# Intelligent SMS Spam Filtering Using Topic Model

Jialin Ma[1,2*], Yongjun Zhang[1,2], Jinling Liu[1], Kun Yu[1],
1. Huaiyin Institute of Technology
Huaian, China
2. Hohai University
Nanjing, China

XuAn Wang
Engineering University of CAPF
Xi'an, China

***Abstract*—Nowadays, spam messages have been overflowing in many countries. They seriously violate personal rights, and may even harm the national security. The existing filtering techniques usually uses traditional text classifiers, which are more suitable to deal with normal long texts; therefore, it often faces some serious challenges, such as the sparse data problem and noise data in the SMS message. This research work proposes a message topic model (MTM) for SMS spam filtering. The MTM derives from the famous probability topic model. Although the MTM is based on probability topic model, it is different from the famous standard Latent Dirichlet Allocation (LDA) in the following aspects: (1) For the purpose of overcoming the sparsity problem in SMS message classification, first, the standard K-means algorithm is used to classify the training data into rough classes, then, aggregates all the spam messages of a class into a single document. (2) Symbol semantics is taken in account. Some preprocessing rules and background terms are considered to make the model more appropriate to fully represent SMS spam. Finally, we compare the MTM with the SVM and the standard LDA on the public SMS spam corpus. The experimental results show that the MTM is more effective for the task of SMS spam filtering.**

***Keywords—SMS Spam; Topic Model; LDA; MTM***

## I. INTRODUCTION

SMS (short message service) Spam messages are unsolicited and unwanted messages, including advertisements, frauds, erotic services, etc. They disturb normal life, consume the resource of mobile communication equipment, and lead to congestion of communication network [1,6,7]. With the rapid growth of mobile users and business in the world, SMS has become not only an important way in daily work and life but also a massive commercial industry due to its convenience and inexpensive price [1,2]. At the same time, criminals abuse SMS to defraud and earn economic or political benefits. This situation has become more serious over the years, especially in India, Pakistan, and China where SMS lacks of effective supervisions [3,8]. Reports from the Chinese National Spam Report Center show that Chinese mobile phone users received 12 spam messages every week in average, and the total number of spam messages reached 120 million in 2014 [4]. In the USA, more than 69% of the surveyed mobile users claimed to have received text spam in 2012 [5,29].

Most of the current SMS spam filtering technologies are transplanted from email spam filtering [9-13]. They can be generally summed up as follows: (a) the high frequency of messages sending from phone numbers will be intercepted in the operators' server. However, this way has two fatal defects: can't deal with pseudo base-stations and possibly hinder the normal mass SMS messages; (b) users define black and white lists or sensitive keywords by themselves in their mobile phones' apps. But this would have negative impact on users experience [7,27]; (c) Content based SMS filtering is considered as one of the most effective method, but it restrict by the state of art rele-vant to text mining, machine learning, and Natural Language Processing (NLP) [3].

In our study, we present a Message Topic Model (MTM) which is based on the probability theory of latent semantic analysis, and more suitable for the task of SMS spam filtering. Compared with the existing SMS spam filtering technologies and standard topic model—Latent Dirichlet Allocation (LDA), the MTM can eliminate the sparsity problem in SMS message classification and pay attention to kinds of symbols which usually appears in SMS spam. Therefore, the MTM is fitter for SMS filtering.

## II. MESSAGE TOPIC MODEL (MTM)

In order to represent spam messages more fully, except for considering the symbol terms, we also divide the terms into topic terms and background terms in the spam. The similar inspiration for different applications has appeared in the works: [22,23], for different applications.

Conventional topic models, like PLSA[17] and LDA[15,18], reveal the latent topics within the text corpus by implicitly capturing the document-level word co-occurrence patterns. Therefore, directly applying these models on short texts will suffer from the severe data sparsity problem (i.e. the sparse word co-occurrence patterns in each short document[14,19,26,28].

For the purpose of overcoming the sparse problem in the SMS message, first we use unsupervised clustering algorithm—K-Means to cluster training spam messages into rough classes, then aggregate all the spam messages of a class as a single document. Word co-occurrence patterns will be captured in these aggregated pseudo-documents.

Fig. 1(a) shows the graphical model for the "standard topic model" (LDA). $D$ is the number of documents in the corpus and document d has Nd words. The process includes two steps: first, assign a topic number from document-topic distribution $\theta$; then, draw a word from topic-word distribution

$\varphi$. All documents share $T$ topics. Document-topic and topic-word distributions all obey the multinomial distributions, and each of them is governed by symmetric Dirichlet distribution. $\alpha$ and $\beta$ are hyper-parameters of symmetric Dirichlet priors for $\theta$

978-1-5090-4124-4/16 $31.00 © 2016 IEEE
DOI 10.1109/INCoS.2016.47

and $\varphi$. Parameters $\theta$ and $\varphi$ can be obtained through a Gibbs sampling.

Fig. 1(b) shows a similar general structure for SMS spam filtering, named message topic model (MTM). Let $\varphi$ denote the word distribution for topics and $\varphi_b$ denote the word distribution for background terms. Let $\theta$ denote the topic distribution for the whole corpus. Let $\lambda$ denote Bernoulli distribution which controls tIhe variable $x$ for the choice between background terms and topic terms. Variable $z$ and $w$ represent a topic number and a word respectively. $\Phi$, $\theta$, and $\varphi_b$ all obey multinomial distributions, each of term is drawn from symmetric Dirichlet ($\beta$), Dirichlet ($\beta_b$), and Dirichlet ($\alpha$) respectively. $\lambda$ is drawn from Beta ($\gamma$).

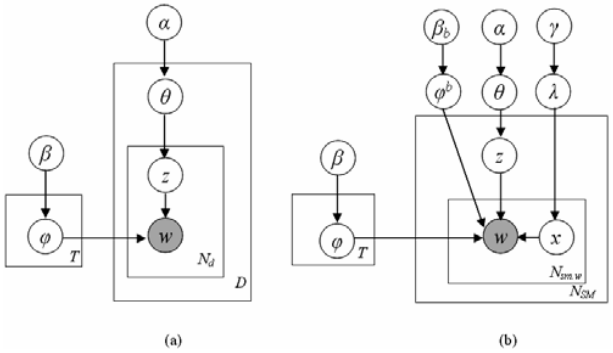


Fig. 1. Graphical models for (a) Latent Dirchlet Allocation, (b) Message Topic Model (MTM)

The conditional probability of a term $w$ in the SMS spam corpus can be described as:

$$p(w) = p(x=0)\sum_{z} p(z)p(w \mid z) + p(x=1)p(w \mid x=1) \qquad (1)$$

In Eq. 1, $w$ is topic term when $x$=0, $w$ is background term when $x$=1.

A SMS spam generative probability is expressed as:

$$p(\vec{w}) = \prod_{i=1}^{N_{sm.w}} (p(x=0)\sum_{z} p(z)p(w_i \mid z) + p(x=1)p(w_i \mid x=1)) \quad (2)$$

In Eq. 2, $N_{sm.w}$ denotes the number of terms in a SMS spam.

**Algorithm 1 MTM training**

Step1: initialization:.distribute '0' or '1' randomly for ever term in the training spam messages.
Step2: distribute a topic number randomly for these terms in a SMS Spam that has been distributed with '0'.
Step3: rescan the training data, for every word $w_i$,
if $x_i$=0, use formula(3) to sample and update $x_i$ and its topic number
if $x_i$=1, use formula(4) to sample and update $x_i$
Step4: repeat this process until Gibbs sampling convergence
Step5: Calculate all parameters of the MTM by these samples

The training process is mainly to acquire the samples ($z$, $w$) and ($x$, $w$) in accordance with the real corpus via Gibbs sampling. All the parameters in the model can be eventually calculated through sample statistics. The simple process of training is described as algorithm 1.

## III. Experimental Results

In order to evaluate the proposed method, we integrated tw o SMS datasets: SMS Spam Collection v.1 and DIT SMS Spa m Dataset. The first dataset comes from UCI machine learning repository, which has been used in relevant research reference [24]. In particularly, the details about this corpus have been intr oduced in the work references [2,20]. It can be downloaded fre ely in web site: http://archive.ics.uci.edu/ml/datasets/SMS+Spa m+Collection. DIT SMS Spam Dataset (http://www.dit.ie/com puting/research/resources/smsdata/) only has SMS spam data. Except for collecting messages from scraping two UK public c onsumer complaints websites, it is also include 639 SMS spam from SMS Spam Collection v.1. More information about DIT d ataset can be seen in the paper[3].The details about the two dat asets are shown in TABLE I.

TABLE I. ABOUT THE TWO DATASETS

| Dataset name | Hams | | Spams | | Total |
|---|---|---|---|---|---|
| *SMS Spam Collection v.1* | *4827* | *86.6%* | *747* | *13.4%* | *5574* |
| *DIT SMS Spam Dataset* | *0* | *0* | *1353* | *100%* | *1353* |

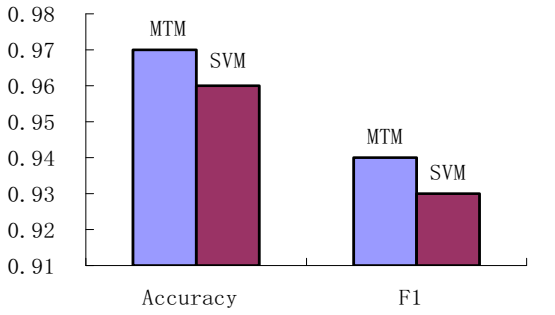


Fig. 2. F1 and accuracy for MTM and SVM

Symmetric Dirichlet priors in the MTM estimate with $\alpha$ = 50/ $T$ and $\beta$= 0.01, which are common settings in the literature reference [25]. Because $\beta_b$ is weakly equal to $\beta$, also set $\beta_b$ =0.01. $\gamma$ is a prior of Bernoulli distribution, we set $\gamma$=0.5, it refers *t*o another similar study[16,21]. We used four-fifths of 1353 SMS spam to train, and the rest were mixed with 500 hams which were randomly drawn from 4872 hams, then, ran a 10-fold cross validation.

In the study of reference [3], SVM has been verified as outperforming other evaluated classifiers in the task of SMS spam filtering, such as Basic Naive Bayes (NB), K-Nearest Neighbors (KNN) etc. Therefore, SVM can be used as a baseline for comparison in our study. Compared MTM with SVM, the experimental results show in Fig. 2, both of them obtained high TABLE II shows the top-20 terms when $T$=8.

TABLE II. THE TOP-20 TERMS IN EACH TOPIC

| Topics | Annotation | Top-20 terms |
|---|---|---|
| *Topic1* | *Commercial advertising* | *Free, call, Tel, prize, miss, customer, discount, gift, shop, weekend, holiday, new, recent, you, award, cost, service, no, replay, msg.* |
| *Topic2* | *The winning cheat* | *Win, reward, call, Tel, valid, free, info, congratulate, cash, contact, receive, claim, expire, account, u, you, money, code, msg, to.* |
| *Topic3* | *Telecom operators value-added advertising* | *Call, update, ringtone, music, service, free, please, send, msg, message, wap, URL, link, subscription, mobile, replay, text, you, user, discount.* |
| *Topic4* | *Financial fraud* | *Please, claim, debt, expire, Tel, data, credit, reclaim, bank, charge, loan, bid, msg, message, contact, attempt, info, pay, entitle, expire.* |
| *Topic5* | *Premium fraud* | *Tel, call, msg, money, free, replay, cash, accident, claim, compensate, money, replay, record, entitle, indicate, owe, u, you, get, have.* |
| *Topic6* | *Software recommended advertising* | *Game, new, wap, play, win, now, club, URL, update, free, download, music, last, cost, to, ur, number, will, register, recepipt.* |
| *Topic7* | *Erotic services* | *Chat, u, you, age, try, sex, female, luv, replay, wait, invite, her, friend, Tel, secret, girl, alone, service, naked, call.* |
| *Topic8* | *Parcel fraud* | *Tel, call, you, u, urgent, a, parcel, await, please, delivery, service, customer, annoce, large, to, internet, free, for, now, date.* |

## IV. CONCLUSIONS

In this paper, a Message Topic Model (MTM) which is based on the probability theory of latent semantic analysis is proposed. The MTM can eliminate the sparse problem in SMS message classification and pay attention to kinds of symbols which usually appears in SMS spam. It is more suitable for the task of SMS spam filtering. When compared with the existing SMS spam filtering technologies and standard topic model—Latent Dirichlet Allocation (LDA), the MTM is more suitable for SMS filtering.

## ACKNOWLEDGMENT *(Heading 5)*

The authors would like to thank Almeida, TA. and his team for providing the Corpus of the SMS Spam Collection v.1, and Delany SJ, et al. for collecting the DIT SMS Spam Dataset. In addition, this work was supported by the Graduate Student Scientific Research and Innova-tion Project of Jiangsu Province, China (Grant No.: KYLX15_0494); the Fundamental Research Funds for the Central Universities of China; University Science Research Project of Jiangsu Province (15KJB520004), Science and Technology Projects of Huaian (HAS2015033, HAG2015060, HAG2014028).

## REFERENCES

[1] [Wu et al., 2008] N. Wu, M. Wu, and S. Chen, "Real-time monitoring and filtering sysem for mobile SMS", IEEE Conference on Industrial Electronics & Applications, 2008, pp. 1319 - 1324.

[2] [Almeida et al., 2011] T. Almeida, A. Hidalgo, J. M. G., and A.Yamakami, "Contributions to the study of SMS spam filtering: new collection and results", Paper presented at the Proceedings of the 11th ACM symposium on Document engineering, 2011.

[3] [Delan et al. et al., 2012] S. J., Delany, M. Buckley, and D.Greene, "SMS spam filtering: Methods and data", Expert Systems with Applications, vol. 39(10), pp. 9899-9908.

[4] [CNSR, 2014] Chinese National Spam Report Center, 2014, http://www.12321.cn/pdf/2014sbnsjdxdc.pdf

[5] [Jiang et al., 2013] N. Jiang, Y. Jin, A. Skudlark, and Z.-L.Zhang, "Understanding SMS spam in a large cellular network: characteristics, strategies and defences", Research in Attacks, Intrusions, and Defences, pp. 328-347: Springer.

[6] [Sohn et al., 2012] D.-N. Sohn, J.-T. Lee, K.-S. Han, and H.-C. Rim, "Content-based mobile spam classification using stylistically motivated features", Pattern Recognition Letters, vol. 33(3), 2012, pp. 364-369.

[7] [Wadhawan and Negi, 2014] A. Wadhawan, and N. Negi, "A Novel Approach For Generating Rules For SMS Spam Filtering Using Rough Sets", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME, vol. 3(7), 2014.

[8] [Qihoo 360, 2014] Qihoo 360, 2014, http://zt.360.cn/2015/reportlist.html?list=1

[9] [Liu and Wang, 2010] W. Liu, and T. Wang, Index-based Online Text Classification for SMS Spam Filtering. Journal of Computers, vol. 5(6), 2010.

[10] [Gómez Hidalgo et al., 2006] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C.García, "Content based SMS spam filtering", Paper presented at the Proceedings of the 2006 ACM symposium on Document engineering, 2006.

[11] [Deng et al., 2013] J. Deng, H. Xia, Y. Fu, J. Zhou, and Q. Xia, "Intelligent spam filtering for massive short message stream", COMPEL - The international journal for computation and mathematics in electrical and electronic engineering, vol. 32(2), 2013, app. 586-596.

[12] [Rosen-Zvi et al., 2004] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," Paper presented at the Proceedings of the 20th conference on Uncertainty in artificial intelligence, 2004.

[13] [Chan et al., 15] P. P. K. Chan, C.Yang, D. S.Yeung, and W. W. Y. Ng, "Spam filtering for short messages in adversarial environment," Neurocomputing, 2015, pp. 167-176.

[14] [Ponte et al., 1998] J. M. Ponte, and W. B.Croft, "A language modeling approach to information retrieval," Paper presented at the Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

[15] [Blei, 2012] D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55(4), 2012, pp. 77-84.

[16] B. Schölkopf, J. Platt, and T. Hofmann, "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model," Advances in Neural Information Processing System, Proceedings of the Twentieth Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December, 2006, pp.241-248..

[17] [Hofmann, 1999] T. Hofmann, " Probabilistic latent semantic indexing," Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999.

[18] [Blei et al., 2003] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, 2003,pp. 022.

[19] [Hong and Davison, 2010] L. Hong, and B. D. Davison, "Empirical study of topic modeling in Twitter," Proceedings of the Sigkdd Workshop on Social Media Analytics, 2010, pp. 80-88.

[20] Ho TP, Kang H-S, Kim S-R, "Graph-based KNN Algorithm for Spam SMS Detection," J UCS. 2013;19(16):2404-19.

[21] [Zhao et al., 2011] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," Paper presented at the In ECIR, 2011.

[22] [Chemudugunta, 2007] C. Chemudugunta, P. Smyth, M. Steyvers, "Modelling General and Specific Aspects of Documents with a Probabilistic Topic Model," Vol. 19, 2007, T Press.

[23] [Li et al., 2010] P. i, J. Jang, and Y. Wang, "Generating templates of entity summaries with an entityaspect model and pattern mining," Proceedings of Acl, 2010, pp.640-649.

[24] [Ahmed et al., 2015] I. Ahmed, R. Ali, D. Guan, Y.-K. Lee, S. Lee, and T. Chung, “Semi-supervised learning using frequent itemset and ensemble learning for SMS classification,” Expert Systems with Applications, vol. 42(3), 2015, pp. 1065-1073.

[25] [Heinrich, 2004] G. Heinrich, “Parameter estimation for text analysis,” Technical Report, 2004.

[26] C. Lai, and C. Moulin."Semantic indexing modelling of resources within a distributed system,"International Journal of Grid & Utility Computing , 2013, pp. 21-39.

[27] S. Pllana, S. Benkner, F. Xhafa, and L. Barolli, “A novel approach for hybrid performance modelling and prediction of large-scale computing systems,” International Journal of Grid & Utility Computing, vol. 1(4), pp.316-327(12).

[28] Abul. Bashar, "Graphical modelling approach for monitoring and management of telecommunication networks." International Journal of Space-Based and Situated Computing, 2015, pp. 65-75.

[29] Polk Tim, and Sean Turner, "Security challenges for the internet of things," Workshop on Interconnecting Smart Objects with the Internet, Prague,2011.