

Graph-Based Learning Model for Detection of SMS Spam on Smart Phones

Muhammad Zubair Rafique
Center of Excellence in Information Assurance
King Saud University
Riyadh, Kingdom of Saudi Arabia
E-mail: rafique.zubair@gmail.com

Muhammad Abulaish, *SMIEEE*
Center of Excellence in Information Assurance
King Saud University
Riyadh, Kingdom of Saudi Arabia
E-mail: mAbulaish@ksu.edu.sa

Abstract—Short Message Service (SMS) has been increasingly exploited through spam propagation schemes in recent years. This paper presents a new method for graph-based learning and classification of spam SMS on mobile devices and smart phones. Our approach is based on modeling the content and patterns of SMS syntax into a directed-weighted graph through exploiting modern composition style of messages. The graph attributes are then used to classify spam messages in real-time by using KL-Divergence measure. Experimental results on two real-world datasets show that our proposed method achieves high detection accuracy with less false alarm rate to detect spam messages. Moreover, our approach requires relatively less memory and processing power, making it suitable to deploy on resource-constrained mobile devices and smart phones.

Index Terms—SMS spam detection; Graph-based SMS modeling; Probabilistic classification; Smart phones.

I. INTRODUCTION

In recent years, Short Message Service (SMS) has become one of the most popular adopted data communication service on mobile devices. The service is utilized by hundreds of millions users daily¹ because of its ease of usage, simplicity, prompt delivery, and in some cases significantly cheap rates as compared to voice services of cellular networks. SMS usage grows beyond traditional *texting* and is now being extensively used in authentication (e.g., mobile banking, one time password delivery, etc.), information retrieval systems (e.g., TV shows), smart phones configuration (Over-The-Air (OTA) configuration) and social web site alerts (e.g., Facebook, Twitter, etc.).

On the heels of wide usage and dependency, the service is increasingly abused in terms of both social and security threats ranging from simple advertising and political campaigns to more sophisticated threats like phishing, scam and stealing of personal information [2]. The gravity of problem can be analyzed by a survey report, which shows that the number of spam SMS exceeds more than 50% of the total received SMS [1]. Alone in UK, 66% of mobile phone users received spam text messages [2]. It is also worth to note that a large volume of spam SMS originates from cellular companies themselves containing information related to new offers and deals [2]. With this level of nuisance, it is necessary and useful to

stop spam SMS on personal mobile devices and smart phones through deploying a light-weight spam detection system.

Unlike email, spam SMS not only intrudes the privacy of the users, but also adds to their annoyance because in most mobile phones its arrival is indicated through an alert tone. The traditional approaches of filtering spam email can not be applied straight forwardly to detect spam SMS on mobile devices and smart phones. SMS significantly differs in writing style and is usually written by using different variants of words and short abbreviations that cannot be detected by traditional keywords-based algorithms [3]. In addition, the email related spam detection schemes require large memory and processing resource, and consequently they are not suitable for detecting spam on resource-constrained mobile devices and smart phones.

In this paper, we propose a novel graph-based supervised learning method that models the syntactical features of SMS to classify spam messages. Instead of applying any shallow or deep parsing technique, we transform the messages into a set of space-delimited tokens. This set constitute the node-set of the graph. Directed edge between a pair of nodes is derived on the basis of their occurrence sequence. Each node and edge of the graph is assigned a weight as a function of occurrence probabilities of tokens. The proposed method is efficient and is capable of detecting spam SMS including scams, selling financial services, offers and promotions from telecom operators, phishing, and push messages to download malware. More specifically, the proposed method aims to build a graph of words from the labeled messages stored on users mobile (training data) and compute the probability of occurrence and joint distribution among the words of spam and benign messages. The probabilities are then used to classify new incoming messages by using KL-Divergence. The proposed scheme presented in this paper takes into consideration the *modern composition* of SMS including short abbreviated messages and common word adulteration techniques.

We evaluate our proposed scheme on two real-world datasets of benign and spam messages. The first one is a publicly available dataset *NUS SMS Corpus (NSC)*, which contains 1000 benign and 425 spam messages collected from the volunteers of National University of Singapore. The second dataset, collected by our research team for one of our previous

¹A recent market survey shows that more than 8.1 trillion text messages will be sent over carrier networks worldwide in year 2011 [1].

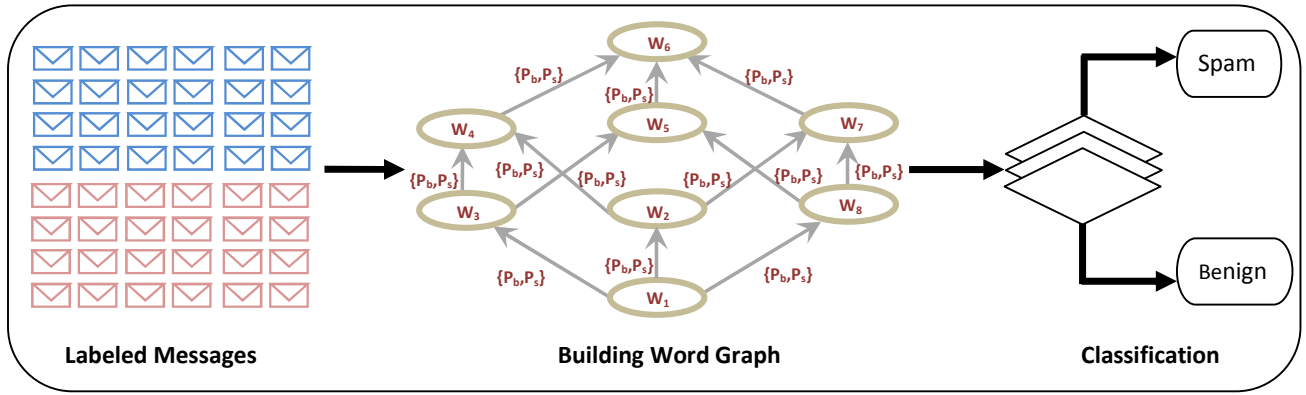


Fig. 1. Architecture of the proposed graph-based SMS classification system

works [4], is a collection of more than 5000 real-world benign and 2000 spam messages. The experimental results show that our method achieves high detection accuracy with less false alarm rate to detect spam messages. To the best of our knowledge, our system is the first method to detect spam messages on mobile devices and smart phones using graph-based learning model with the ability to generalize on different writing styles of SMS composed in different languages.

The rest of the paper is organized as follows. Section II presents an overview of the proposed graph-based SMS classification model to detect spam messages on mobile devices and smart phones. In Section III, we discuss the details of graph-based learning model we have adopted for our approach followed by the illustration of classification scheme in section IV. We discuss and analyze the real-world datasets and results of our study in section V. In section VI, we briefly describe the related work in the field of SMS spam detection. Finally, we conclude the paper with an outlook to our future research.

II. PROPOSED ARCHITECTURE

In this section, we describe our filtering method to detect spam messages received by the users on his/her mobile phone. We intend the system to detect spam messages of different nature by learning the patterns of the stored messages in users' inbox. The proposed method can be used to detect spam messages and store them into a spam folder without user notification through ring tone or vibration alert. We adopt following design goals for our spam filtering system:

- **Modular** – The spam detection process should be modular in nature so that it could operate on different architecture of mobile devices and smart phones.
- **Syntax independent** – The learning method should be language and writing style independent so that it could incorporate the modern composition style of SMS, which contains short abbreviated words and non-language conformance phrases.
- **Resource efficient** – The memory and processing requirement for the learning and classification process should be

in accordance with the available resources in the resource-constrained mobile devices.

- **Accurate** – The classification system should be able to detect spam messages with low false alarm rate so that the benign messages could not be marked as spam.

A. System Flow

Figure 1 shows the system flow of our approach to detect spam messages at application layer² of mobile devices and smart phones. The main goal is to silently detect the spam message and placed it in a spam folder.

In our approach, the system requires mobile phone users to label messages in the inbox as benign or spam. Using the labeled messages the graph is constructed by extracting space delimited tokens from the messages. The main purpose of using space delimited tokens is to take in account the modern composition style of SMS composed using short abbreviations or by combining different alphabets and special characters. It is worth to note that these compositions of SMS are not actually the part of any formal language and hence can not be detected by using traditional keyword-based detection schemes [4].

Once the graph is build using the previously stored messages, we compute the probability of occurrence of specific token in both benign and spam messages. Furthermore, we compute the link probability of nodes both in spam and benign messages. The computation of link probabilities between nodes of graph helps us to include an extra information about the sequence of tokens used in different composition of benign and spam messages, as discussed in Section III.

²The architecture of current smart phones and mobile devices is composed of two layers – the application layer and the access layer. The SMS is received at the access layer, by the GSM modem, from a Short Message Service Center (SMSC). The OS of mobile device operates at application layer and received SMS from GSM modem through telephony stack. The SMS received from the modem (140 bytes) is usually in the Protocol Description Unit (PDU) format [5] and is decoded to human readable text at application layer. Once the message is decoded, the OS notified the user through ring tone or vibration alert for an incoming message. Our goal is to sniff the message and perform classification before it is notified to the mobile user. More details on SMS and smart phone architecture can be found in [5].

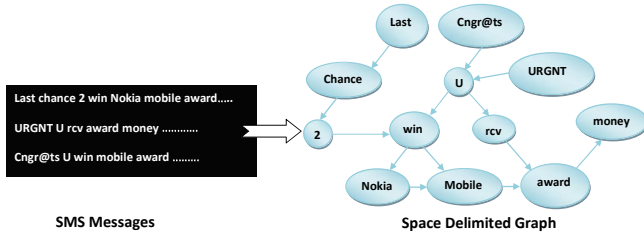


Fig. 2. A sample directed graph constructed from space delimited tokens.

The computed probabilities are then given to classification module as an input which operates as a binary classifier to classify messages as either spam or benign. The training phase (i.e., graph construction and probability calculation) occurs off-line and is independent of real-time detection of spam messages. Mobile phone users can update the training model any time depending on his/her feasibility. However, the real-time detection is not in user control and it starts functioning as soon as the model is trained. The decision is made by simply calculating the divergence between the probabilities of links and nodes in the graph for the new incoming messages.

III. GRAPH-BASED MODEL

Graph structures have been extensively used in text classification systems and they have application from simple deterministic classification problems to more advance probabilistic classification of objects. Our approach is based on the construction of graph from the space delimited tokens of messages. The implicit set of tokens obtained from the messages correspond to all possible words separated by space in a particular message. Since the length of SMS message is 140 bytes only, the substrings (tokens) obtained from messages are limited and is bounded by the length of SMS message. Let S be the set of labeled messages from the user's inbox and T is a finite set of tokens extracted from the messages. The function ∇ operates on each incoming message and transform it into a set of token as $T = \{t_1, t_2, \dots, t_k\}$.

Once the tokens are extracted from the messages, a directed Graph G is constructed for the model learning and classification purpose. The nodes (N) of the graph represent the extracted token from SMS message and edges E are the links between the consecutive tokens. The classification is based on the features of both extracted tokens and properties of the links between the tokens. The token's attributes can be useful to determine the presence of particular substring (token) in a message, whereas link's properties can provide a valuable information about the sequential patterns of the tokens. Figure 2 shows three example messages and a sample directed graph constructed from them.

A. Node Attributes

The attributes of nodes in the token graph provide a basic description of objects in the construction of the graph. We compute the probability of occurrence of each particular token during the training phase. Since each node in the graph

contains unique substring (token) value, whenever a particular token is found in other messages during training, only the corresponding probability values are updated and no new node is added. Each node (t_i) is represented using a pair of real values attributes $\langle P_{b_i}, P_{s_i} \rangle$, where P_{b_i} is the probability of occurrence of t_i in benign messages and P_{s_i} is the probability of occurrence of t_i in spam messages.

B. Edge Attributes

The motivation behind computing edge attributes is to model the sequential patterns of tokens appearance in both spam and benign messages that are available for training purpose. To achieve this task, we compute the out-going transition probabilities between the nodes of the graph in training phase. Let $O(N_i)$ is the set of outgoing neighbors of a token t_i . With each node N_i , such that $N_i \rightarrow t_i$, we compute the transition frequency of edges on the set of outgoing neighbors ($O(N_i)$) in both spam and benign messages. For a given graph with k nodes, the adjacency matrix, given below, demonstrates the probability distribution of different possible node-pairs.

$$A = \begin{bmatrix} (P_{b(0,0)}, P_{s(0,0)}) & (P_{b(0,1)}, P_{s(0,1)}) & \dots & (P_{b(0,k)}, P_{s(0,k)}) \\ (P_{b(1,0)}, P_{s(1,0)}) & (P_{b(1,1)}, P_{s(1,1)}) & \dots & (P_{b(1,k)}, P_{s(1,k)}) \\ \vdots & \vdots & \ddots & \vdots \\ (P_{b(k,0)}, P_{s(k,0)}) & (P_{b(k,1)}, P_{s(k,1)}) & \dots & (P_{b(k,k)}, P_{s(k,k)}) \end{bmatrix}$$

Note that $P_{b(i,j)}$ corresponds to the transition probability of the movement from node N_i to N_j in benign messages, whereas $P_{s(i,j)}$ corresponds to the transition probability of the movement from node N_i to N_j in spam messages. These links probabilities correspond to the conditional distribution of tokens in both benign and spam messages. This provides a valuable information about the sequential patterns of the tokens in benign and spam messages used in learning the model.

The conditional distribution has an advantage over traditional n-gram method [6] in which information is obtained through joint distribution of words. Also, in such systems it is difficult to choose an appropriate value of n a priori. A small value of n increases the probability of false detection, whereas a large value of n significantly increases the processing and memory overhead. The increase in processing and memory overhead demands large resources making the approach impractical for resource-constrained devices like mobile phones. The conditional distribution of tokens in our approach reduces the underlying memory and processing requirements, making it feasible for resource-constrained mobile devices.

IV. CLASSIFICATION OF SPAM MESSAGES

Once the graph is built, and the node and edge attributes are calculated, our goal is to detect incoming spam messages in a real-time. To achieve this task, we choose statistical measures to compute the distance/divergence of incoming SMS from the attributes of our training model. Statistical measures are chosen due to low processing overheads making the method suitable for classifying spam messages on resource-constrained mobile devices and smart phones [7]. In this study,

we employ well-known information theoretic measure KL-Divergence (or relative entropy) to detect spam SMS message in real-time.

KL-Divergence is central to information and statistical theory and computes the distance between two probability distributions [8]. It has been successfully used in document classification by comparing the distribution of terms in text documents [9]. With a set of extracted tokens from an incoming message, KL-Divergence can be used to solve the *two-sample* problem where the target is to detect whether two set of samples (tokens) have been drawn from the same distribution (i.e., same category of SMS – benign or spam) [8].

Each incoming message is transformed to space delimited token array using ∇ function (see Section III). The extracted array of tokens are then used to compute divergence scores from the graph. We compute two types of divergence score from the generated graph – KL_b and KL_s . KL_b represents the distance of incoming message from benign attributes ($P_{b_i}, P_{b(i,j)}$) of the training graph, whereas KL_s represents the distance of incoming message from spam attributes ($P_{s_i}, P_{s(i,j)}$) of the graph. Mathematically, the two divergence scores can be obtained using equations 1 and 2.

$$KL_b = \sum_{i=0, j=i+1}^{k-1} \theta_{t_b(i,j)} * \log_2 \left(\frac{\theta_{t_b(i,j)}}{\theta_{t_{b_i}} * \theta_{t_{b_j}}} \right) \quad (1)$$

$$KL_s = \sum_{i=0, j=i+1}^{k-1} \theta_{t_s(i,j)} * \log_2 \left(\frac{\theta_{t_s(i,j)}}{\theta_{t_{s_i}} * \theta_{t_{s_j}}} \right) \quad (2)$$

Note that the summation runs over the index of extracted token array of incoming message and k represents the length of the extracted token array. The functions $\theta_{t_b(i,j)}$ and $\theta_{t_s(i,j)}$ are the edge attributes ($P_{b(i,j)}, P_{b_i}$) between the tokens t_i and t_j , whereas $\theta_{t_{b_i}}$ and $\theta_{t_{s_i}}$ are the node attributes of token t_i in the generated graph model. It may be the case that a new incoming message contains few tokens that are not present in the training model of the graph. In this case, the value of the θ function is set to zero for the corresponding token, i.e., $\theta_{t_i} = 0$, $\theta_{t_j} = 0$ and $\theta_{t(i,j)} = 0$. The resulting values of KL-Divergence indicates how closely the new message follows the probability distribution of benign and spam messages.

Through our pilot studies, we found that the KL_b divergence value of particular message is normally greater than KL_s when message belongs to benign category. Similar pattern holds for spam messages in which KL_s is usually greater than KL_b . Therefore, we can detect messages by simply comparing the two scores defined through equations 1 and 2. But, in order to avoid the wrong detection of benign messages as spam during classification, the minimum threshold value α is added to the basic boolean expression $KL_s > KL_b + \alpha$.

V. EVALUATION STRATEGY

In this section, we present our approach for evaluating the graph model for detection of SMS spam on mobile devices and smart phones. We first discuss the real-world benign and spam datasets used for evaluating the proposed model. Afterwards,

we define the key metrics used to carry out a comprehensive analysis of the proposed model. Finally, we discuss the results of our experiments on the real-world datasets.

A. Benign and Spam Dataset

In order to analyze the detection capability of our proposed approach to detect spam messages, we have used two real-world datasets containing spam and benign SMS messages. The first dataset (hereafter termed as dataset-1) is publicly available dataset *NUS SMS Corpus (NSC)*³, which contains 1000 legitimate benign SMS collected from the volunteers at National University of Singapore. The dataset also contains 425 spam messages manually extracted from Grumbletext⁴. The dataset is also used by [10], [11], and [12] to evaluate the effectiveness of different spam detection methodologies.

The second dataset (hereafter termed as dataset-2), used in our previous work [4], is a collection of more than 5000 real-world benign and 800 spam SMS. This dataset contains a diverse set of messages received by people belonging to different socioeconomic background including – teenagers, researchers, students, professionals, teachers and senior citizens. The dataset also contains additional 1200 spam messages extracted manually from Grumbletext. The collection of two datasets give us wide variety of messages composed in various styles by different cell phone user communities.

B. Experiments and Results

The goal of our experiments is to analyze the classification accuracy of the proposed model on the real-world SMS datasets. To this end, we have developed a prototype system of the graph-based model and classification module presented in Section III and Section IV, respectively. For our experiments, the datasets are divided into two separate sets for training and testing. The training data is used for generating the graph model, while the test set is used to determine the accuracy of the classification module.

We extracted benign and spam messages in equal sets for experiments to generate the graph of space delimited tokens and computed the attributes of nodes and edges in the graph. We performed k -fold cross validation, in which we divide the extracted set of messages into k partitions. $k - 1$ partitions are used for training and the rest is used for testing. The procedure is repeated on all the extracted sets in a particular dataset and the reported results are averaged. To gain better knowledge of classification potential of the proposed approach, we validate the system for three different values (3, 7, and 10) of k .

Based on the classification results, we calculated the true positives TP (number of spam messages the classification module identifies as spam), the false positives FP (number of benign messages the classification module identifies as spam), true negatives TN (number of benign messages the classification module identifies as benign), and the false negatives FN (number of spam messages the classification module identifies

³<http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/>

⁴Grumbletext is a UK-based consumer complaint web site for online reporting of spam SMS message <http://www.grumbletext.co.uk/>

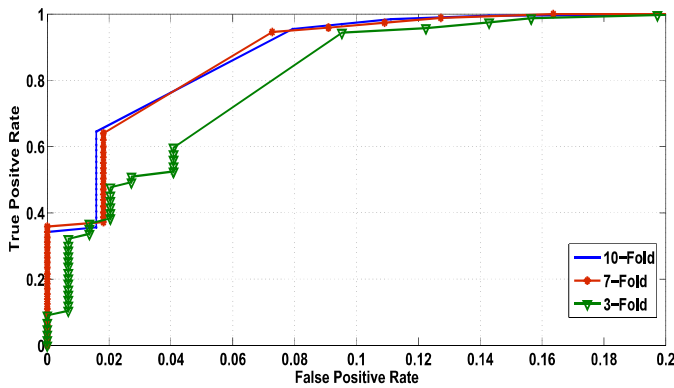


Fig. 3. ROC analysis on *dataset-1*

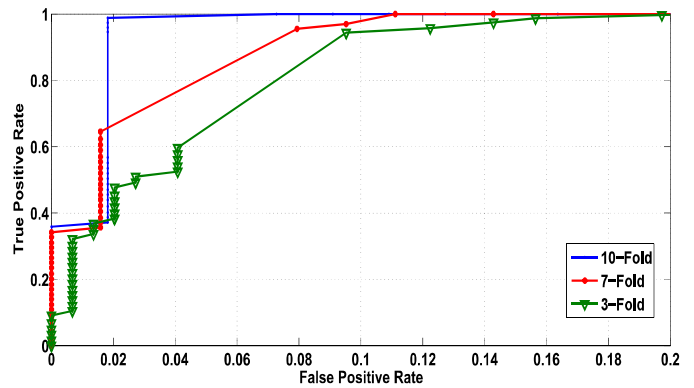


Fig. 4. ROC analysis on *dataset-2*

as benign). Using these values, we calculated true positive rate (TPR) and false positive rate (FPR) using equations 3 and 4, respectively.

$$TPR = \frac{TP}{TP + FN} \quad (3)$$

$$FPR = \frac{FP}{FP + TN} \quad (4)$$

The TPR shows the spam detection accuracy of the proposed model, whereas the FPR indicates the false detection of benign messages. Both TPR and FPR values assist us in carrying out standard ROC [13] analysis to evaluate the effectiveness of the proposed model for detecting spam SMS on mobile devices.

ROC curves are extensively used by AI community to carry out the trade-off analysis between TPR and FPR of a given model. The ROC analysis of our experiments on both of the datasets, *dataset-1* and *dataset-2*, are shown in Figures 3 and 4, respectively. It can be observed from the results presented in Figure 3 that our system achieves 98% detection rate with false alarm rate of less than 0.08 during the classification of spam SMS in 10 fold cross validation on *dataset-1*. Similar results are obtained when classification is performed using 7 fold cross validation. However, the 3 fold cross validation result shows degradation in both TPR and FPR in classifying the spam and benign SMS. The same pattern of classification can be observed in Figure 4 representing the ROC curve for *dataset-2*. A higher TPR indicates the better classification of spam messages received by the mobile users on their mobile phone. The low FPR indicates that less benign messages will be moved into the spam folder without user notification. With these results, we believe that our approach is highly suitable for detection of spam messages on smart phones and mobile devices.

VI. RELATED WORK

Cellular operators have recently devised Open Mobile Alliance (OMA) to prevent spam on cellular networks received by the users [14]. The procedure mainly functions at the cellular operator side and provides effective protection against massive spam schemes. On contrary, the techniques for the

personalized detection of spam messages for mobile devices are usually an adaption of email spam detection algorithms and usually incorporate grammar-based keywords features along with n-gram of characters for the classification of spam messages [12]. These approaches based on content analysis used machine learning algorithms on extracted feature set for classification of spam messages. For instance, [15] used common machine learning algorithms like SVM, KNN, or Naïve Bayes for the classification of SMS spam. Also, the author of [16] used a Hidden Markov Model (HMM) to detect spam SMS at the access layer of mobile devices. Recently, in [4] the authors proposed a method based on evolutionary algorithms to detect spam SMS at access layer of mobile devices. However, no previous study has exploited graph based methods for the detection of SMS spam on mobile device.

VII. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel graph-based spam detection architecture that tokenizes SMS messages and exploits their occurrence and sequential patterns to detect spam messages on mobile devices. The proposed scheme is evaluated on two real-world datasets containing both benign and spam messages and the achieved accuracy for detecting SMS spam is very appealing. The other important contribution of the proposed scheme lies in its real-time deployability to classify spam messages by using well-known statistical measure, KL-Divergence. The focus of our future work will be to study the effect of the proposed model on different languages. We also plan to extend this work to develop a more robust approach for spam SMS detection based on the concept drift phenomenon and incremental learning.

REFERENCES

- [1] IronPort, "Case-Study: IronPort Helps a Nationwide Carrier Stop Wireless Threats," http://www.ironport.com/pdf/ironport_case_study_wireless.pdf.
- [2] Text4ever, "White Paper: UK spam study," Oct. 2009, <http://www.txt4ever.com/study/spamstudy.pdf>.
- [3] D.-N. Sohn, J.-T. Lee, and H.-C. Rim, "The contribution of stylistic information to content-based mobile spam filtering," in *Proceedings of the ACL-IJCNLP Conference Short Papers.*, 2009.

- [4] M. Rafique, N. Alrayes, and M. Khan, "Application of evolutionary algorithms in detecting sms spam at access layer," in *Proceedings of the 16th annual conference on Genetic and evolutionary computation*. ACM, 2011, pp. 1787–1794.
- [5] ETSI-GSM, "03.40 Technical realization of the SMS," Apr. 1998, <http://www.3gpp.org/ftp/Specs/html-info/0340.htm>.
- [6] F. Ahmed, H. Hameed, M. Z. Shafiq, and M. Farooq, "Using spatio-temporal information in API calls with machine learning algorithms for malware detection," in *Proceedings of the 2nd ACM workshop on security and artificial intelligence*. NY, USA: ACM, 2009, pp. 55–62. [Online]. Available: <http://doi.acm.org/10.1145/1654988.1655003>
- [7] M. Rafique, M. Khan, K. Alghathbar, and M. Farooq, "A framework for detecting malformed sms attack," *Secure and Trust Computing, Data Management and Applications*, pp. 11–20, 2011.
- [8] F. Pérez-Cruz, "Kullback-leibler divergence estimation of continuous distributions," in *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*. Ieee, 2008, pp. 1666–1670.
- [9] M. Gamon, "Graph-based text representation for novelty detection," in *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 17–24.
- [10] G. V. Cormack, J. M. Gómez Hidalgo, and E. P. Sánz, "Spam filtering for short messages," in *Proceedings of the 16th ACM Conference on information and knowledge management*. NY, USA: ACM, 2007, pp. 313–320.
- [11] —, "Feature engineering for mobile (SMS) spam filtering," in *Proceedings of the 30th annual conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2007, pp. 871–872.
- [12] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García, "Content based SMS spam filtering," in *Proceedings of the 2006 ACM symposium on Document engineering*. NY, USA: ACM, 2006, pp. 107–114. [Online]. Available: <http://doi.acm.org/10.1145/1166160.1166191>
- [13] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine Learning*, vol. 31, 2004.
- [14] GSMWorld, "GSMA to Address Spam and Fraudulent Messaging Threats for Consumers," Mar. 2010, <http://gsmworld.com/newsroom/press-releases/2010/4797.htm>.
- [15] M. Healy, S. Delany, and A. Zamolotskikh, "An assessment of case-based reasoning for short text message classification," in *Proceedings of 16th Irish Conference on Artificial Intelligence and Cognitive Science*, ser. AICS '05, Sep. 2005, pp. 257–266.
- [16] M. Z. Rafique and M. Farooq, "'Be Liberal in What You Recieve' on your mobile phone," in *Proceedings of 20th Virus Bulletin Conference*, ser. VB '10, Vanc., Canada, 2010.