SMS Spam Detection

Saurabh Jaiswal, Saurabh Jain, Saurabh Verma
Computer Science and Engineering Dept., IIT Roorkee

Abstract - The growth of mobile phone users has lead to a dramatic increasing of SMS spam messages. Recent reports clearly indicate that the volume of mobile phone spam is dramatically increasing year by year. In practice, fighting such plague is difficult by several factors, including the lower rate of SMS that has allowed many users and service providers to ignore the issue, and the limited availability of mobile phone spam-filtering software. Probably, one of the major concerns in academic settings is the scarcity of public SMS spam datasets, that are sorely needed for validation and comparison of different classifiers. Moreover, traditional content-based filters may have their performance seriously degraded since SMS messages are fairly short and their text is generally rife with idioms and abbreviations. In this paper, we use a new real, public and non-encoded SMS spam collection that is the largest one as far as we know. Additionally, we compare the performance achieved by several established machine learning techniques. In summary, the Support Vector Machines(linear SVM) outperforms other evaluated techniques and, hence, it can be used as a good baseline for further comparison.

Index Terms—Mobile phone spam; SMS spam; spam filtering; text categorization; classification

I. INTRODUCTION

Short Message Service (SMS) is the text communication service component of phone, web or mobile communication systems, using standardized communications protocols that allow the exchange of short text messages between fixed line or mobile phone devices. They are commonly used between cell phone users, as a substitute for voice calls in situations where voice communication is impossible or undesirable. Such way of communication is also very popular because in some places text messages are significantly cheaper than placing a phone call to another mobile phone. SMS has become a massive commercial industry since messaging still dominates mobile market non-voice revenues worldwide. According to Portio Research[1], the worldwide mobile messaging market was worth USD 179.2 billion in 2010, has passed USD 200 billion in 2011, and probably will reach USD 300 billion in 2014. The same study indicates that annual worldwide SMS traffic volumes rose to over 6.9 trillion at end-2010 to break 8 trillion by end-2011. The increasing popularity of SMS has led to messaging charges dropping below US$ 0.001 in markets like China, and even free of charge in others. Furthermore, with the explosive growth in text messaging along with unlimited texting plans it barely costs anything for the attackers to send malicious messages. This combined with the trust users inherently have in their mobile devices makes it an environment rife for attack. As a consequence, mobile phones are becoming the latest target of electronic junk mail, with a growing number of marketers using text messages to target subscribers. SMS spam (sometimes also called mobile phone spam) is any junk message delivered to a mobile phone as text messaging. Although this practice is rare in North America, it has been very common in some parts of Asia. According to a Cloudmark report[2],

[1] http://www.portioresearch.com/MMF11-15.html

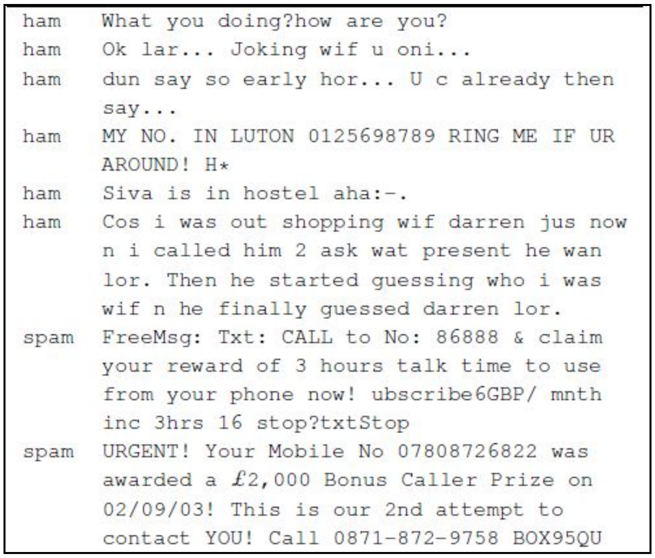
2. http://www.cloudmark.com/en/article/
mobile-operators-brace-for-global-surge-in-mobile-messaging-abuse

3. http://news.cnet.com/8301-1009_3-57494194-83/
protect-yourself-from-smishing-video/

the amount of mobile phone spam varies widely from region to region. For instance, in North America, much less than 1% of SMS messages were spam in 2010, while in parts of Asia up to 30% of messages were represented by spam. The same report reveals that financial fraud and spam via text messages is now growing at a rate of over 300 percent year over year. In fact, in a more recent report[3] by the same firm, it is stated that about 30 million smishing (SMS Phishing) messages are sent to cell phone users across North America, Europe, and the U.K. Smishing is part of the much larger SMS spam problem. In the U.S. alone, there has been and almost 400 percent increase in unique SMS spam campaigns in the first half of the year 2012. Besides being annoying, SMS spam can also be expensive since some people pay to receive messages. Moreover, there is a limited availability of mobile phone spam-filtering software and other concern is that important legitimate messages as of emergency nature could be blocked. In this paper we will be comparing several machine learning methods(namely Linear SVM and SGD Classifier with different set of features) on the basis of their accuracy on predicting whether a SMS is Spam or genuine.

II. The Dataset

- A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.
- A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.
- A list of 450 SMS ham messages collected from Caroline Tag's PhD Thesis.
- Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages.



ham	What you doing?how are you?
ham	Ok lar... Joking wif u oni...
ham	dun say so early hor... U c already then say...
ham	MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham	Siva is in hostel aha:-.
ham	Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam	FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam	URGENT! Your Mobile No 07808726822 was awarded a £2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Fig 1.Examples of messages present in the SMS Spam Collection

III. Methodology

We have generated a feature vector of seven features to train and test the machine learning algorithms.

The various features used are listed below :-

1. **Number of characters** in a SMS.
2. **Number of unique character** in a SMS.
3. **Weighted unique characters**:- The ratio of unique number of chars in sms divided by total number of chars in SMS.
4. **Word count** :- Number of words in a SMS.
5. **Is repeated** :- if a word in the sms has frequency greater than 0.5 then it will return 1 else 0.
6. **Repetition count top 3** :- It will return the

sum of frequencies of top 3 most occurred words in a SMS.

7. **Length of longest word** :- It will give the length of the longest word.

The Machine learning methods we used are :-

1. Linear SVM classifier.
2. Stochastic Gradient Descent (SGD) classifier.

Initially we tested these two classifiers with only first five features.The accuracy we got from both methods was quite impressive, then we added the two new features (6 and 7), the accuracy of Linear SVM was unaffected by the increase in features but the accuracy of SGD classifier improved a little bit.

Then we used *term frequency/inverse document frequency* (TF-IDF) representation.

Assume that there are m documents and F_{ij} is the number of times the word k_i appears in document dj

. The normalised term frequency is defined as:

$$\mathrm{TF}_{i,j} = \frac{F_{ij}}{\max_z F_{zj}},$$

where the maximum is computed over all frequencies F_{zj} for document dj. In other words, for a particular document we find the number of occurrences of the most frequent keyword and then normalise the frequencies of the keywords by this count. Frequent keywords may not be useful and hence one also uses inverse document frequency, defined for a keyword k_i as

$$\mathrm{IDF}_i = \log \frac{m}{n_i},$$

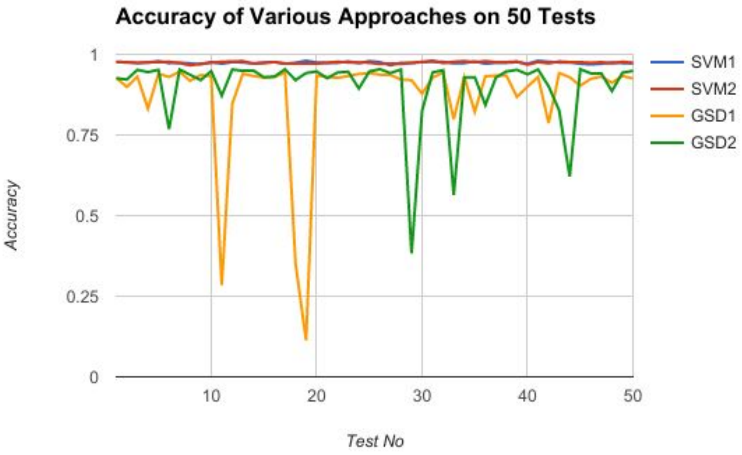
where n_i is the number of times keyword ki occurs in all documents and log is the logarithm function. In words, the inverse document frequency is the inverse of the proportion of times a certain keyword occurs in all documents (a value between 0 and 1), and then the logarithm of that quantity. Putting the parts together, the TF-IDF weight for a keyword k_i in document d_i is then $X_{ji}=TF_{i,j}$×IDF_i and each document can be represented using a vector of keyword weights. In this way similar vectors correspond to similar documents and the representation is now ready for a variety of machine learning algorithms.

The accuracy we got when we trained SVM from this data the accuracy was higher than the previous approaches.

IV. Experiments Results

The accuracy of various approaches on 50 Tests (80% Training and 20% Testing Data) is shown in the graph below.

0

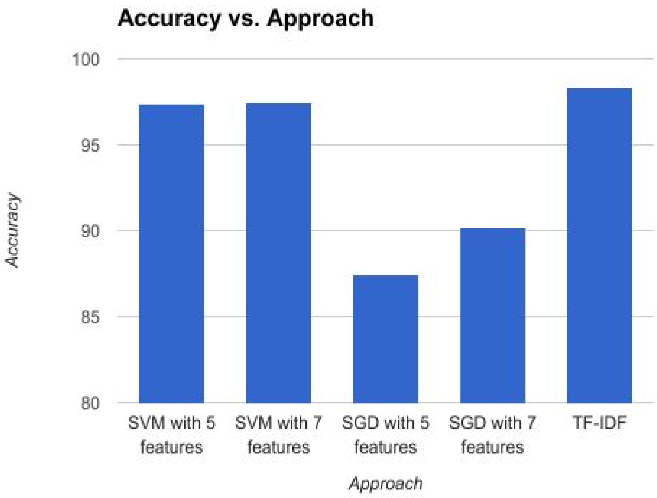


Graph 1 : Accuracy of SVM1, SVM2, SGD1, SGD2 on 50 Tests generated by dividing Data Set into Training (80%) and Testing (20%) Sets randomly 50 times.

The overall average accuracy of all approaches is tabulated in Table 1 and showed in Graph 2.

Approach	Accuracy
SVM with 5 features	97.4 %
SVM with 7 features	97.5 %
SGD with 5 features	87.5 %
SGD with 7 features	90.2 %
SVM with TF-IDF	98.3 %

Table 1 : Overall Average Accuracy of all approaches



Graph 2 : Overall Accuracy of Various Approaches

V. Conclusion

We observe that the accuracy in SVM didn’t increase significantly (increase of 0.1%) by adding 2 extra features. But in case of SGD the accuracy increases by 2.7 % after adding 2 extra features. So we conclude in case of SVM lesser no of features may give a good accuracy but in SGD we require more features to get better accuracy.

The best classifier turns out to be TF-IDF with 98.3% which is also justifiable by the fact the frequency of a particular word is taken into account while classifying. Because spam messages tend to have some specific words common like reward, claim, prize etc.

Acknowledgment

We would like to thank Professor Paarth P. Roy, IIT Roorkee for his valuable guidance and support during this course of this paper.

References

[1] T. Almeida, J. Gómez Hidalgo, and A. Yamakami, “Contributions to the Study of SMS Spam Filtering: New Collection and Results,” in Proceedings of the 2011 ACM Symposium on Document Engineering, Mountain View, CA, USA, 2011, pp. 259–262.

[2] J. M. Gómez Hidalgo, T. A. Almeida, and A. Yamakami, “On the Validity of a New SMS Spam Collection,” in Proceedings of the 2012 IEEE International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 2012, pp. 240–245.

[3] J. M. Gómez Hidalgo, “Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization,” in Proceedings of the 17th ACM Symposium on Applied Computing, Madrid, Spain, 2002, pp. 615–620.

[4] L.Zhang,J.Zhu,and T.Yao,“An Evaluation of Statistical Spam Filtering Techniques,” ACM Transactions on Asian Language Information Processing, vol. 3, no. 4, pp. 243–269, 2004.

[5] G. Cormack, “Email Spam Filtering: A Systematic Review,” Foundations and Trends in Information Retrieval, vol. 1, no. 4, pp. 335–455, 2008.