International Journal of Information Processing, 8(4), 68-78, 2014
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

RePC-SSMSM: Repetitive Preprocessing and Clustering Approach for Filtering Spam SMS Messages using Naive Bayes Classifier

Asha S Manek[a], P Deepa Shenoy[b], M Chandra Mohan[a], K R Venugopal[b], L M Patnaik[c]

[a]Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India, Contact: ashas100@gmail.com.

[b]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001 India.

[c]Honorary Professor, Indian Institute of Science, Bangalore, India.

With the use of internet over mobile phones, SMS spam is one of the text messaging forms among the many forms of spams which are increasing day by day. As the popularity of mobile phones surged in the early 2000s, frequent users of text messaging began to see an increase in the number of unsolicited commercial advertisements being sent to their mobile phones through text messaging, so this can be particularly infuriating for the recipient. Hence it is important to filter these spam SMS messages. The proposed model RePC-SSMSM filters SMS spam using repetitive preprocessing and clustering approach and the performance has been evaluated across the identified parameters against other existing models. Our results shows 99.73% accuracy using Naive Bayes classifier, thus demonstrating the efficiency of the proposed technique over other models in this area of research.

Keywords : Clustering, Naive Bayes Classifier, Pre-Processing Technique, RePC-SSMSM, SMS Spam.

1. INTRODUCTION

Short Message Service (SMS) is a most popular medium of interaction and communication today for the mobile users to communicate with friends, family and acquaintances. Mobile phone spam is described as mobile spamming, SMS spam, text spam, m-spam or mspam [1]. With the increase of various smart phones and with the rapid development and wide usage of various apps, SMS (Short Message Service), SPIM (Spam on Instant Messaging Services) on the mobile internet, spam increases explosively.

Today, the mobile messaging channel is considered as "clean" and "trusted" in most parts of the world. The level of trust means almost all messages received by subscribers are opened and read and because of the ease of use of Smartphones, numbers are easily dialed or links clicked on further exposing the subscriber to risk. Also, unlimited free text plans of sending SMS and the various methods of billing available made mobile messaging system exposed to and very attractive target for spammers. While SMS spam is not perceived as a major issue in some regions it now constitutes 20-30% of all SMS traffic in Asian markets such as China and India. Both countries have regulated to restrict the number of messages each subscriber can send in one day but this is not containing the problem. As the economic benefits of SMS spam continues to grow, unguarded networks being embattled first by the spammers. Previously unaffected markets will suffer an increase in attacks by using advanced methods and techniques to avoid detection. A Spam message has gone beyond the extent of simple spam messages to fraudulent scams, mobile viruses, phishing, spyware and

6. CONCLUSIONS AND FUTURE WORK

In this work, we proposed repetitive preprocessing and clustering technique for filtering spam SMS. Our main goal is to find the 10 keywords using clustering method from both ham and spam messages present in the dataset. We evaluate the usefulness of these features in spammer detection using traditional classifiers like W-Random Forest, Naive Bayesian, Support Vector Machine, W-Random Tree and W-IBK schemes using the publicly available SMS Spam Corpus v.0.1 and SMS Spam Corpus v.0.1 Big dataset we have collected. The results and the performance analysis show the best performance is achieved 99.73% and 98.48% accuracy with Nave Bayes classifier respectively. Based on our dataset, our features provide slightly better classification results when compared to those suggested in [17] or [18].

The dataset does not contain the phone number of the sender who sends the spam messages. So, in future we can use this information along with network they use in detecting SMS spam. We also hope to include larger dataset for evaluation as well as wall-post datasets from other online social networking sites.

REFERENCES

1. http://en.wikipedia.org/wiki/Mobile_phone_spam
2. Accident Claim Text Scam. Kathirvel.com. 7 July 2010. Retrieved 29 March 2012.
3. NY Times article on UCAN case against Sprint
4. UCAN report on Sprint SPAM SMS settlement. Ucan.org. 5 October 2006. Retrieved 29 March 2012.
5. Narayan, Akshay and Prateek Saxena. The Curse of 140 Characters: Evaluating the Efficiency of SMS Spam Detection on Android, *In Proceedings of the Third ACM workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 33–42, 2013.
6. Coskun, Baris and Paul Giura. Mitigating SMS Spam by Online Detection of Repetitive Near-Duplicate Messages, *In IEEE International Conference on Communications (ICC)*, pages 999–1004. 2012.
7. Uysal, Alper Kursat, S Gunal, Semih Ergin and E Sora Gunal. Detection of SMS Spam Messages on Mobile Phones, *In 20th Conference on Signal Processing and Communications Applications (SIU)*, pages 1–4, 2012.
8. Jiang, Nan, Yu Jin, Ann Skudlark and Zhi-Li Zhang. Greystar: Fast and Accurate Detection of SMS Spam Numbers in Large Cellular Networks Using Gray Phone Space, *In USENIX Security*, pages 1–16, 2013.
9. Gomez Hidalgo, Jose Maria, Guillermo Cajigas Bringas, Enrique Puertas Sanz and Francisco Carrero Garcia. Content Based SMS Spam Filtering, *In Proceedings of the 2006 ACM symposium on Document Engineering*, pages 107–114, 2006.
10. Mahmoud, Tarek M, and Ahmed M Mahfouz. Sms Spam Filtering Technique Based on Artificial Immune System, *IJCSI International Journal of Computer Science Issues*, 9(1), 2012.
11. Shirani-Mehr, Houshmand. SMS Spam Detection using Machine Learning Approach.
12. Delany, Sarah Jane, Mark Buckley and Derek Greene. SMS Spam Filtering: Methods and Data, *Expert Systems with Applications*, 39(10):9899-9908, 2012.
13. Xu, Qian, Evan Wei Xiang, Qiang Yang, Jiachun Du and Jieping Zhong. SMS Spam Detection using Noncontent Features, *IEEE Intelligent Systems*, 27(6):44–51, 2012.
14. http://www.esp.uem.es/jmgomez/smsspamcorpus/.
15. Manning, Christopher D, Prabhakar Raghavan and Hinrich Schtze. Introduction to Information Retrieval, *Cambridge: Cambridge University Press*, 1, 2008.
16. Ted Dunning. Statistical Identication of Language. Technical report, Computing Research Lab, New Mexico State University, 1994.
17. Valles, Enrique and Paolo Rosso. Detection of Near-Duplicate User Generated Contents: The SMS Spam Collection, *In Proceedings of the 3rd ACM International Workshop on Search and Mining User-Generated Contents*, pages 27–34, 2011.
18. Almeida, Tiago, Jos Mara Gmez Hidalgo and Tiago Pasqualini Silva. Towards SMS Spam Filtering: Results under a New Dataset, *International Journal of Information Security Science*, 2(1):1–18, 2013.

Asha S Manek received the B.E (EC) degree from Nagpur University in 1993. She completed her M.E degree in Information Technology from University Visvesvaraya College of Engineering, Bangalore University, Bangalore in 2008. She is pursuing her Ph.D from JNTUH University, Hyderabad in Computer Science and Engineering under the guidance of Dr. P Deepa Shenoy and Dr. M Chandra Mohan. She has published five research publications in various International Conferences.

Dr P Deepa Shenoy is currently working as a professor in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore. She did her doctorate in the area of Data Mining from Bangalore University in the year 2005. Her areas of research include data mining, soft computing, biometrics and social media analysis. She has published more than 100 papers in refereed International conferences and journals. She is also a senior member of IEEE and currently serving as student activity chair, IEEE Bangalore section.

Dr M Chandra Mohan received the B.E. (EEE) degree from Osmania University in 1994. He worked as an Assistant Engineer in AP State Electricity Board(APSEB) for 7 years (1994-2001). Mean while he completed his MTech. (CSE) from Osmania University in 2000 by availing study leave from APSEB. He is working in JNT University since 2001. Presently he is working as a Professor in Dept of CSE in JNTUH College of Engineering Hyderabad, JNT University Hyderabad. He is the recipient of 3 Gold Medals from Osmania University at graduate level by securing University first rank. He received his Ph.D degree in Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTUH) in 2010. His research interests includes Software Engineering and Image Processing.

Venugopal K R is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D in Economics from Bangalore University and Ph.D in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 51 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems *etc.*. During his three decades of service at UVCE he has over 400 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.

L M Patnaik is currently Honorary Professor, Indian Institute of Science, Bangalore, India. He was a Vice Chancellor, Defense Institute of Advanced Technology, Pune, India and was a Professor since 1986 with the Department of CSA, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute he has over 700 research publications in refereed International Journals and refereed International Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to High Performance Computing and Soft Computing. His areas of research interest have been Parallel and Distributed Computing, Mobile Computing, CAD for VLSI circuits, Soft Computing and Computational Neuroscience.