A New Fine-grain SMS Corpus and Its Corresponding Classifier Using Probabilistic Topic Model

Jialin Ma[1,2], Yongjun Zhang[1,2], Zhijian Wang[1], Bolun Chen[2]

[1] College of Computer and Information, Hohai University, Nanjing, 211100, China

[e-mail: majl@hyit.edu.cn]

[2] Huaiyin Institute of Technology, Huai'an, 223003, China

[e-mail: 13511543380@139.com]

*Corresponding author: Jialin Ma

Received March 26, 2017; revised August 4, 2017; revised September 7, 2017; accepted October 24, 2017; published February 28, 2018

Abstract

Nowadays, SMS spam has been overflowing in many countries. In fact, the standards of filtering SMS spam are different from country to country. However, the current technologies and researches about SMS spam filtering all focus on dividing SMS message into two classes: legitimate and illegitimate. It does not conform to the actual situation and need. Furthermore, they are facing several difficulties, such as: (1) High quality and large-scale SMS spam corpus is very scarce, fine categorized SMS spam corpus is even none at all. This seriously handicaps the researchers' studies. (2) The limited length of SMS messages lead to lack of enough features. These factors seriously degrade the performance of the traditional classifiers (such as SVM, K-NN, and Bayes). In this paper, we present a new fine categorized SMS spam corpus which is unique and the largest one as far as we know. In addition, we propose a classifier, which is based on the probability topic model. The classifier can alleviate feature sparse problem in the task of SMS spam filtering. Moreover, we compare the approach with three typical classifiers on the new SMS spam corpus. The experimental results show that the proposed approach is more effective for the task of SMS spam filtering.

The authors would like to thank all students and colleagues who contributed their SMS messages for collecting the HIT Spam Corpus. Specially, we would thank very much for our term members (Chunxia Jin, Hui Zhang, Changhui Yu) who expended a lot of effort and time to label the spam messages. In addition, this research was in part supported in part by the Chinese National Natural Science Foundation under grant No.61602202, and Natural Science Foundation of Jiangsu Province under contract BK20160428.

http://doi.org/10.3837/tiis.2018.02.004 **ISSN : 1976-7277**

Keywords: Spam SMS corpus, Topic Model, LDA, SMTM

1. Introduction

With the rapid growth of mobile phone users, the spammers have been transferring their target from email to SMS (Short Message Service). Due to its convenience and inexpensive price, a mass of criminals utilize SMS to defraud for economic or political benefits [2][11]. A report from the Chinese National Spam Report Center shows 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. In the USA, more than 69% of the surveyed mobile users claimed to have received text spams in 2012 [20]. In actual situation, the spam messages have been causing disaster in many countries, especially in India, Pakistan, and China where SMS lacks of effective supervisions [11].

Despite all of this, because the different national conditions and development stages, the standards of filtering SMS spam are different from country to country. For example, advertisement SMS spams without permission are deemed to harass users and violate personal rights in developed countries, but they are permitted in many developing countries, even as one of a important value-added service for the mobile operators [11]. In fact, we can easily observe that the contents of spam messages are varied. They consist of different categories, and usually include selling ads (such as house sale, financial product, mall sales, mobile operators, education and training), fraud (such as pretend landlord or friends, winning fraud, pretend banks, finance and investment fraud, credit card fraud, pretend court, pretend mobile operators, pretend airlines) and others (eroticism messages, fake invoices, etc.).

Nevertheless, the existing SMS spam filtering methods usually treat the SMS spam as a binary-class problem, which couldn't provide for fine-grain filtering. In addition, any machine learning approach (such as SVM, K-NN, and Bayes) is highly depending on the quality and quantity of the training data. Therefore, the scale and quality of SMS message corpus decide the accuracy of classifiers. However, high quality and large-scale SMS spam corpus is very scarce, fine categorized SMS spam corpus is even none at all at present. So far as we know, the largest corpus (DIT SMS Spam Dataset) only has 1353 spam messages [11].

Furthermore, the existing technologies about SMS spam filtering are facing the two serious challenges:

- The limited length of SMS messages lead to lack of enough features in traditional vector space model (VSM) which is usually apt to long text. These seriously degrade the performance of the traditional classifiers.

● SMS spam messages are usually injected with more kinds of abbreviations, symbols, websites, variant words, and distort or deform sentences than legitimate messages, which are usually used deliberately by the spammers to avoid anti-spam system filtering. The current studies often ignore these important characteristics.

In this paper, we present a new fine categorized SMS corpus which is unique and the largest one as far as we know. The corpus includes 13,078 Chinese real messages consist of 8,874 legitimate and 4,204 illegitimate SMS messages. In particularly, we had completed manual classification for the illegitimate message in the corpus, including 7 fine categories. In addition, we present a novel approach for the task of SMS spam filtering. In order to alleviate the feature sparse problem in the process of SMS spam classifying, we propose a *Short Message Topic Model* (*SMTM*) to enrich feature information. The *SMTM* is based on the probability topic model theory of latent semantic analysis. The proposed approach can alleviate the feature sparse problem in the process of SMS spam classifying. Moreover, non-linguistic features and pre-processing rules are used in our method. Finally, we compare the approach with three typical classifiers on the new SMS spam corpus. The experimental results show that the approach is more effective for the task of SMS spam filtering.

The paper is organized as follows: Section 2 reviews the related works about SMS spam corpuses, SMS spam filtering and topic model. Section 3 introduces the new SMS corpus in detail. Section 4 presents our *SMTM* and the proposed classifier for SMS spam filtering. Section 5 shows the experiments and discussion. Finally, we conclude and discuss further research in Section 6.

2. Related Works

This section reviews the situation of mainly existing SMS spam corpuses, and then briefly reviews the studies of spam filtering and some important research works about topic model in the past.

2.1 The Existing SMS Spam Corpuses

The main way to acquire SMS data is to via the contribution of mobile users or some website. At present, there are several commonly used SMS spam corpuses:

The first source is UCI machine learning repository (SMS Spam Collection v.1)[1], which has been used in relevant researches [11][16]. This corpus has been collected from several free or free for research sources at the internet. More useful information about the SMS Spam Collection v.1 can be found in three papers [2,3][15]. It can be downloaded at the following page of the UCI Repositoryfreely[1].

The second source is DIT SMS Spam Dataset[2]. This dataset is a corpus of 1,353 unique spam SMS text messages collected by scraping messages from two UK public consumer

complaints websites. The corpus covers the period from late 2003 to the middle of 2010. The information and research work can found in the paper [11].

[1] http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
[2] http://www.dit.ie/computing/research/resources/smsdata/.
[3] http://wing.comp.nus.edu.sg:8080/SMSCorpus/overview.jsp.

The last is NUS SMS Corpus[3]. It is collected by the group of researchers from the School of Computing of the National University of Singapore. They resurrected their earlier project from 2004 for SMS collection in October 2010 and focuses on collecting both English and Chinese SMS messages. This work is to expand their 2004 English SMS corpus and address the need for a public Chinese SMS corpus. They had created array of methodologies to collect SMS from contributors continuously and have collected 45718 English SMS messages and 31465 Chinese SMS messages as of January 2016 [8]. The related information about NUS SMS Corpus can be seen in their website[3].

Table 1. Properties and statistics of the main SMS corpuses

Source	Language	Total number	Category	Number	Proportion
SMS Spam Collection v.1	English	5,574	Spam	747	13.4%
			Ham	4,827	86.6%
DIT SMS Spam Dataset	English	1353	Spam	1353	100%
			Ham	0	0
NUS SMS Corpus	English	45,718	Untagged	—	—
	Chinese	31,465	Untagged	—	—

These three corpuses are the most frequently-used in the studies of SMS message. Obviously, the scales of them are not enough, especially the number of spam message. The most of spam data only have 1353 messages in the DIT SMS Spam Dataset. Although, the researchers of NUS can constantly enlarge their SMS corpus via their mechanism, the lack of their corpus is untagged dataset. Therefore, the untagged dataset is unavailable for the research of SMS spam filtering. **Table 1** give more detailed information of those SMS corpuses.

2.2 Related Works about SMS Spam Filtering and Topic Model

Most researchers still focus on the study of SMS spam filtering in recent years, although SMS is not as popular as social media, for example, Twitter, Facebook. The main reason is we still have been annoying by SMS spam. The researchers usually used many kinds of traditional and improved machine learning algorithms to detect spam messages [7][9][13] [19][22][25][28]. Almeida, Gomez Hidalgo, and Yama kami (2011) offered a new SMS spam corpus, and compared the performances between several established machine learning

algorithms (Naive Bayes, SVM, k-NN, etc.). Fifteen results achieved by combinations of classifiers and tokenizes were shown in their work in order to give a baseline. They also verified that SVM is the best classifier in SMS spam filtering task. In resent three years, some representative works appeared [5,6][16][20,21][27]. They paid more attentions on the new machine learning techniques.

Even though most researchers have been devoting to use or to improve kinds of standard classifiers for identifying SMS spam. However, the limited length of SMS messages leads to lack of enough features. These seriously degrade the performance of the traditional classifiers (such as SVM, K-NN, and Bayes). The traditional classifiers usually use vector space model (VSM) to represent text which is apt to long text. It is easily to cause similarity-drift when it is used for SMS messages. Moreover, it is well-known that VSM is lack of contextual and semantic information.

Latent Semantic Analysis (LSA) was first to bring semantic dimensionality between the text and words [26]. Then Probabilistic Latent Semantic Analysis (PLSA) is the further improvement of the LSA [17]. In PLSA, a document is regarded as a mixture of topics, while a topic is a probability distribution over words. Latent Dirichlet Allocation (LDA) was proposed in the first time [5], which added Dirichlet priors in the distributions. LDA is a more completely generative model and achieves great successes in text mining and other artificial intelligence domains [4]. Many LDA variations have been generated in recent years, especially using for social media mining [18]. In order to alleviate sparse problem, some researchers utilize external knowledge base or corpuses to augment the information [29]. In addition, Author-Topic Model (ATM) [24] and Twitter-LDA are famous models which take different strategies to aggregate the short texts into long texts [30]. Although many researchers are eager to use topic model for social media mining, only Modupe, Olugbara, and Ojo (2013) directly used Author-LDA it in SMS spam filtering in their work. This is most related to our work, but author information is usually missing in SMS dataset. Therefore, this approach has limitation and isn't used commonly.

3. The New SMS Corpus

Reliable dataset is significant in any scientific research. However, in the studies of mobile spam filtering, there are fewer corpuses available. Because of legal and privacy reasons, it is difficult to collect SMS messages from mobile phone users or mobile network operators.

3.1 Collection of the new SMS corpus

In order to fill the gap of the scarce dataset in study of SMS spam filtering and benefit all the researchers in this field, we have collected a new fine categorized SMS corpus[4] which is unique and the largest one as far as we know. The new SMS corpus is a set of tagged Chinese real messages. It contains 8,874 legitimate and 4,204 illegitimate, a total of 13,078 SMS messages. In particularly, we had completed manual classifying for the illegitimate

messages in the corpus, including 7 fine categories. All manual work was done by our team members from the Faculty of Computer and Software Engineering of Huaiyin Institute of Technology (HIT). They expended a lot of effort and time to label the spam messages.

This corpus has been collected from 2012 to 2015. All SMS spam messages were collected from the contributors who come from the Huaiyin Institute of Technology, including students, teachers, our team members and their friends. The contributors were made aware that their contributions were going to be made publicly available. The raw data is more than 6,000 spam messages. Finally, it remains 4,204 spam messages via duplicate removal. All messages were tagged by our team with 7 fine categories which was a very hard and time-consuming task.

[4] http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection

3.2 Statistics about the new SMS corpus

Table 2. Basic statistics of *HIT SMS Spam Corpus*

Category	Category ID	Fine-category	Number	Proportion(%)	Subtotal	Total
Ham	0	—	8874	67.9%	8874	13,078
Spam	1	Scam	205	1.6%	4,204	
	2	Commercial ads	2138	16.3%		
	3	Real estate ads	711	5.4%		
	4	Acting invoice	353	2.7%		
	5	Education and job	300	2.3%		
	6	Financial and investment	256	2.0%		
	7	Miscellaneous categories	241	1.8%		

Compared with other SMS corpuses, the new SMS corpus (named HIT SMS Spam Corpus) is not only the largest, but also has fine category of spam data. It can be downloaded freely4. More information about the corpus is showed in the **Table 2**.

Table 3. Token statistics for *HIT SMS Spam Corpus*

Category	Category ID	Tokens	Tokens Types	Avg length /Msg	Total Tokens	Total Tokens Types
Ham	0	175,586	15,733	20	175,586	15,733
Spam	1	5,920	1,897	29	138,004	20,948
	2	74,261	13,846	35		
	3	24,801	5,344	35		
	4	7,720	1,230	22		
	5	9,748	3,049	32		

	6	8,441	1,830	33		
	7	7,113	2,549	30		

We had used the tokenizer (jieba-0.38) to conduct word segmentation for the corpus. **Table 3** shows the statistics about the tokens extracted from the HIT SMS Spam Corpus. A **token** is the technical name for a sequence of characters—such as 'sale', 'ad', or ':)'. A **token type** is a unique item of token occurrences in the corpus.

4. Our Method

In this section, we present our classifier for SMS spam filtering in detail, which is based on the probability topic model.

4.1 Topic Model Construction

Conventional topic models, like PLSA and LDA, reveal the latent topics within the text corpus by implicitly capturing the document-level word co-occurrence patterns [5][14]. Nevertheless, directly applying these models on SMS corpus will suffer from the severe data sparse problem [30]. Although some topic models, such as Author-LDA or Twitter-LDA, were successfully used to mining micro log [29], except for short, SMS message are obviously different from micro blog in the two aspects: (1) SMS spam are usually filled up with many kinds of abbreviations, symbols, websites, variant words, and distort or deform sentences etc. Those non-linguistic tokens also represent the intentions of the senders. (2) SMS data usually isn't the same as micro blog which usually can easily acquire rich information, such as author, headlines, comments, and guest posts.

We proposed an aggregation strategy in training stage for SMS spam messages filtering which is different form Author-LDA. The specific process is as the following:

(1) After word segmenting, we not only retained the normal word but also non-linguistic tokens which are usually dropped in other text analysis.

(2) *TF-IDF* was used to extract features in order to acquire important terms.

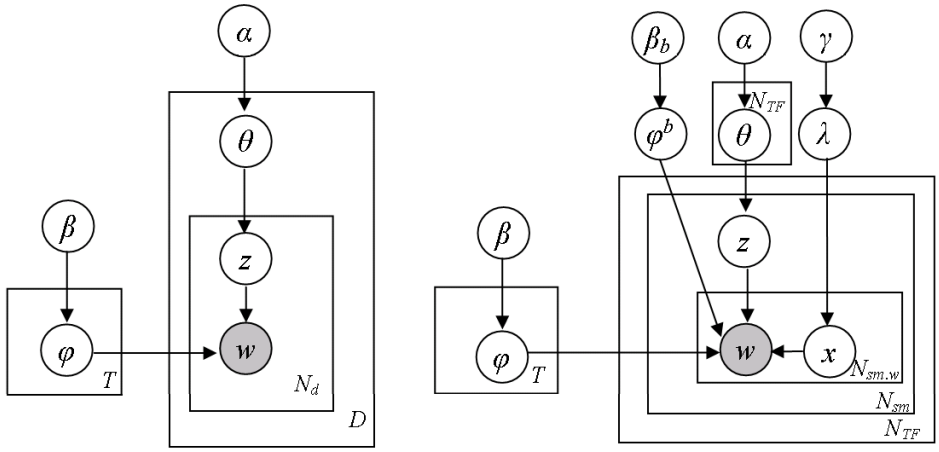
(3) Select first N terms in the result of (2).

(4) For each term in the N terms, a term profile is generated by aggregating SMS messages that contain this term.

This term profiles can form certain normal texts. Then we proposed a *Short Message Topic Model* (*SMTM*) based on the term profiles. Therefore, it can eliminate the sparsity in topic model training. The graph model of *SMTM* is showed in **Fig. 1**.

Fig. 1(a) shows the graphical model for the standard LDA. D is the total number of documents in the corpus, and N_d is number of words in d. The process includes two steps: first, assign a topic number from document-topic distribution θ; then, draw a word from topic-word distribution φ. 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. α and β are hyper-parameters of symmetric Dirichlet priors for θ and φ.



(a) (b)

Fig. 1. Plate notation of (a) Latent Dirchlet Allocation [5], (b) SM Topic Model(*SMTM*)

Fig. 2(b) shows a similar general structure: *Short Message Topic Model* (*SMTM*). z represent a topic, and *w* represent one of term in a term profile. $N_{sm,w}$ is the count of term in a term profile. N_{sm} is the count of SMS messages in a term profile. N_{TF} is total number of term profile. Let φ denote the word distribution for topics and φ_b denote the word distribution for background terms. Let θ denote the topic distribution of term profile. Let λ denote the Bernoulli distribution which controls the indicator x for the choice between background terms and topic terms. Φ, θ, and φ_b all obey multinomial distributions, each of term is drawn from symmetric Dirichlet (β), Dirichlet (β_b), and Dirichlet (α) respectively. λ is drawn from Beta (γ).

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

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

In Eq. (1), *w* is topic term when x=0, *w* is background term when x=1.

Because the real-world SMS messages usually contain with many noisy and redundant terms, we brought in a indicator x to select informative features in *SMTM*. Moreover, we also have a assumption that a single SMS message only has one topic due to a important fact: a SMS message is short. This inspiration is similar with the work: Zhao et al. (2011) which was used in twitter mining. The specific generation process can be described as follows:

1. Draw $\varphi^b \sim$ Dir(β_b) and $\lambda \sim$ Beta(γ)
2. For each topic t=1,…,T
   (a) draw a topic-word distribution $\varphi_t \sim$Dir(β)

3. For each term profile n=1,…,N_{TF}
   (a) Draw θ_n~Dir(α)
   (b) for each SMS message m=1,…N_{sm}
      (i) draw $z_{n,m}$ ~Multi(θ_n)
      (ii) for each term t=1,…,$N_{sm,w}$
         Draw x~B($N_{sm,w}$, λ)
         If x=0 draw w~Multi(φ)
         else if x=1 draw w~Multi(φ^b)

4.2 Parameter Inference

Gibbs sampling is an effectively and widely applicable Markov chain Monte Carlo algorithm for latent variable inference. Therefore, we adopt Gibbs sampling to infer parameters in the *SMTM*. $\boldsymbol{w}$ is used to denote all terms in the SMS spam corpus. $\boldsymbol{y}$ and $\boldsymbol{z}$ are hidden variables. If x=0, the standard LDA mechanism is used to generate the terms; otherwise, if x=1, the terms are generated via the background term distribution. In *SMTM*, when given the values for other variables, latent variable z can be sampled in Gibbs sampling as the following formulas:

$$P(z=t/\mathbf{z}_{\neg},x,\mathbf{w}) \propto \frac{n^{ms}_{(tp,t)}+\alpha}{n^{ms}_{(tp,*)}+T\alpha} \times \left(\frac{\Gamma(n^{*}_{t}+N_W\beta)}{\Gamma(n^{*}_{t}+n^{*}_{(tp,t,ms)}+N_W\beta}\right) \cdot \prod_{n=1}^{N_W} \frac{n^{w}_{t}+n^{w}_{(tp,t,ms)}+\beta}{n^{w}_{t}} \quad (2)$$

When given z=t, for the indicator y=0:

$$P(x=0/\mathbf{z},\mathbf{y}_{\neg},\mathbf{w}) \propto \frac{n^{w}_{t}+\beta}{n^{*}_{t}+N_W\beta} \cdot \frac{n_0+\lambda}{n_*+2\lambda} \quad (3)$$

If y=1:

$$P(x=1/\mathbf{z},\mathbf{y}_{\neg},\mathbf{w}) \propto \frac{n^{w}_{b}+\beta_b}{n^{*}_{b}+N_W\beta_b} \cdot \frac{n_1+\lambda}{n_*+2\lambda} \quad (4)$$

Table 4. The meaning of symbols of formula (2),(3) and (4)

symbols	meaning
¬	exclude the current *ms* of *t*.
N_W	the total number of terms in corpus.
ms	a SMS message.
tp	a term profile.
$n^{ms}_{(tp,t)}$	count of *ms* assigned to topic *t* for *tp*.
$n^{ms}_{(tp,*)}$	total count of *ms* in *tp*.
n^{*}_{t}	total number of terms are assigned to *t*.
n^{w}_{t}	count of term *w* is assigned to topic *t*.
$n^{*}_{(t,ms,tp)}$	total number of terms for *t* in the *ms* of the *tp*
$n^{w}_{(t,ms,tp)}$	count of *w* for *t* in the *ms* of the *tp*.
n^{w}_{b}	count of *w* is assigned to the background terms.

n_0	count of terms assigned as topic terms.
n_1	the number of terms assigned as background terms.

Algorithm 1, the training process of *SMTM*:

1. Initialize:
   (a) For each *term* in the corpus, assign x=0 or 1 randomly
   (b) For each *term* in the corpus
      If x=0,then
         assigns a topic number [1,….T] randomly for the *term*
         and make sure all topic terms in a *ms* share the same topic
2. For each *term* in the corpus
      If x=0, then
         execute sampling by formula (2) and update the topic number
         for this term and other topic terms in the *ms*
         sampling by formula (3) and update x
      Else sampling by formula (4) and update x
   Loop step 2 until Gibbs sampling is convergence
3. Acquire message-topic and topic-term distribution by count

4.3 Classifier Construction

In order to construct a classifier using for the task of fine-grained SMS spam filtering, we use the new *HIT SMS Spam Corpus* which has categorizing labels. It can help to label the topics which can be acquired by the *SMTM*. Intuitively, if a topic is associated with SMS spam messages in a category many times, then the topic likely belong to that category. Therefore, we utilized the proposed *SMTM* to learn topics in the Spam corpus. Then we can acquire a topic-term distribution matrix and each message can be assigned a specific topic. Let N_c^{ms} denotes the total number of SMS messages in category c. $n_{(t,c)}^{ms}$ is the number of messages that was assigned to topic t in c. The probability of topic t given category c as the following:

$$p(t/c)=\frac{n_{(t,c)}^{ms}}{N_c^{ms}} \tag{5}$$

Algorithm 2, the SMS spam filtering classifier based on *SMTM*

1. execute algorithm 1 for the training corpus resulting in the *SMTM*
2. execute algorithm 1 again for test SMS messages, then infer its topics
3. utilize formula (5) to obtain category-topic distribution matrix
4. For each *ms* in the test SMS corpus
      If the probability value of this message-topic >μ then

judge this is Spam message
$ms \in Max_c$ in the category-topic

5. Experiments and Discussion

In this section, we performed series of experiments on the new SMS Spam corpus (*HIT Spam Corpus*) with the classifier we proposed. The main content include pre-processing measures, analyzing thresholds, and comparing *SMTM* with the standard LDA. In additional, we also compared the performance of the proposed classifier with the traditional and typical three classifiers，the k-Nearest Neighbors (K-NN), Naive Bayes (NB) and the Support Vector Machine (SVM) for the task of SMS spam filtering.

5.1 Preprocessing

In order to promote the performance, we reserved non-linguistic tokens appearing in the SMS dataset. Other researchers, like T. A. Almeida et al. (2011), also considered that traditional preprocessing and feature selection techniques tend to hurt filters' accuracy. However, we define some replacement rules, as being shown in **Table 5** to improve the quality of the tokens.

Table 5. Preprocessing rules

Terms string in messages	Examples	Replacement rules
￥, $, £,€+number	￥88.8	{money}
Number+%, number+percent	20 percent discount	{discount}
http://+string, WWW.+string	http://www.388.cz.cc/	{url}
Ring, TEL, call+number	ring02073162414	{tel}
Time，date	21st May 2005	{time}
Long string Numbers	6222022010019460000	{account}
String+@+string)	xqz@163.com	{email}

5.2 Term Profile Generating

After word segmenting and pre-processing, we acquired 15,548 unique terms, and then adopted TF-IDF to compute features' value. The top TF-IDF values of terms were reserved to form term profiles. For each term in the top *N*, we aggregated SMS messages that contain this term into a training term profile. It is observed that data in the fine categories are imbalance. Therefore, we pick the terms following the proportion of SMS messages counts between fine categories and the whole Spam corpus. The **Table 6** shows the constitution of the term profiles.

In order to determine parameter *N* (the number of term profiles), we adopted an indicator-Coverage Rate to weigh how many top terms should be picked to generate term

profiles. The definition of Coverage Rate (*CR*) as follow:

$$CR=M/T \tag{6}$$

M: the number of messages which were included in any one of term profiles.

T: the total number of messages in the corpus.

Table 6. the proportion of term profiles in the categories

Fine-category ID	1	2	3	4	5	6	7
Proportion (%)	4.9%	50.9%	16.9%	8.4%	7.1%	6.1%	5.7%

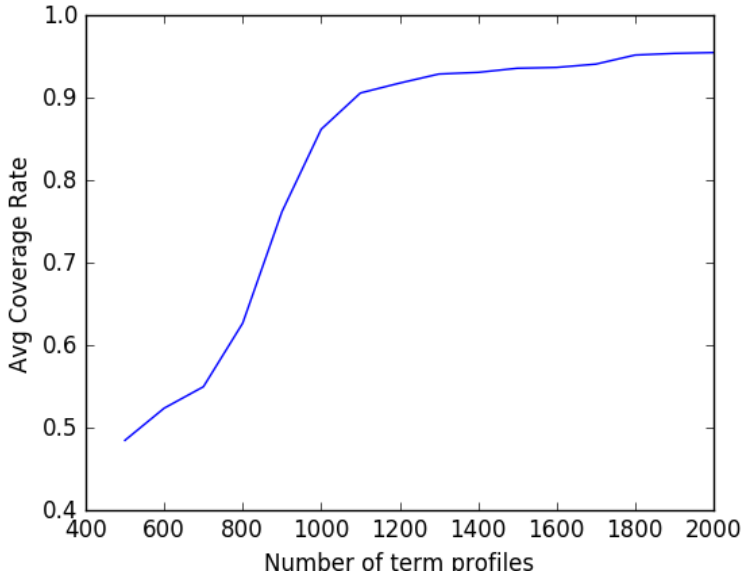


Fig. 2. Avg Coverage Rate Curve of term profiles in the categories

Fig. 2 shows the Average Coverage Rate (*ACR*), in which *N* is from 500 to 2000. It can be seen in the curve that the ACR is constantly increasing alone with the *N*. However, the ACR become trend to stable when *N* reach 1100. It is reflected that enlarge term profiles couldn't help to enhance ACR when it reach to about 90%. On the contrary, excessive emphasis on ACR could increase the messages repetitive rate of messages in the term profiles which will hurt the performance of *SMTM*. Therefore, *N* =1100 is the best value for this corpus.

5.3 Parameter Estimation and Topic Class

In order to discover topics from the HIT SMS Spam corpus (only for spams), we conducted three types of experiments: fist, directly conducted the standard LDA on the corpus; second, via the proposed aggregation strategy, we trained LDA on the term profiles (Term Profile + LDA); third, we utilize the *SMTM* completely. In the experiments, we ran

1000 iterations of Gibbs sampling using the Gibbs LDA++ toolkit3, α = 50/ T, and β= 0.01, which are common settings in the literature [14]. Because β_b is weakly equal to β, also set β_b =0.01. γ is a prior of Bernoulli distribution, thus we set γ=0.5, it refers to another similar study [7]. In order to find the optimal parameter T, We set the number of topics from 10 to 100 and adopted common criterion of perplexity to estimate the quality of topics [14].

$$perplexity = exp^{-(log(p(W)))/(N)} \tag{7}$$

$$P(W) = \sum z\ p(z|d) * p(w|z) \tag{8}$$

In Eq.(7) and (8), $P(W)$ is the probability of all terms appearing in the test corpus, w is one of term in W, z in one of the topics, N is the total number of terms in the test corpus. The three way experiments' results of perplexity are shows in **Fig. 3**:

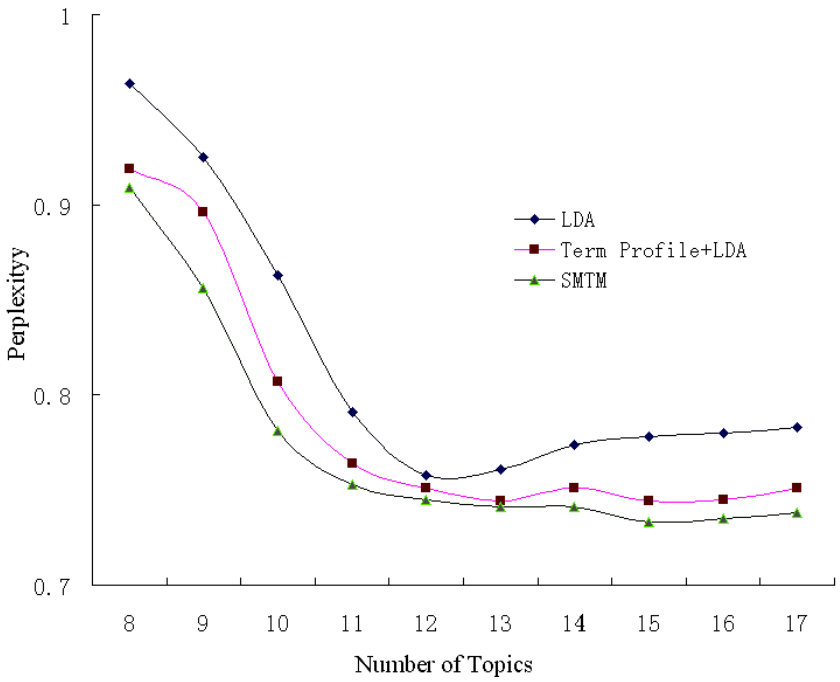


Fig. 3. The perplexity of topic from 8 to 17 for different models

We trained the three models in the same conditions. **Fig. 3** reflects the perplexities of the three models decline alone with the increase of topics. Although the inflection points of the three curves are different, the curves are relatively flat around 12. Therefore, we consider 12 is the appropriate value for T in the next experiments.

In addition, we utilize Eq. (5) to label the category of these topics. Apparently, the term profiles also inherit the labels from their SMS messages. We invited three language professionals to judge the results of topic labels.

Table 7. comparisons the results of topic classes

Method	Agreement (matched/topics)
Standard LDA	53.4%
Term Profile+LDA	68.7%
SMTM	80.5%

Table 7 shows the average judgment results of the labels for these topics discovered by each model. It is obvious that the term profiles is beneficial for eliminate sparse problem for SMS messages. *SMTM* can acquire more high quality topics than other two models. Furthermore, the term profiles are also helpful for standard LDA.

Table 8 presents the topics terms which were acquired by *SMTM*. It can be obviously observed from the results that a group term for a topic is around a semantic center. Specifically, topic0 is about real estate advertisement; topic1 is about commercial advertisement; topic2 is advertisement about mobile operators; topic3 is about recruitment information; topic4 is about education; topic5 is about erotic services; topic6 is about entertainment industry; topic7 is about automobile service; topic8 is about clothing shop; topic9 is about loan; topic10 is about house decoration; and topic11 is advertisement about financing.

Table 8. cases of topic terms acquired by *SMTM*

Topic0	Term	international	investment	book	M^2	opening	garden	villa
	probability	0.0085	0.0052	0.0051	0.0051	0.0049	0.0039	0.0039
Topic1	Term	promotion	discount	May Day	address	gift	All stores	appliance
	probability	0.01454	0.0079	0.0069	0.0057	0.0053	0.0049	0.0050
Topic2	Term	free	phone	com	http	www	acquire	download
	probability	0.0143	0.0078	0.0062	0.0051	0.0029	0.0028	0.0026
Topic3	Term	interview	recruit	free	immediate	salary	hour	telephone
	probability	0.0013	0.0008	0.0007	0.0006	0.0006	0.0005	0.0004
Topic4	Term	registration	telephone	child	teacher	study	educate	free
	probability	0.0036	0.0034	0.0031	0.0028	0.0027	0.0023	0.0019
Topic5	Term	address	enjoy	discount	drinks	belle	excite	service
	probability	0.0062	0.0051	0.0049	0.0039	0.00341	0.0022	0.0020
Topic6	Term	telephone	new	global	sport	entertain	blockbuster	movie
	probability	0.0023	0.0017	0.0014	0.0010	0.0010	0.0005	0.0004
Topic7	Term	consult	hot line	welcome	service	car	maintenance	4S store
	probability	0.0128	0.0114	0.0111	0.0111	0.0032	0.0010	0.0008
Topic8	Term	welcome	new style	shopping	customer	presence	fashion	our store
	probability	0.0067	0.0043	0.0041	0.0034	0.0024	0.00212	0.0016

Topic9	Term	company	shalom	contact	transact	loan	pledge	procedure
	probability	0.0162	0.0130	0.0116	0.0042	0.0042	0.0041	0.0038
Topic10	Term	telephone	decoration	experience	address	brand	owner	model
	probability	0.0013	0.0006	0.0005	0.0005	0.0003	0.0003	0.0002
Topic11	Term	product	income	finance	deadline	Financing	client	treasure
	probability	0.0027	0.0019	0.0019	0.0011	0.0009	0.0009	0.0009

5.4 Evaluation Measures

The related concepts and formulas are defined as the follows:

TP (true positive): the number of samples correctly classified to a specific class.

TN (true negative): the number of samples correctly rejected by a class.

FP (false positive): the number of examples incorrectly classified to a class.

FN (false negative): the number of examples incorrectly rejected by a class.

$$Precision(P) = \frac{TP}{TP + FP} \qquad Recall(R) = \frac{TP}{TP + FN} \qquad (9\text{-}10)$$

$$F_1 = \frac{2RP}{R + P} \qquad Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN} \qquad (11\text{-}12)$$

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \qquad (13\text{-}14)$$

Our basic task is to identify the unlabeled spam SMS dataset; this can be regard as binary-class problem. Therefore, we use ROC (Receiver Operating Characteristic) curve to measure the classifiers' performance. ROC curve consists of *FPR* as abscissa and *TPR* as ordinate.

The further task is multi-classification. We adopted a common evaluation indicator: F_1 measure, which is a synthetic indicator. It can reflect the mutual effects of both precision and recall. For the task of multi-classification, we adopt *Macro-F1* and *Micro-F1*. *Macro-F1* is the arithmetic mean for each category. *Micro-F1* is arithmetic average of each performance index for all instance documents in the dataset.

5.5 Experiment and Comparison

(a) For binary-classification task

Firstly, we conducted the experiments in order to investigate the effect of the preprocessing measures which were proposed in the section 5.1. The experiments used the *SMTM* and were ran a 10-fold cross validation. Moreover, we set μ>=0.0012, the message was judged as SMS spam. The *SMTM* trains on three quarters of spam messages, and the quarter of spam messages were mixed with the same amount of hams which random sample in the 8,874 hams for testing. We use ROC curve to measure the *SMTM* performances for the two cases:

- Normal preprocessing: this is normal preprocessing measures that remove the symbols and stop words.
- Proposed preprocessing: this is used our preprocessing method which introduced in section 5.1.

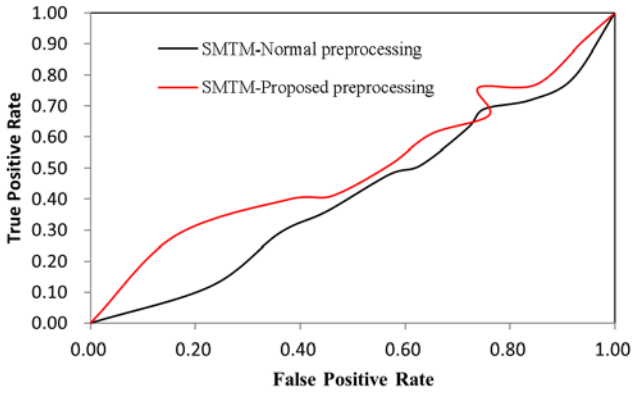


Fig. 4. ROC curves comparison on normal preprocessing and proposed preprocessing

We can see in **Fig. 4** that the ROC curve used the proposed preprocessing nearly keep above the unused case. It indicates that non-linguistic tokens appearing in the SMS dataset is valuable for SMS Spam identifying. The performance of classifiers can be promoted by using these replacement rules which are presented in **Table 5**. It is also proved that non-linguistic features in SMS spam imply rich semantic information and the traditional preprocessing measures or feature selection techniques are really hurting the classifiers' performance. Therefore, the next experiments of our work will all use this preprocessing.

Secondly, the next experiments is in order to compare the performance of *SMTM* with the traditional and typical three classifiers，the k-Nearest Neighbors (K-NN), Naive Bayes (NB) and the Support Vector Machine (SVM) for the task of binary-class. Because they need different parameters when make a decision for SMS Spam identifying, we normalize them so that we easily compare them. The experiments about KNN, NB, and SVM are with the help of python public toolkits: NLTK, and scikit-learn. Their ROC curves are showed in **Fig. 5**.

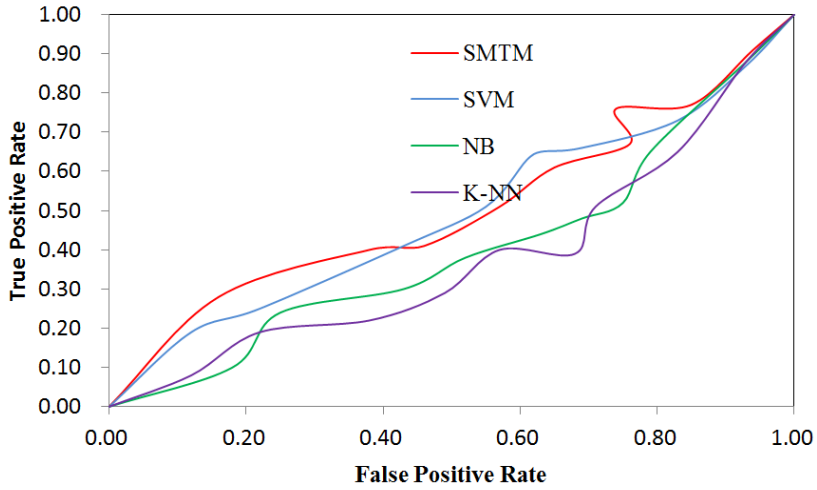
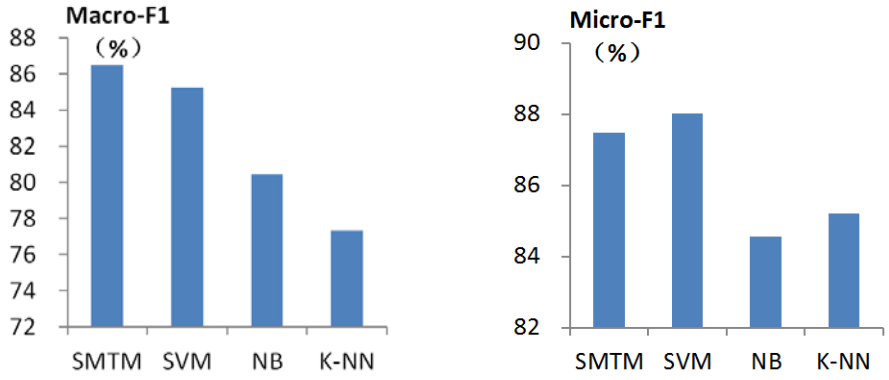


Fig. 5. ROC curves comparison for the four classifiers on proposed preprocessing

Fig. 5 reflects the *SMTM* is overall outperforms of the other three classifiers. But the SVM is close to our *SMTM*. The curves of NB and K-NN are nearly more or less, but are weaker than *SMTM* and SVM. This result indicates greatly improve the performance of classifiers is very difficult in the state of the art for the task of SMS spam filtering. The fundamental reason is the shorter of SMS message, thereby lack of adequate semantic information. The other reason is the spammers deliberately make variant content of SMS spam to avoid anti-spam system filtering. Although, the measure of preprocessing we proposed is effect, it is not perfect.

(b) For multi-classification task

We also implemented the four algorithms: *SMTM*, K-NN, NB and LibSVM for the task of multi-classification on our new SMS Spam corpus (*HIT Spam Corpus*). The experimental environment and data is the same as the above. Synthetic indicator *Macro-F1* and *Micro-F1* for the four classifiers are showed in **Fig. 6**.



(a) *(b)*

Fig.6. the comparison of Macro-F1 and Micro-F1 for the four classifiers

It is notable in **Fig. 6** that the *SMTM* and SVM achieved the best experimental results and outperformed the other two classifiers, almost reached 86% and 88% respectively. The *SMTM* is better than SVM in *Macro-F1*, but weaker in *Micro-F1*. In additional, K-NN is weaker than NB in *Macro-F1*, but better in *Micro-F1*. We speculate K_NN is unsatisfied with unbalanced data. On the whole, the performance of the *SMTM* is close to SVM, and NB is closed to K-NN. In order to further compare the *SMTM* and SVM, We calculated the evaluation indicators specifically for them. It is showed in **Table 9**.

The statistical data in **Table 9** indicates that the *Accuracy* and F_1 about the *SMTM* are better than the SVM in most categories. The fundamental reason is that the modeling process of the *SMTM* is different from the standard LDA. It utilizes the aggregated term profiles to eliminate the feature spare problem. By contrast, SVM couldn't deals with the sparse data well. However, the two classifiers are all weak in some categories, such as '5', '7'. The reason is those categories are lack of features, and the HIT SMS Spam corpus is imbalance in the categories. Despite all these, we found that the performance of *SMTM* would be decline when dealing with the data that has the mixed topics. This phenomenon can be also seen in **Table 9**, but the SVM is better or near to *SMTM* in the small sample categories '5', '6' '7'.

Table 9. comparison the *SMTM* with SVM

Category	Fine-category	SVM		*SMTM*	
—	—	*Accuracy*（%）	F_1（%）	*Accuracy*（%）	F_1（%）
Ham	0	90.1	88.6	91.6	90.2
spam	1	91.4	87.1	93.1	88.3
	2	90.7	84.3	89.4	86.7
	3	90.4	88.6	94.2	90.7
	4	91.8	88.9	91.9	92.8
	5	88.3	84.5	87.9	84.9
	6	86.9	81.6	87.4	80.8
	7	85.1	78.4	84.5	77.4

6. Conclusions and Future Work

In order to fill the gap of the scarce dataset in study of SMS spam filtering and benefit all the researchers in this field, we contribute a new fine categorized SMS corpus which is unique and the largest one as far as we know. In addition, we proposed a classifier for fine grained SMS spam filtering, which is based on the probability topic model. This classifier can alleviate the feature sparse problem in task of SMS spam filtering. The characteristic of the approach is not only can identify spam messages but also can provide fine-grained spam

filtering. Despite all that, the *HIT SMS Spam Corpus* we collected is still not very enough and has the problem of class-imbalance. Besides, the challenge of feature sparse for per SMS message in task of SMS spam filtering is still severe. The *SMTM* we proposed is not acquired total and breakthrough progress when compare with the traditional classifiers. But the idea of fine-grained SMS spam filtering is novel and practical in this specific field of research.

Future work should be considered to improve the efficiency of the *SMTM*. It is well known the spam filtering task often needs to complete in real time in practical applications. The high efficiency of online dynamic machine learning algorithms is desired in the future. Moreover, we plan to continue to collect the SMS spam messages and extend the *HIT SMS Spam Corpus* in the future.

References

[1] Ahmed, I., Ali, R., Guan, D., Lee, Y.-K., Lee, S., & Chung, T., "Semi-supervised learning sing frequent itemset and ensemble learning for SMS classification," *Expert Systems with Applications, 42*(3), 1065-1073, 2015. Article (CrossRef Link)

[2] Almeida, T., Hidalgo, J. M. G., & Silva, T. P., "Towards sms spam filtering: Results under a new dataset," *International Journal of Information Security Science, 2*(1), 1-18, 2013. Article (CrossRef Link)

[3] Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A., " Contributions to the study of SMS spam filtering: new collection and results," in *Proc. of Paper presented at the Proceedings of the 11th ACM symposium on Document engineering*, 2011. Article (CrossRef Link)

[4] Blei, D. M., "Probabilistic topic models," *Communications of the ACM, 55*(4), 77-84, 2012. Article (CrossRef Link)

[5] Blei, D. M., Ng, A. Y., & Jordan, M. I., Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022, 2003. Article (CrossRef Link)

[6] Chan, P. P. K., Yang, C., Yeung, D. S., and Ng, W. W. Y., "Spam filtering for short messages in adversarial environment," *Neurocomputing, 155*, 167-176, 2015. Article (CrossRef Link)

[7] Chemudugunta, C., Smyth, P., Steyvers, M., "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model," *MIT Press*, Vol. 19, 2007. Article (CrossRef Link)

[8] Chen, T., and Kan, M.-Y, "Creating a live, public short message service corpus: the NUS SMS corpus," *Language Resources and Evaluation*, vol. 47, no. 2, 299-335, 2013. Article (CrossRef Link)

[9] Cormack, G. V., Gómez Hidalgo, J. M., and Sánz, E. P., "Spam filtering for short messages," in *Proc. of Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, p. 313-320, 2007. Article (CrossRef Link)

[10] Cormack, G. V., Hidalgo, J. M. G., and Sánz, E. P., "Feature engineering for mobile (SMS) spam filtering," *Paper presented at the Proceedings of the 30th annual international ACM SIGIR*

conference on Research and development in information retrieval, 871-872, 2007. Article (CrossRef Link)

[11] Delany, S. J., Buckley, M., and Greene, D., "SMS spam filtering: methods and data," *Expert Systems with Applications*, vol. 39, no. 10, 9899-9908, 2012. Article (CrossRef Link)

[12] Deng, J., Xia, H., Fu, Y., Zhou, J., and Xia, Q., "Intelligent spam filtering for massive short message stream," *COMPEL - The international journal for computation and mathematics in electrical and electronic engineering*, vol. 32, no. 2, 586-596, 2013. Article (CrossRef Link)

[13] Endres, D. M., & Schindelin, J. E., "A new metric for probability distributions," *IEEE Transactions on Information theory*, vol. 49, no. 7, 2003. Article (CrossRef Link)

[14] Gómez Hidalgo, J. M., Bringas, G. C., Sánz, E. P., and García, F. C., "Content based SMS spam filtering," in *Proc. of Paper presented at the Proceedings of the 2006 ACM symposium on Document engineering*, p. 107-114, 2006. Article (CrossRef Link)

[15] Heinrich G., "Parameter estimation for text analysis," *Technical Report*, 2004. Article (CrossRef Link)

[16] Hidalgo, J. M. G., Almeida, T., and Yamakami, A., "On the validity of a new SMS spam Collection," in *Proc. of Paper presented at the Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, 2012. Article (CrossRef Link)

[17] Ho, T. P., Kang, H.-S., and Kim, S.-R., "Graph-based KNN Algorithm for Spam SMS Detection," *J. UCS*, vol. 19, no. 16, 2404-2419, 2013. Article (CrossRef Link)

[18] Hofmann T., "Probabilistic latent semantic indexing," in *Proc. of Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999. Article (CrossRef Link)

[19] Hong, L., and Davison, B. D, "Empirical study of topic modeling in Twitter," in *Proc. of Proceedings of the Sigkdd Workshop on Social Media Analytics*, 80-88, 2010. Article (CrossRef Link)

[20] Hu, X., & Yan, F., "Sampling of mass SMS filtering algorithm based on frequent time-domain area," in *Proc. of Kwledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on*, 2010. Article (CrossRef Link)

[21] Jiang, N., Jin, Y., Skudlark, A., and Zhang, Z.-L, "Understanding sms spam in a large cellular network: characteristics, strategies and defenses," *Research in Attacks, Intrusions, and Defenses*, pp. 328-347, Springer, 2013. Article (CrossRef Link)

[21] Kang, S.-S, "A Normalization Method of Distorted Korean SMS Sentences for Spam Message Filtering," *KIPS Transactions on Software and Data Engineering*, vol. 3, no. 7, 271-276, 2014. Article (CrossRef Link)

[22] Liu, W., and Wang, T. x., "Index-based Online Text Classification for SMS Spam Filtering," *Journal of Computers*, vol. 5, no. 6, 2010. Article (CrossRef Link)

[23] Modupe, A., Olugbara, O. O., & Ojo, S. O., "Investigating topic models for mobile short messaging service communication filtering," *Paper presented at the Proceedings of the World Congress on Engineering*, 2013. Article (CrossRef Link)

[24] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., "The author-topic model for authors and documents," in *Proc. of Paper presented at the Proceedings of the 20th conference on Uncertainty in artificial intelligence*, p. 487-494, 2004. Article (CrossRef Link)

[25] Sohn, D.-N., Lee, J.-T., and Rim, H.-C, "The contribution of stylistic information to content-based mobile spam filtering," in *Proc. of Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 321-324, 2009. Article (CrossRef Link)

[26] Thomas K. Landauer, P. W. F., Darrell Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, p. 259–284, 1998. Article (CrossRef Link)

[27] Wadhawan, A., & Negi, N., "A Novel Approach For Generating Rules For SMS Spam Filtering Using Rough Sets," *International Journal of Scientific & Technology Research, 3*(7), p. 80-86, 2014. Article (CrossRef Link)

[28] Wu, N., Wu, M., and Chen, S, "Real-time monitoring and filtering system for mobile SMS," in *Proc. of IEEE Conference on Industrial Electronics & Applications*, p. 1319 – 1324, 2008. Article (CrossRef Link)

[29] Yan, X., Guo, J., Lan, Y., and Cheng, X., "A biterm topic model for short texts," in *Proc. of Paper presented at the Proceedings of the 22nd international conference on World Wide Web*, 2013. Article (CrossRef Link)

[30] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., and Li, X., "Comparing Twitter and Traditional Media Using Topic Models," *Paper presented at the In ECIR*, p. 338-349, 2011. Article (CrossRef Link)

Jialin Ma is a PhD student of College of Computer and Information, Hohai University, Nanjing, China. He also is a senior experimenter of Huaiyin Institute of Technology, Huaian, China. His research interests focus mainly on natural language processing (NLP), data mining, and machine learning.

Yongjun Zhang is a PhD student of College of Computer and Information, Hohai University, Nanjing, China. He also is a lecturer of Huaiyin Institute of Technology, Huaian, China. His research interests focus mainly on natural language processing (NLP), data mining, and machine learning.

Zhijian Wang is a professor and doctoral supervisor of College of Computer and Information, Hohai University, Nanjing, China. His research interests focus mainly on software reuse technology and software system integration technology.

Bolun Chen is the Professor in Huaiyin Institute of Technology. He received his Ph.D. degree from Nanjing University of Aeronautics and Astronautics in the year 2016. His research interests include link prediction, recommender systems, data mining, and so on. At present, he is a reviewer of the Journal of Information Science and World Wide Web.