J. Basic. Appl. Sci. Res., 4(4)275-279, 2014
© 2014, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

SMS Spam Detection Using Simple Message Content Features

Dr. Ghulam Mujtaba[1], Majid Yasin[2]

[1]Assistant Professor, Electrical Engineering Department, Comsats Institute of Information Technology, Abbottabad, Pakistan
[2]Electrical Engineering Department, Comsats Institute of Information Technology, Abbottabad, Pakistan

Received: January 30 2014
Accepted: March 11 2014

ABSTRACT

Short message services (SMS) spam is increasing as more people exchange SMS messages very frequently. It is desirable to be eliminated for a number of reasons. This work describes a mobile station based approach where the spam sms would be identified and removed as soon as it is received at the mobile device. Four features are derived from each sms message and using these features a trained machine learning algorithm can classify an unknown message to be spam or ham. These features are the size of the message and existence of frequently occurring monograms in the message, existence of frequently occurring diagrams in the message and message class. The performance of Naïve bayes algorithm is shown to be better than other algorithms explored. The other algorithms are Artificial Neural Networks and Decision Tree classifier.

KEYWORDS: sms spam, m-spam, classification, machine learning, identification.

1. INTRODUCTION

Short messaging Services (SMS) are an important means of communication today between millions of people around the world. SMS services, which are a must-have service nowadays for telecom operators, transmit their messages using standardized communications protocols. At the same time we also face a serious problem caused by SMS spamming. Spam is the use of electronic messaging systems to send unsolicited bulk messages, especially advertising, indiscriminately [1]. While the most widely recognized form of spam is e-mail spam, SMS spam is also increasing resulting in resource consumption and annoyance at the recipient, like email SPAM. SMS spam is a form of spamming directed at the short messaging service which usually contain marketing materials, much like email spam. It is described as mobile spamming, SMS spam, text spam, m-spam or mspam. In 2012, in the US alone more than 69% of the mobile users had received SMS spam [2]. A study by the security firm Cloudmark showed that 66% of UK mobile subscribers had received SMS spam [3]. Other than advertisement, the SMS spam is also being used in other threats, like phishing, and identity theft [3].

The spam in SMS causes annoyance at the user, resource consumption of the mobile device and in some communities even the receiver is charged for getting the SMS.

Hence it is of very much significance that these spam messages be removed as soon as they are received at the mobile station, if not before that. The issue of mobile phone spam has not received that much attention by the research community as the more familiar email spam. In [4] a scheme is proposed that monitors an SMS platform from a central vantage point within the mobility network and detects potential spam campaigns when it sees unusually high number of messages with similar content being transmitted. a service-side solution to the spam sms problem is provided that uses graph data mining to distinguish likely spammers from normal senders without checking a message's contents .Wang et. All have proposed a scheme based on behavior based social network and temporal (spectral) analysis to detect spam [5].Rick and Peter proposed a method to filter spam SMS by including a SMS message differentiating module in the routing node [6].

In this paper, a simple and fast framework for the identification of SPAM sms at the mobile station node is presented. It relies on the fact that spamsms is essentially a problem of text classification [7]. It should be kept in mind that simplicity of the scheme is very important because of the limited resources available on a mobile device. It is desirable to have a system with less computational load and less memory and battery requirements even though there is a little sacrifice on the accuracy. Even if a subscriber gets a spam sms not filtered by the system occasionally at the rate of say 5-10 % that would not be as much a problem as a very complicated SPAM filter which is consuming resources at a high rate.

The approach presented here is based on machine learning, and the simplicity of the approach is that it requires only four features extracted from the SMS messages. It has been shown empirically that these four features are

***Corresponding Author:** Dr. Ghulam Mujtaba, Assistant Professor, Electrical Engineering Department, Comsats Institute of Information Technology, Abbottabad, Pakistan. Email: gmujtaba@ciit.net.pk.

sufficient to filter the spam sms from the non-spam or 'ham' messages. It will be able to do the classification in real-time.

2. PROPOSED METHODOLOGY

This work is based on machine learning approach for classification of the incoming SMS at the mobile phone device itself. Since the SMS spam is different from the more familiar email spam in several aspects i.e. it does not contain a mailing list of recipients, the message is less than 160 bytes only and the mobile spam filtering system is to be implemented on resource limited mobile phones, in terms of processing as well as the battery life.

Considering these differences, a mobile spam filtering scheme is designed that meets four requirements. It operates at the access layer of a mobile phone, so that a spam SMS is silently moved to the spam folder without any intervention of the user. It is trained with relatively small number of SMS messages (of the order a few thousands. It has a high detection rate with minimum false alarm rate, and its resource requirements, memory and processing powerare significantly small to make it suitable for deployment on resource constrained mobile phones.

Effectiveness of the filtering framework is evaluated on a large SMS message collection including legitimate and spam messages. Following the evaluation, remarkably accurate classification results are obtained for both spam and legitimate messages.

2.1 Collection of SMS data

First a dataset of real SMS messages was collected using several mobile networks with a number of volunteers who agreed to share their SMSes for this work. This way, about 6600 SMSes including both Spam and Ham messages was compiled. This collection of SMS was then used in the experiments. The following table 3.1 is a portion of the messages bank.

Sr.#	SMS Status	SMS text
1	**Ham**	Fine if thats the way you feel.
2	**Spam**	Jazz Karaoke! Mobilink Jazz presents a unique technology with which you can sing along with your favorite songs.
3	**Ham**	I'm going to try for 2 months ha ha only joking
4	**Ham**	Just forced myself to eat a slice. I'm really not hungry tho. This sucks
5	**Ham**	Lol you are always so convincing.
6	**Ham**	K tell me anything about you.
7	**Spam**	*Jazz New SIM Offer* New family members of Jazz can enjoy 100 FREE Minutes and 100 free sms on daily usage of just Rs. 10+tax.
8	**…**	…….
9		**And so on**

Table-1 Aportion of the SMS dataset

2.2 Feature Extraction

The next step is to obtain distinctive features from the messages, both spam and ham. In order to keep the system simple and fast, the following features were selected for classification. The results show that these features are distinctive enough to tell the spam and ham messages apart. Each message in represented by a set of these four features.

1. No. of characters each SMS contain: This feature is a count of the characters contained in a message, based on the observation that spam messages have more characters usually than ham.

2. Matching spam words. In a list the words, say 50 words, which have highest frequency in the spam messages is obtained. Then a match of the words in the subjected message is made to the words in the list. A single match would render this feature as 1; no match found would return 0 on this feature. The frequent words list looks like:

{'FREE! Discount jazz indigo vodafone @PKR @USD +tax www offer…..'}

3. Match found in a digram list: In the next feature, combination of words in the given SMS are matched against a list of frequently occurring digrams obtained from the sms bank used for training. The digram list looks like:

{'Activate now| Advertisement portal| SMS portal| SMS alert|Buy now| Dial now|freesms| free minutes| friends & Family| GPRS users| Non-GPRS users| on-net| off-net| reply with| |/minute| MMS Package| plus tax|/mint| to subscribe| for subscribtion||call customer| free camera| free mobile| free message| O2 customers | Unlimited calls| Unlimited SMS…..'}

4.SMS Class (Output attribute)

- 1 means Spam SMS
- 0 means Ham/Valid SMS

2.3 Applying Machine Learning:

Several algorithms are investigated for supervised machine learning based classification of an SMS as a spam or ham message. In supervised learning the data used for training of the algorithm is labeled as to which class it belongs. Using the labeled data, the algorithm learns the relationship between the feature sets and the output, and hence it then classifies the unlabeled data from the learned relationship.

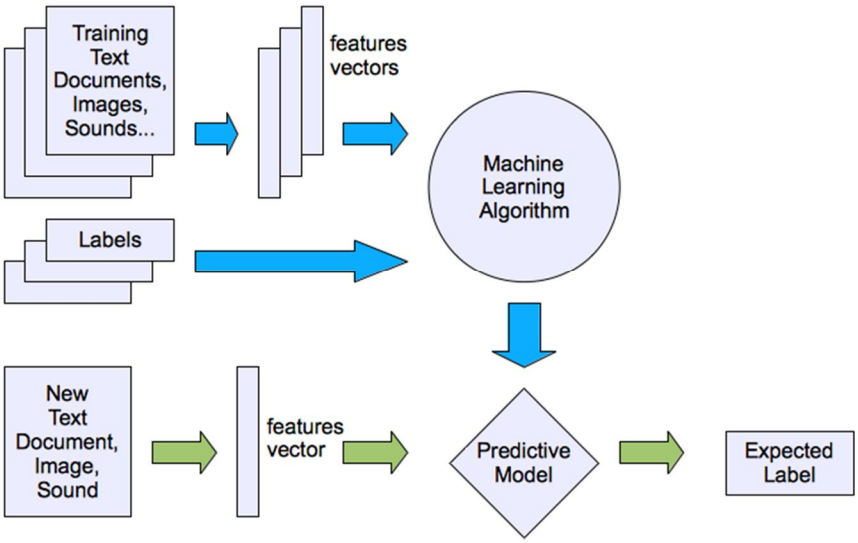

Fig1: Supervised Learning Schematic

3. RESULTS OF MACHINE LEARNING ALGORITHMS

In this section, the results of applying three different machine learning algorithms are presented. In this chapter, it is empirically demonstrated that this data can be utilized for the identification of the SPAM messages using machine learning classification algorithms or classifiers. Various classifiers were used at this stage and their performance was analyzed. WEKA (Waikato Environment for Knowledge Analysis) [8] machine learning software is used because it provides many different algorithms for data mining and machine learning which are easily useable by people who are not data mining specialists, besides being an open source and freely available software under GNU public license [9]. The classifiers used in the experiments are: naïve Bayes, Neural Network (Multilayer Perceptron) and C4.5 decision Tree. These classifiers were taken simply to be representative of a group/class of supervised classifiers in the WEKA software. Since their performance proved satisfactory for the requirements of this work, investigating the use of other classifiers was left for future work. From the available data, 66 % of the dataset is selected for training of the machine learning algorithm and the remaining set is used for the testing of the algorithm.

3.1 Classification using Bayesian Algorithm

This method has been proposed in 1998 to detect spam which is a relevant event based rule and the possibility of an event which would occur in the future can be inferred from previous occurrences of that event, as a result it can be applied to text classification [15]. NaiveBayes in Weka implements the probabilistic Naïve Bayes classifier. A Bayes classifier combines prior knowledge with observed data to assign a posterior probability to a class based on its prior probability and its likelihood given in the training data. A Naive Bayes classifier assumes conditional independence between attributes and assigns the maximum aposterior probability (MAP) class to new instances.

Summary

Total Number of Instances 2244

Attributes: 4

- No. of Char

- SPAM Words matched (Y/N)
- Combination of SPAM words Matched (Y/N)
- SMS Status (Output attribute)

Correctly Classified Instances	2040	92.953 %
Incorrectly Classified Instances	204	08.047 %

3.2 Classification using Multilayer Perceptron Algorithm

The Multilayer Perceptron is a neural network architecture that trains using back propagation and classifies instances. In [10], it is stated that "A multilayer perceptron is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. It is a modification of the standard linear perceptron in that it uses three or more layers of neurons (nodes) with nonlinear activation functions, and is more powerful than the perceptron in that it can distinguish data that is not linearly separable." Feed forward means that data flows in one direction from input to output layer through one or more hidden layers (forward). This type of network is trained with the back propagation learning algorithm. MLPs are extensively employed for pattern classification, prediction and approximation. The network used here consists of three layers, an input layer which has 3 input nodes corresponding to the 3 attributes of the training data. The 4th feature is the class or output attribute. The network has one hidden layer and an output layer with 10 nodes. In Weka, the default value for hidden layers is 'a', which means one hidden layer, with the number of nodes being the sum of input nodes and output nodes divided by 2. So in this experiment there are 2 nodes in the hidden layer, because the number of input nodes is 3 and the number of output nodes is 1 whose sum is 4. Some other default parameters for this network are: learning Rate -- The amount the weights are updated is 0.3, momentum -- Momentum applied to the weights during updating is 0.2, no. of epochs or passes through training data is set to 500 by default. The nodes in this network are all sigmoid. The details of these configuration parameters can be found in [11] by Witten and Frank. The summary of results obtained by using the MLP algorithm is given below:

Attributes: 4

Total Number of test Instances	2244	
Correctly Classified Instances	2004	89.3048 %
Incorrectly Classified Instances	240	10.6952 %

3.3 Classification using Weka Decision Tree

The C4.5 classifier is under the trees category of the WEKA classifiers. The C 4.5 Decision Tree was developed by Ross Quinlan [12].This is a divide and conquer algorithm like the other tree based algorithms. The data is partitioned recursively until every leaf has only the instances of one class or until further partitioning is impossible, when two cases have same values for each attribute but their class is different. Hence, if there are no conflicting cases, the decision tree will be able to classify every training instance correctly, which is "over-fitting", which generally leads to loss of prediction accuracy in most applications[13]. The over fitting problem is overcome usually by removing some of the structure of the decision tree after it has been produced, also known as pruning or sometimes by a stopping condition that prevents some cases from being subdivided. "After a decision tree is produced by the divide and conquer algorithm, C4.5 prunes it in a single bottom-up pass." [14] Here the Weka implementation of C4.5 algorithm, also known as J48 decision tree is used for this data set. The result of applying the decision tree algorithm is given below:

Total Number of Instances	2244
Correctly Classified Instances	2004 (89.3048 %)
Incorrectly Classified Instances	240 (10.6952 %)

4. Conclusions

From the results, the NaiveBayes algorithm has better performance than other two. This simple scheme can be programmed into an application which can reside on a smart mobile phone device. Hence a spam filter approach which is lightweight and simple is introduced in this work which has an accuracy of around 93%. In further extension of this work, other machine learning algorithms can also be explored and compared with the three given in this work.

Acknowledgment:

The authors declare that they have no conflicts of interest in this research.

REFERENCES

1. Spam - Definition and More from the Free Merriam-Webster Dictionary. Merriam-webster.com. 2012-08-31. Retrieved 2013-07-05.
2. http://abcnews.go.com/blogs/technology/2012/08/69-of-mobile-phone-users-get-text-spam/.
3. Text4ever. White Paper: UK spam study, Oct. 2009. http://www.txt4ever.com/study/spamstudy.pdf
4. QianXu; Xiang, E.W.; Qiang Yang; Jiachun Du; JiepingZhong, SMS Spam Detection Using Noncontent Features, Intelligent Systems, IEEE , vol.27, no.6, pp.44,51, Nov.-Dec. 2012
5. Wang, C et. all (2010), A behavior-based SMS antis pam system, IBM Journal of Research and Development, 3:1-3:16
6. Rick L. Allison, & Peter J. Marsico, US Patent Document -6S19932 Methods and systems for preventing delivery of unwanted short message service (SMS) messages, (Nov 2004).
7. Paul Graham, (August 2002), A plan for spam,http://paulgraham.com/spam.html
8. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* 11, 1 (November 2009), 10-18.
9. http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf, last accessed on 14th December, 2010
10. Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function Mathematics of Control, Signals, and Systems (MCSS), 2(4), 303–314.
11. Ian H. Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) June 2005
12. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
13. Quinlan, J. R. 1986. Induction of decision trees. Machine Learning, 1, 81-106.
14. Patil, Purushottam R., Revankar, Pravin and Joshi, Prashant,the Application of Data Mining for Direct Marketing, 2009.
15. Beyrami, Sisi and Derakshi, Review of Some Machine Learning Algorithms for Spam Filtering. Journal of Basic and Applied Scientific Research, 2013, 3(1s):25-30.