tactics, rendering previous features and methods for spam detecting no longer effective. In this paper, we propose to incorporate different new content based features to improve the performance of SMS spam detection. The effectiveness of the proposed features is empirically validated using multiple classification methods. The results demonstrate that the proposed features can improve the performance of SMS spam detection.

Keywords

Spam Detection, SMS, Classification, Content Features

Introduction

Spams are unwanted messages which can be transmitted over a communication media such as SMS. According to TIME[1] and Digital Trends[2], 6 billion out of 7 billion people in the world have access to cellphones and it is going to increase to 7.3 billion by 2014. Thus, the number of cellphones will soon outgrow the world population. In 2012, there were more than 6 billion daily Short Message Service (SMS) exchanges over mobile phones just in the US[3], and the rate of SMS spams increased by 400 percent. These spam messages not only waste network resources but also increase the cost for mobile phone users and even lead to cyber attacks such as phishing. Therefore, there is a strong need for SMS spam detection.

There are two major types of methods for detecting SMS spam: collaborative based and content based methods. The first one is based on the feedbacks from users and shared user experience. The second one is focused on analyzing the textual content of messages. This research adopts the second approach, which is more popular due to the difficulty in getting access to the data about usage and user experience. Content spam detection can be further classified into dynamic and static approaches. For instance, Lee et al. (2011) distinguished between spammers and non-spammers in twitter by tracking twitter users activities over a period of seven months. There are a host of studies on spam review detection by extracting features from the content of online consumer reviews (Delany et al. 2006; Delany et al. 2012; Ott et al. 2013; Ott et al. 2011). In particular, the research on SMS spam detention has focused on the content-based method. However, previous studies (Almeida et al. 2013; Almeida et al. 2011) have only considered words or tokens without looking into deep-level semantics. The choice of words and tokens can be easily manipulated by spammers. As a result, these detection methods have limited use because they are incapable to deal with constantly evolving spamming tactics.

To address the limitations of the state of research on SMS spam detection, we propose a content-based method that leverages lexical semantics. Instead of relying on individual words, our proposed method uses

[1] http://newsfeed.time.com/2013/03/25/more-people-have-cell-phones-than-toilets-u-n-study-shows/

[2] http://www.digitaltrends.com/mobile/mobile-phone-world-population-2014/

[3] ttp://www.cnn.com/2012/12/03/tech/mobile/sms-text-message-20/

semantic categories of words as features, which allows us to handle variations in word choices by spammers. In addition, using categories of words as features also helps to reduce the feature space, which in turn improves the efficiency of spam detection that has significant implications for SMS users. An empirical evaluation of the proposed methods has shown promising results.

This paper is organized in the following sections. In the section 2, we review the literature in SMS spam detection. Section 3 offers details about overall framework. Section 4 and 5 talks about the experiment, the performance evaluation, and analysis derived from our results. Finally we discuss conclusion and future work in the last section.

Related Work

In view of the similarity between the spams in emails and in SMS, researchers have extended the content-based detection methods for emails to SMS (e.g., Xiang et al. 2004; Gomez et al. 2006; Longzhen, An et al. 2009). However, there are some unique characteristics associated with short messages that distinguish it from emails such as limited number of characters in SMS (Kobus et al. 2008; Ling 2005) and lack of structural patterns in content presentation. The above difference has important implications for spam detection such as the selection of input features.

A variety of classification techniques have been applied to SMS spam detection. One of the earliest work in SMS spam detection (Xiang et al. 2004) used Support Vector Machines (SVMs). The same classification method was used by (Gomez et al. 2006), which selected tokens using Information Gain (IG). Longzhen, An et al. (2009) applied a k-nearest neighbor algorithm (k-NN) along with other spam detection methods on a dataset which contained 750 spam and ham SMS. Liu and Wang (2010) used the frequency of some text units as input attributes by using 6 classification methods including NB, kNN, kNN45, SVM, and ITC. Jie et al. (2010) added a weight to words to increase the cost of higher false positive. Almeida et al. (2011) tested 13 classification algorithms on a dataset that contained more than 5500 SMS (747 spam and 4827 ham). Their results show that SVM combined with the alphanumeric tokenization performed the best. The tokenization includes separating non-alphanumeric and alphanumeric characters. They finally extracted 81,000 tokens from short messages. Delany et al. (2012) modified the method and dataset used in (Almeida et al. 2011). This study focused on spam class evaluation factors but did not report F-Measure and ROC area. The overall accuracy showed that their method did not perform so well for non-spam SMS as did previous studies. Almeida et al. (2013) improved the evaluation factor of their previous work slightly by introducing new features.

In summary, the literature review shows that SVMs and Naive Bayes are the two most popular methods for spam classification (Xiang et al. 2004; Gomez et al. 2006; Nuruzzaman et al. 2011; Yadav et al. 2011; Xu et al. 2012). The best performance reported in the SMS spam detection literature to date is 97.59%, 95.48%, 0.899, and 2.09 for accuracy, spam caught percentage, Matthews Correlation Coefficient (MCC), and blocked hams percentage, respectively, produced by using 81,000 features. Two limitations are noted from these results. One is high false-positive rate and the other is low efficiency because of the large number of input features. In addition, they did not use evaluation metrics such as F-measure. We aim to address these limitations in this research by significantly reducing the number of input features, and providing an overall assessment of a detection method by adopting evaluation metrics such as F-measures.

A Content-based SMS Spam Detection Method

In this section, we present a conceptual framework of our proposed approach.

Problem Formulation

In SMS, there are a set of k text messages $TM = \{tm_1, \dots, tm_k\}$. Each message is limited to 160 characters consisting of words, number, etc. Messages can be about any topic (see samples messages in Figure 1.).

The tasks of SMS spam detection is to predict whether tm_i is a spam (A) or non-spam (B) by using a classifier c. The problem is formulated below:

$$c: tm_i \rightarrow \{spam, non - spam\}$$

To support the classification, we need to first extract a set of n features $F = \{f_1, \dots, f_n\}$ from TM.

A	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
A	Ok lar... Joking wif u oni...
B	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
A	U dun say so early hor... U c already then say...
A	Nah I don't think he goes to usf, he lives around here though
B	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

Figure 1: Sample short text messages (Almeida et al. ,2011)

A Framework

We propose a framework for detecting spam messages in SMS (see Figure 2.). This framework is composed of two major components: feature extraction and classification. The goal of feature extraction is to transform the input data into a set of features. It is very important in text analysis because it has a direct effect on machine learning techniques to distinguish classes or clusters; moreover, it is hard to find good features in unstructured data. In feature extraction step, we extracted two categories of features which will be introduced in detail in the following sections. In addition, we explored a wide range of classification algorithms from Random Forest to Naive Bayes and different test options to evaluate our proposed framework.

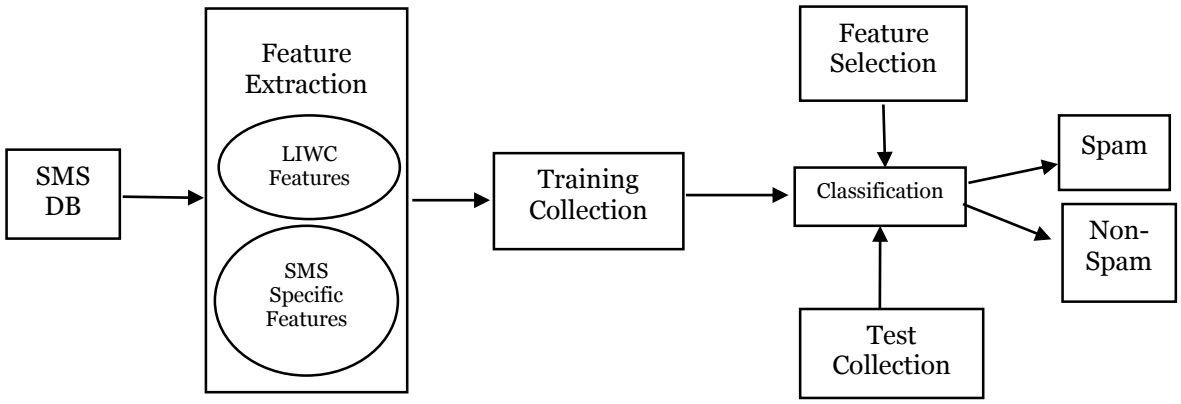

Figure 2: A Framework of Content based SMS Spam Detection

Extracted Features

The quality of a classifier depends on the quality of features. We explored two broad categories of features: SMS-specific content-based features and LIWC-based features.

SMS Specific (SMSS) Features

Some of SMS-specific content based features (see Table 3) were identified from the related literature (Lee et al. 2010; Stringhini et al. 2010a; Stringhini et al. 2010b; Wang 2010) including feature IDs 2,4,5,and6, some others are found by reading SMS and finding high frequency words in dataset including feature IDs 1,3, and 7, and finally the rest of features were calculated by measuring different combination of rates.

ID	Feature	ID	Feature
1	# Capital Words (CW)	9	The Rate of SW to UW
2	# Spam Words (SW)	10	The Rate of SW to S
3	# SMS Segments (S)	11	The Rate of URL to S
4	#Unique Words (UW)	12	The Rate of CW to UW
5	# URL	13	The Rate of CW to S
6	# SMS Frequency	14	The Rate of SW to S
7	# Using word "Call"	15	The Rate of UW to S
8	The Rate of URL to S	16	The Rate of URL to UW

Table 3: SMS Specific Features

LIWC Features

Linguistic Inquiry and Word Count (LIWC)[4] is an analysis tool for analyzing 80 different features including linguistic processes such as number of pronouns, psychological processes which has different subsets such as affective processes which has another subset such as degree of positive of negative feelings, current concerns such as degree of leisure, Spoken features such as degree of assent, and punctuation such as number of colons. In addition to the raw features collected from LIWC, we also incorporated additional features (230 features) based on various combinations of the raw features. For instance, we derived the relative polarity by examining the difference between the positive and negative feelings scores. To the best of our knowledge, this kind of sentiment features has not yet been explored for SMS spam detection.

Table 4 lists the top 20 features of LIWC features in SMS. This ranking is based on chi-square attribute selection method. This method measures the lack of independency between data. This table shows that linguistic features such as verbs, punctuations, words which are in dictionaries, and pronouns play an important role in spam detection. In addition, the topics such as money, leisure, death, and achievement, and the affective processes such as positive feeling can be a good indicator for spam detection.

An Experiment

In this section, we empirically evaluate our proposed method for SMS spam detection.

Dataset

We leveraged publically available datasets[5] in this research. These dataset was collected from the Grumbletext website[6], the National University of Singapore[7], a PhD thesis[8] and another website[9]. The dataset contains 5,574 manually labeled short messages, which can be broken down into two groups (see Table 1 and Figure 2):

- Spam: A collection of 747 SMS spam messages.
- Non-Spam (Ham): A Collection of 4,827 non-spam messages.

[4] http://www.liwc.net/

[5] http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

[6] http://www.grumbletext.co.uk/

[7] http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/

[8] http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

[9] http://www.esp.uem.es/jmgomez/smsspamcorpus/

ID	Top 20 LIWC Combined Features
1	The rate of score of verbs to the score of all words
2	The difference between the scores of "Money" and the score of "Death"
3	Score of "Punctuations"
4	Score of Dictionary Words
5	Score of "Pronouns"
6	The rate of the score of "Money" to the score of "Current Concern"
7	The difference between the scores of "semicolons" and "all punctuations"
8	The difference between the scores of "Question Mark" and " Apostrophe"
9	Score of "Exclamation Marks"
10	The difference between the scores of "Causation" and "Exclusive"
11	The rate of the score of "Exclusive" to the score of "Cognitive Processes"
12	The difference between the scores of "Leisure" and "Home"
13	The difference between the scores of "Certainty" and "Inhibition"
14	The rate of the score of "Achievement" to the score of "Home"
15	The rate of the score of "Parentheses" to the score of "Dashes"
16	Score of "i"
17	Score of "we"
18	The rate of the score of "positive" to the score of "Affective Processes"
19	Total Score of "Spoken categories"
20	The difference between the scores of "Insight" and "Discrepancy"

Table 4: Top-20 Selected Features

	Amount	Percentage
Non-Spams	**4827**	**86.6%**
Spams	**747**	**13.4%**
Total	**5574**	**100%**

Table 1: SMS Spam Dataset Statistics

Evaluation Metrics

The output of classification algorithms are presented as a confusion matrix, as shown in table 2. For evaluating the performance of spam detection, we measured precision, recall, F-measure, accuracy (ACC), Area Under the ROC curve (AUC), Spam Caught (SC), Blocked Hams (BH), False Negatives (FN), True

Positives (TP), and Matthews Correlation Coefficient (MCC). The evaluation metrics were defined based on the confusion matrix, as shown in equation (1)-(5).

		Predicted	
		Spam	**Non-Spam**
Actual	**Spam**	a	b
	Non-Spam	c	d

Table 2: Confusion Matrix

$$Precision(P) = \frac{a}{a+c} \quad (1)$$

$$Recall(R) = \frac{a}{a+b} \quad (2)$$

$$Accuracy = \frac{a+d}{a+b+c+d} \quad (3)$$

$$F-measure = 2 * \frac{PR}{P+R} \quad (4)$$

FN (c in the table) occurs when a spam is classified as non-spam, and TP (a in the table) occurs when a spam correctly classified as spam. ROC curves plot FP on the X axis vs. TP on the Y axis to find the tradeoff between them; therefore, the closer to the upper left indicates better performance. MCC is used to determine the quality of classification methods, ranging between -1 and +1 with +1 indicating the best performance.

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (5)$$

Classification Methods

In order to determine whether a SMS is spam or non-spam, we adopted supervised machine learning algorithms. We tested 40 classification algorithms with different testing settings by using Weka[10] , a machine learning tool. We chose chi-square as the feature selection method, which is among the most effective ones (Yang et al. 1997). Finally, we chose different evaluation test sets including different splits between training and test datasets including $\frac{1}{5}$, $\frac{1}{4}$, and $\frac{1}{3}$ as well as 10-fold cross validation.

Results

The classification accuracy ranges between 92% and 98% across various algorithms. Among the algorithms, boosting of Random Forest and SVM algorithms showed the best performance. The best results from using different categories of features are reported in table 5.

Discussion and Conclusion

The recent surge of mobile phone use makes the emerging communication media such as SMS particularly attractive for spammers. The challenge of detection spams in SMS is due the small number of characters in short text messages and the common use of idioms and abbreviation. The research that does exist on SMS spam detection has only focused on word distribution but has yet to examine explicit semantic categories of text expressions. In the current study, we proposed to employ categories of lexical semantic features in the detection of SMS spam. Our experiment results show that incorporating semantic categories improve the performance of SMS spam detection. We also compared the results of our method with those from the state-of-the-art research in this area.

[10] http://www.cs.waikato.ac.nz/ml/weka/

Method	LIWC+ SMSS Features	LIWC+ SMSS Features	LIWC+ SMSS Features	LIWC Features	SMSS Features
Classifier	Boosted-Random Forest	Boosted-Random Forest	SVM	SVM	Random Forest
#Features	320	320	150	250	16
SC%	89.1	88	89	89.1	75.5
BH%	0.1	0.1	0.6	0.3	2.2
MCC	0.934	0.926	0.914	0.923	0.768
Acc%	98.47	98.27	97.99	98.21	94.69
TPR	0.985	0.983	0.98	0.982	0.947
FNR	0.094	0.103	0.095	0.095	0.214
Precision	0.985	0.983	0.982	0.982	0.945
Recall	0.985	0.983	0.98	0.982	0.947
F-Measure	0.984	0.982	0.98	0.982	0.946
AUC	0.968	0.960	0.942	0.923	0.952

Table 5: Classification Performance

The findings of this study show that combining different types of features can lead to improvement of the classification performance. Specifically, using both LIWC and SMS-specific content based features can significantly improve the performance of spam detection. We conducted a comparison between the results of the current study and those of the best performance reported in the SMS spam literature (Almeida et al. 2013; Almeida et al. 2011; Delany et al. 2012) , as shown in Figures 4 and 5. Among the previous results, the best Acc%, SC%, MCC, and BH% are 97.59, 95.48, 0.899, and 2.09. Therefore, our proposed method improves these evaluation metrics to 98.47, 89.1, 0.934 and 0.1. The comparison results show that using our proposed content-based features improves the performance of state-of-the-art in SMS spam detection.

The comparison also shows that we used much fewer features than the state-of-the-art method for SMS spam detection by using a small fraction of the (i.e., 320 in total) 81,000 attributes used in previous studies.

The features identified in this study may be applied to improve spam detection in other types of communication media such as emails, social network systems, and online reviews. These features will also help to improve mobile users' aware of spam and their knowledge on how to detect spams in SMS.

There are some interesting future research issues such as incorporating dynamic features by tracking the usages of different words over time and testing the generality of the proposed features in other communication media such as online review and social networks. In addition, there is space for further improving the performance of spam detection.

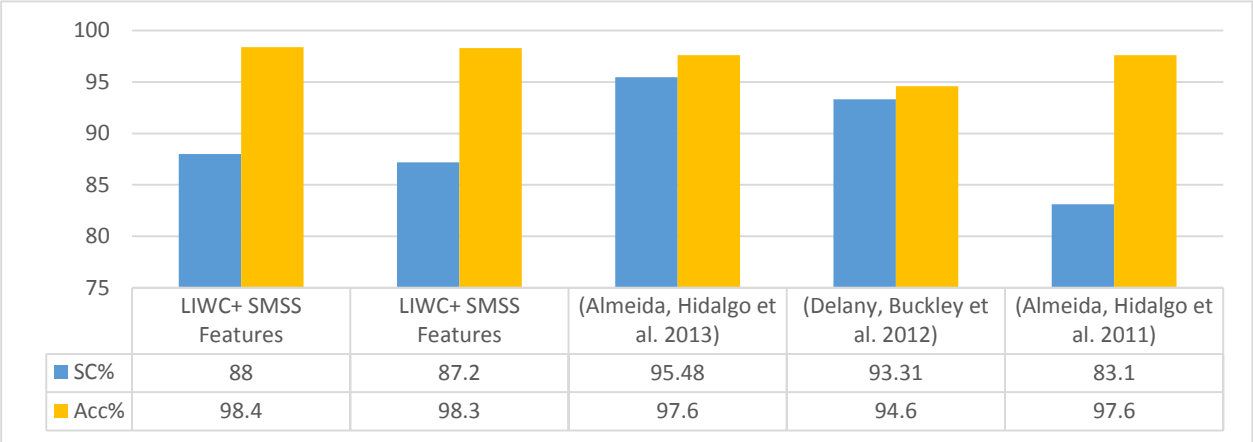

Figure 4: SC% and Acc% Comparison

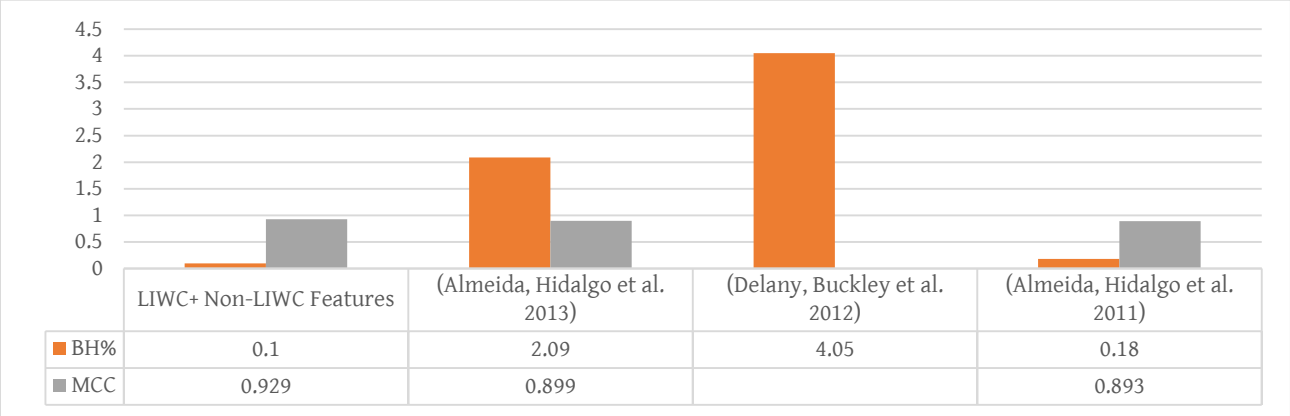

Figure 5: BH% and MCC Comparison

REFERENCES

Almeida, T., Hidalgo, J.M.G.m., and Silva, T.P. "Towards SMS Spam Filtering: Results under a New Dataset," *International Journal of Information Security Science* (2:1) 2013, pp 1-18.

Almeida, T.A., Hidalgo, J.M.G., and Yamakami, A. "Contributions to the study of SMS spam filtering: new collection and results," Proceedings of the 11th ACM symposium on Document engineering, ACM, 2011, pp. 259-262.

Delany, S.J., Buckley, M., and Greene, D. "SMS spam filtering: Methods and data," *Expert Systems with Applications* (39:10) 2006, pp 9899-9908.

Delany, S.J., Buckley, M., and Greene, D. "SMS spam filtering: Methods and data," *Expert Systems with Applications* (39:10) 2012, pp 9899-9908.

Gomez Hidalgo, J.M.a., Bringas, G.C., SÃinz, E.P., and GarcÃa, F.C. "Content based SMS spam filtering," Proceedings of the 2006 ACM symposium on Document engineering, ACM, 2006, pp. 107-114.

Jie, H., Bei, H., and Wenjing, P. "A Bayesian approach for text filter on 3G network," Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on, IEEE, 2010, pp. 1-5.

Kobus, C., Yvon, F.o., and Damnati, G.r. "Normalizing SMS: are two metaphors better than one?," Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 2008, pp. 441-448.

Lee, K., Caverlee, J., and Webb, S. "Uncovering social spammers: social honeypots+ machine learning," Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, 2010, pp. 435-442.
Lee, K., Eoff, B.D., and Caverlee, J. "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," ICWSM, 2011.
Ling, R. "The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians," in: *Mobile Communications*, Springer, 2005, pp. 335-349.
Liu, W., and Wang, T. "Index-based online text classification for sms spam filtering," *Journal of Computers* (5:6) 2010, pp 844-851.
Longzhen, D., An, L., and Longjun, H. "A new spam short message classification," Education Technology and Computer Science, 2009. ETCS'09. First International Workshop on, IEEE, 2009, pp. 168-171.
Nuruzzaman, M.T., Lee, C., and Choi, D. "Independent and personal SMS spam filtering," Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on, IEEE, 2011, pp. 429-435.
Ott, M., Cardie, C., and Hancock, J.T. "Negative Deceptive Opinion Spam," Proceedings of NAACL-HLT, 2013, pp. 497-501.
Ott, M., Choi, Y., Cardie, C., and Hancock, J.T. "Finding deceptive opinion spam by any stretch of the imagination," *arXiv preprint arXiv:1107.4557*) 2011.
Stringhini, G., Kruegel, C., and Vigna, G. "Detecting spammers on social networks," Proceedings of the 26th Annual Computer Security Applications Conference, ACM, 2010a, pp. 1-9.
Stringhini, G., Kruegel, C., and Vigna, G. "A Study on Social Network Spam," Graduate Student Workshop on Computing, 2010b, p. 43.
Wang, A.H. "Detecting spam bots in online social networking sites: a machine learning approach," in: *Data and Applications Security and Privacy XXIV*, Springer, 2010, pp. 335-342.
Xiang, Y., Chowdhury, M., and Ali, S. "Filtering mobile spam by support vector machine," CSITeA'04: Third International Conference on Computer Sciences, Software Engineering, Information Technology, E-Business and Applications, International Society for Computers and Their Applications (ISCA), 2004, pp. 1-4.
Xu, Qian, Evan Wei Xiang, Qiang Yang, Jiachun Du, and Jieping Zhong. "Sms spam detection using noncontent features." *IEEE Intelligent Systems* 27, no. 6 (2012): 44-51. Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., and Naik, V. "SMSAssassin: Crowdsourcing driven mobile-based system for SMS spam filtering," Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, ACM, 2011, pp. 1-6.
Yang, Y., and Pedersen, J.O. "A comparative study on feature selection in text categorization," ICML, 1997, pp. 412-420.