Investigating Topic Models for Mobile Short Messaging Service Communication Filtering

Abiodun. Modupe, Oludayo O. Olugbara, Sunday O. Ojo

Abstract — **This research work investigates the use of Latent Dirichlet Allocation (LDA), a generative topic modeling technique to extract latent features arising from mobile Short Messaging Service (SMS) communication for automatic discovery of user interest. This involves integrating temporal ordering of SMS into a generative process in an iterative manner. The mobile SMS documents are partitioned into segments, wherein the discovered topics in each segment are propagated to influence the discovery of latent features. The proposed technique filters malicious mobile SMS communication and shows that topic models can effectively detect distinctive latent features to support automatic content filtering overtime. The practical implication of SMS communication filtering is apparent in designing systems that proactively detect information security threats to mobile subscribers and operators. The system can assist in optimum decision making, for instance in a scenario where an imposture attempts to sneak confidential information from unsolicited messages send to a subscriber or an operator.**

Index Terms— **Internet, Mobile SMS, SMS Spam, Latent Dirichlet Allocation, filtering.**

I. Introduction

The principal objective of this research work is to investigate the use of Latent Dirichlet Allocation (LDA) topic modelling technique for filtering malicious messages communicated through mobile SMS. LDA is a generative model that integrates conversations as mixture of topics of interest. The advent of Information Technology (IT) and increasing computational power of mobile devices have made mobile messaging service market to rapidly grow as a profitable business for mobile operators. Text messaging is the most widely used mobile data service on the planet with 72% being an active users of the Short Messaging Service (SMS) worldwide [1][2]. The users are switching from traditional communication such as chatting, blogging and email to short text messages, which are promptly available on most mobile communication devices. Unlike traditional postal services, which could take several days for an addressee to acknowledge receipt, SMS based communication is done swiftly. The security of mobile SMS has been under attack because of malicious activities including phishing attack, online buying, cyber stalking, mobile SMS spam, spoofing, pornography and many SMS-related scams. This has posed global security challenges because the occurrence of these malicious activities has negative security consequences for individuals. The creativity and ingenuity of cyber attackers engaging in social engineering techniques to decoy targeted individuals to reveal private information or incur charges such as revenue linkage have demanded research effort to find optimal ways of countering malicious activities.

The public interest is generally focused on detecting phishing emails [3] and unwanted emails [4], but many of the cellular operators are also concerned with SMS filtering to effectively secure consumer cellular bandwidth. Mobile SMS spam is now ubiquitous because of mobile and internet technologies interconnectivity that enables communication between diverse users of internet systems such as Google, Facebook and Skype. This provides a platform for mobile spamming to occur. Mobile spamming requires proactive solution that exploits the rising computational power of third generation mobile devices. As a result, content filtering techniques based on Bag-of-Word (BoW) [5] [6] have been successfully exploited to filter email messages that are typically in large magnitude.

The filtering of spam messages can significantly contribute to improving the security of mobile SMS and practically increase trust of communication using emerging technologies. The fundamental prevailing research question to solve is "are there enough features in mobile messaging that can be used to identify malicious text messages leading to better messaging personalization?" This necessitates for efficient feature discovery in SMS documents.

II. LDA Model for Mobile SMS Filtering

The model reported in this paper borrows inspiration from the LDA [7]. In the original LDA document collections are represented as a mixture of topics of interest. In the proposed model, unobservable "hidden" features are segments of words that appear concurrently in SMS documents. The documents are tokenized as a stream of words, phrases and symbols in training dataset. This forms a

Manuscript received March, 2013; revised April 10, 2013.

A. Modupe received B.Sc. (Hons) Computer Science with Mathematics and MTech Degree in Information Technology from Department of Information, Tshwane University of Technology, Pretoria. South Africa. He is currently a doctoral student at Tshwane University of Technology. (email: modupea@tut.ac.za).

Oludayo, O. Olugbara received B.Sc. (Hons) in Mathematics with cum laude, M.Sc. in Mathematics with specialization in Computer Science and PhD in Computer Science. He joined Department of Information Technology, Durban University of Technology, Durban, South Africa, where he is currently a Professor of Computer Science and Information Technology. He is a member of Marquis Who's Who in the World, USA and director of research at KZN e-Skills Co-Lab at Durban University of Technology. (e-mail: oludayoo@dut.ac.za).

Sunday, O. Ojo received B.Sc. (Hons) in Computer Science and PhD in Computer Science from University of Glasgow. Dean, Faculty of Information and Communication Technology and Research Professor at the Tshwane University of Technology. He has recently completed a National Innovation Database System Research and Development project, which involved research into a system of Science and Technology Research and Innovation (NSI) and developing an open source relational database system. (email: ojoso@tut.ac.za).

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

number of K dissimilar collection named topics. Each segment of words is typically assigned as topics based on the probability distribution. This implies that each document can have a number of different topics. The assumption of the prevalent topics in LDA is conceptually similar to Author Topic Model (ATM) [8], extending to use more memory to learn distributions from past observed sampling. It utilizes θ and ϕ distributions from previous iterations as a prior probability for subsequent iterations. This helps in classifying whether the message can be filtered successfully as SMS spam or normal SMS messages.

In generally, given a mobile user u and a number k of topics, SMS document of a user u is represented as a multinomial distribution ϕ_k over topics drawn from a Dirichlet prior with a hyper-parameter α. The distribution of topics is represented by a multinomial distribution β_k drawn from a Dirichlet prior with a hyper-parameter β. The generative process of LDA extracts words from SMS document d of a mobile user u to represent a topic $z_{d,u} \in \{1, \cdots\cdots, K\}$ drawn from θ_d. The word extracted from the user document is represented by $w_{d.,u} \in \{1, \cdots\cdots, V\}$, which is also drawn from the distribution $\beta_{z_{d,u}}$. The probability distribution of an SMS document can under this assumption be estimated as the following joint distribution:

$$p(w, z, \theta, \phi | \alpha, \beta) = p(\theta|\alpha)\, p(z|\theta)\, p(\phi|\beta)\, p(w|z, \phi) \qquad (1)$$

The hidden features are marginalized to offer a more simplified model when a corpus w and the hyper-parameters α and β are given and therefore the objective functions of the model parameters and inferred distribution of the latent feature are computed as:

$$p(w|\alpha, \beta) = \int_{\phi} \int_{\theta} \sum_{z_{d,u}} \left(\prod_{u=1}^{U} \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{u,k}^{\alpha_{k+n_{u,k,\cdot}} - 1} \right) \times \left(\prod_{k=1}^{K} \frac{\Gamma(\beta_k)}{\coprod_{k=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \beta_{k,v}^{\beta_{k,v} + n_{\cdot,k,v} - 1} \right) \partial\theta \partial\phi \qquad (2)$$

The collapsed Gibbs sampling [9] is used to optimize the distribution when the number of SMS messages increases and the generated tokens are much higher than a typical document. This allows for a compact representation of the model from a high-dimensionality (when the parameters are large) by learning topics from low-dimensionality document and finding the latent distribution of each SMS message across the topics. That is each SMS message by a user is modelled as a topic-vector, where each dimension is the probability to emit the topic. The LDA algorithm is summarized in Table I as a probabilistic generative process based on Equation (2). It is important to note that word ordering in SMS message collection is not considered.

TABLE I
LDA ALGORITHM

Step1: For each topic *T* drawn from a multinomial distribution over words:

$$\phi_k \approx Dir(\alpha)$$

Step2: For each mobile SMS document d from user u

Step 2.1: Draw a vector of topic proportion with a Dirichlet prior:

$$\theta_d \approx Dir(\beta)$$

Step 2.2:

For each word i :

Step 2.2.1:

(i) Draw a topic index

$$z_{d,u_i} \approx Mult(\theta_d), z_{d,u_i} \in \{1, \cdots\cdots K\}$$

(ii) Draw a word vocabulary index of

$$w_{d,u_i} \approx Mult(\phi_{z_{d,u_i}}), w_{d,u_i} \in \{1, \cdots\cdots V\}$$

III. EXPERIMENTAL RESULTS

In this study, LDA is applied for mobile SMS filtering to provide insight into distinct activities of mobile users. These activities are inferred as a mixture of topics in the SMS collection [10]. Mobile SMS was explored by crawling 1084 legitimate and 1319 spam SMS messages from corpus available at University of California, Irvine (UCI) website (http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection). In addition, 82 spam SMS messages written in English were collected from Grumble text website (http:// mtaufiqnzz.wordpress.com/british-english-sms-corpora/). A total of 1,157 SMS joke and 199 SMS spam messages were collected from Vodacom (http://vodacom.co.za). This gives a total of 3642 plain text SMS messages used for experimentation. In all, innocuous SMS documents constitute 61.50% and 38.50% of the SMS documents constitute spam. Finally, stop-words and punctuations were removed from the SMS messages leaving the body contents for LDA algorithm to process all 3642 SMS documents. The output of the LDA gives 5 distinct segments of semantically related topics from 31929 distinct words obtained from the 3642 SMS documents. The topics are extracted from a single instance at the 200th iteration of the Gibbs sampling with a model distribution propagation parameter. For the extracted topics, illustration is made with the top 5 topical words conditionally generated on the topic for the corpus dataset. Table II shows the proportions of topics that were learned using LDA and the associated word instances as a realistic representation of the corpus dataset.

TABLE II
EVOLUTION OF SAMPLE TOPICS FOR LDA

Topic 1		Topic 2		Topic 3	
Word	Prob.	Word	Prob.	Word	Prob.
free	0.0385	good	0.0329	love	0.0367
text	0.0273	time	0.261	send	0.0328

Topic 4		Topic 5	
Word	Prob.	Word	Prob.
reply	0.0231	call	0.0952
make	0.0199	money	0.0633

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Inferring topic network from topics is significant to discover latent features that uniquely characterize malicious mobile SMS and cohorts that are involved in malicious SMS communication. The associations of topics are calculated according to the strength of words to topics by dividing number of words in a topic by total number of words in corpus. For example, if there were 100 instances of latent feature, say "money" and 13 of those instances were associated to topic 3, then money-to-topic 3 would have a weight of 0.13%. This is directly used as the connection between topics and document to create a word-topic-count as money-8-406. In this example, money has 406 instances in topic 8. This information is used to construct a topological topic network. Fig.1 shows the topic network with intuitive classification of the corpus document evaluated according to the compactness of each topic by the average path length (APL) = 4.31 [11], among the top 30 latent words in the topics. The topic network was constructed using Gephi 0.8.2, an interactive visualization and exploration platform for all kinds of networks, complex systems, dynamic and hierarchical graphs. This work considers topic-to-document and word-to-topic connections of greater than 10% for manageability and to improve statistical significance of the identified topic modules in Gephi. Five segments of words were identified as shows in the topic network with purple, yellow, pink, green and orange colour. The tinning coloured line corresponding to proportion of incidence of the word appearing in the connecting topic.

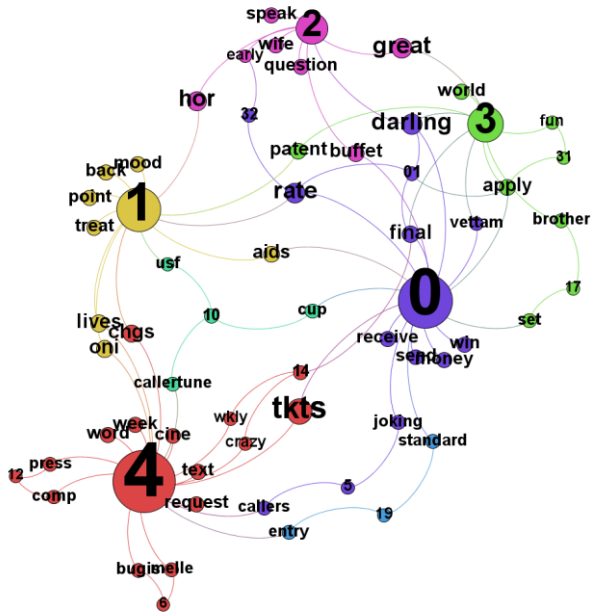

Fig. 1: Network topology to demonstrate relationship between topics.

IV. CONCLUSION

In this paper, a generative technique of documents classification is proposed to iteratively learn the curiosity of a user based on topic-word distributions. This was to filter spam SMS on independent mobile phone. The application of LDA to a sample repository of SMS corpus dataset shows that we can effectively filter malicious user dialogue themes. The ideas and methods presented in this paper will prove useful in identifying criminal communities that are related by co-occurrence and their interlinked subgroup.

REFERENCES

[1] Androulidakis, I. I. (2012). SMS Security Issues. Chap. 5 in Mobile Phone Security and Forensics. Springer briefs in Electrical and Computer Engineering, Springer US, 63-74.

[2] Tekelec. (2007). SMS Security: Malicious Attacks Are Just around the Corner. Are You Protected? 1-14. Morrisville, NC 27560 (USA): Tekelec.

[3] Bergholz, A., De Beer, J., Glahn, S., Moens, M. F., Paaß, G., & Strobel, S. (2010). New filtering approaches for phishing email. Journal of Computer Security, 18(1), 7-35.

[4] Modupe, A., Olugbara, O. O., & Ojo, S. O. (2012). Comparing Supervised Learning Classifiers to Detect Advanced Fee Fraud Activities on Internet. Advances in Computer Science and Information Technology. Computer Science and Information Technology, 87-100.

[5] Wu, L., Hoi, S. C., & Yu, N. (2010). Semantics-preserving bag-of-words models and applications. Image Processing, IEEE Transactions on, 19(7), 1908-1920.

[6] Cormack, G. V., María, J., Hidalgo, G., & Sánz, E. P. (2007). Spam Filtering for Short Messages.

[7] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of machine learning research, 3, 993-1022.

[8] Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence (pp. 487-494). AUAI Press.

[9] Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (pp. 306-315).

[10] Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. In International AAAI Conference on Weblogs and Social Media (Vol. 5, No. 4, pp. 130-137).

[11] Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25(2), 163-177.

ISBN: 978-988-19252-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)