Identifying Valuable Knowledge Topics in Innovation Communities Using Innovation-LDA

Hongting Tang[1], Yanlin Zhang[1], Xianyun Lin[1], Lanteng Wu[2]*

[1] School of Management, Guangdong University of Technology, China
[2] School of Management, Huazhong University of Science and Technology, China
ht_tang@gdut.edu.cn, forest_zhang@163.com, 632176188@qq.com, lanteng_wu@foxmail.com

Abstract

Researchers and practitioners have recognized that user-generated content in the innovation community plays an important role. However, it is challenging to automatically identify valuable knowledge from these unstructured texts. Thus, in this study, we propose an efficient model for extracting innovation-oriented topics and, simultaneously, for assigning discovered topics to each post in the online innovation community. Specifically, we introduce a variant of the latent Dirichlet allocation (LDA) topic model, called the Innovation-LDA model, which comprehensively considers users' interests (reflected by *pageviews* and *replies*) and the structure of threads (e.g., *header* or *body*) to generate the valuable topics. We access the quality of discovered information through statistical fit as well as substantive fit. Based on our experimental results, we can conclude that our proposed method exhibits better performance than that of the contrasted method and can locate more meaningful innovation topics; that is, our innovation-LDA model is capable of not only identifying more rigorous topics for each thread by utilizing the text structure but is also capable of learning more semantic and coherent themes from user interests. This investigation expands topic identification research by providing both a new theoretical perspective and useful guidance for enterprises in product innovation.

Keywords: Topic modeling, Text analysis, Innovation community, Knowledge discovery

1 Introduction

User participation in online communities created by enterprises has contributed to a large number of novel and valuable intellectual achievements [1]. Thus, an online community is a significant source of information for its stakeholders to acquire product information, user experiences, and problem resolutions. In this regard, online communities also contributes significantly to giving enterprises access to a wider source of creativity. Therefore, many corporations, such as Google, Microsoft, and Xiaomi, have created online communities to obtain inspiration for new ideas from current users or from potential customers to improve their product competitiveness. Research has demonstrated that knowledge creation and sharing within the online innovation community can facilitate the development of new products, can improve the quality of product, and can facilitate cooperation among users [2]. In light of this, text analysis of online content from the innovation community has proven to be one of the most complicated and important tasks.

It is important to understand that unstructured texts contain a variety of knowledge features, such as knowledge stock [3], emotional orientation [4], and content readability [5], before exploiting online material in the innovation community. In this work, we are especially interested in another content characteristic, the knowledge topic (knowledge type) [6], which may make it easier for enterprises and users to locate the information they want. In more specific terms, knowledge topics refer to branches of online knowledge pertaining to related products, for example, *Lock Screen*, *Cloud Service*, or *Music*.

However, online content appears as informal (or even messy) unstructured text in the innovation community. Discovering valuable knowledge topics from a large scale of unstructured text is not straightforward work for researchers within the social sciences since it is infeasible to manually tag the online content, especially for large-sized corpora. To address this issue, it is apparent that we should employ automatic text analysis to community content, and the most common methods are the supervised learning methods and the unsupervised learning methods [7].

Supervised learning methods need a predefined set of knowledge topics, and the methods can be easily implemented if the expected knowledge topic set is available [8-9]. It is difficult, however, to determine knowledge topics in advance in most cases. According to our study, the knowledge topics in the innovation community are generally (1) unpredictable, (2) differ between forums, and (3) evolve over time. Obviously, due to the difficulty of obtaining prior knowledge about the exact topics that are contained within each community, supervised methods cannot be used in the assignment of knowledge topics. Furthermore, although users are asked to tag threads utilizing given labels when posting on the community, the majority of these labels are meaningless (e.g., *Others*) or nonrepresentative (e.g., *MiCoin*, a rarely used label). Consequently, it is necessary to propose an efficient method that automatically discovers valuable knowledge topics without manual intervention.

To address this requirement, the unsupervised methods such as the latent Dirichlet allocation (LDA) topic model [10] are clearly a well-established solution. A topic

*Corresponding Author: Yanlin Zhang; E-mail: forest_zhang@163.com
DOI: 10.53106/160792642023112406012

model is a statistical method for identifying sets of topics that depict a collection of documents [11]. However, unsupervised methods are also associated with inherent problems. In particular, unsupervised learning methods work independently without any interaction, which might result in a meaningless outcome for a certain knowledge task [12]. To address this problem, we propose Innovation-LDA, an extension of the LDA model, by incorporating automatic intervention. This approach is based on two reasonable assumptions: (1) The higher the user attention (represented by *pageviews* and *reviews*) of the thread is, the higher the quality and value; And (2) the importance of the *title* and *body* differs in determining the topic distribution of the thread.

Table 1 illustrates a brief overview of our method, by examining three threads. As the original contents are in Chinese, we have translated them into English. By using all the processed texts, the user attention (*pageviews* and *reviews*), and the text attributes (i.e., *title* or *text body*) as input, we propose a solution that automatically generates a set of knowledge topics and simultaneously assigns the generated topics to each thread. By combining this additional information with our proposed Innovation-LDA model, we expect to produce, as the experimental results indicate, more coherent and valuable topics. The following is a summary of the main contributions of our study.

Table 1. Three threads in the online innovation community

Thread_1	Header: New version of desktop was awkward to use. Body: As in the header, the new version of the desktop was awkward to use; it is not convenient to drag the icons.
	Pageviews: 297, Replies: 6.
Thread_2	Header: Suggestion for a function to uninstall the customized input method. Body: It would be helpful if the customized input method could be uninstalled. However, it continues to run in the background even though I have enabled the third-party input method.
	Pageviews: 302, Replies: 5.
Thread_3	Header: Calendar Body: As many jobs work in shifts, I would appreciate it if the calendar could include a shift function.
	Pageviews: 150, Replies: 0.

1) A novel topic model was proposed for identifying knowledge categories from informal unstructured contents of the innovation community; the model can operate automatically and can eliminate manual intervention.

2) The proposed method incorporates user interests (determined by pageviews and replies) as well as text features (i.e., title or body) to determine topic assignments, thereby providing a more coherent and valuable knowledge results.

3) Experimental studies based on real community data demonstrate that the proposed model is more effective at discovering more meaningful knowledge topics than those produced by competing methods.

2 Literature Review

Information acquisition and knowledge exchange have been greatly facilitated by the emergence of social websites. Nevertheless, user-generated content (UGC) on social websites tends to be informal or even inferior. There is no doubt that manual annotation could assist in filtering valuable information. It is, however, difficult to artificially extract valuable knowledge in our case, as the number of tasks and the amount of data are substantial. Therefore, for stakeholders, making the most of these online resources is difficult. Fortunately, researchers have turned to automatic text analysis to reduce human interference in knowledge-mining tasks such as sentiment analysis [13], text genre classification [14], and topic extraction [15]. Among these methods, the most representative are supervised and unsupervised learning methods.

2.1 Supervised and Unsupervised Learning Methods

The supervised learning method is a kind of widely used machine learning algorithm. A supervised model is trained to automatically categorize online texts into different types by using a training set after the researchers manually categorize portions of texts [8]. Because of their labor-saving [16] and easy-to-validate [17] features, supervised learning methods are widely used in feature recognition, ontology construction, and sentiment analysis. For example, Kumar et al. (2018) proposed a hierarchical supervised learning method for the detection of fraudulent posts on online business platforms [18]. Onan and Toçoğlu (2021) introduced a three-layer stacked Bi-LSTM architecture to identify sarcastic texts on social media data [19]. However, the accuracy of supervised learning methods relies heavily on strict rules of artificial annotation, and repetitive annotation efforts are necessary when contextual interests are changed. Hence, its application scope is strongly limited in our fast-changing environment.

Unsupervised learning is capable of automatically identifying potential text features from a dataset without predefined categories, as opposed to a method of supervised learning [20]. Consequently, it could be applied to a variety of text situations in a flexible manner. Moreover, the unsupervised learning approach has been able to reveal subtler and more significant details in the topic detection from online texts when compared to that of existing methods

[21]. Therefore, to achieve better results in content mining, many attempts have been made using unsupervised methods. In Onan (2019), a two-stage unsupervised framework based on improved word embeddings and cluster ensembles was proposed for extracting topics from scientific literature [22]. Deng et al. (2016) devised a model (i.e., TopWords) for automatic Chinese word segmentation in specific domain contents based on an unsupervised learning model, and it performed very well, especially in new word discovery and domain feature detection [23]. Because the innovation community is characterized by an array of expertise and changing interests, we have decided to adopt the unsupervised learning method to obtain more valuable results.

2.2 Latent Dirichlet Allocation Model

A number of unsupervised estimation algorithms have been proposed to detect the hidden features of online texts in the innovation community. Among all algorithms, the topic model is the most commonly used. LSI (Latent Semantic Indexing) was the first classic topic model that was introduced by Papadimitriou [24]. After that, Hofmann (2001) proposed a PLSI (Probabilistic latent semantic indexing) model based on LSI [25]. Accordingly, Blei et al. (2003) extended the PLSI model and introduced an LDA (Latent Dirichlet allocation) model [10]. Due to its simplicity, robustness, and interpretability, the LDA model has recently become the most commonly used topic model.

LDA models have been used extensively in various contexts, including news, corporate annual reports, academic journals, and more [11, 26]. Those content items are formal documents after syntax proofreading and spell checking. The content of online communities, however, appears as free-form texts and contains a variety of indecipherable elements, including abbreviations, emoticons, special characters, misspellings, and grammar mistakes. The noisy and ambiguous nature of online content makes knowledge discovery from the community quite challenging. Previous research has shown that without specific pretreatment, automatic topic models (e.g., LDA), which do well in official formal content, often have poor performance in online contexts [27]. Additionally, the short nature of online text has a tremendous negative impact on topic models' applicability [28].

To overcome these limitations, several extensions to the LDA algorithm have been proposed by scholars in response to the differences in the characteristics of online content. The Author-Topic model, for example, has been developed to overcome the problem of sparsity in online texts by aggregating all the contents generated by the same user as a single document to extract user interests [29]. The Twitter-LDA model, on the other hand, identifies topics for tweets by analyzing user-generated content of background words [30]. To eliminate the noise created by irrelevant texts, the Forum-LDA model differentiates users' serious interests and unserious interests in accordance with the relevance between post content and corresponding comments [31]. Beyond that, there are also several other LDA variants, such as Sent-LDA [11], and local-LDA [32].

Nonetheless, the focus of these studies has been on general communities rather than innovation communities, which have a greater level of professional product knowledge. Topic extraction in the innovation community is more likely to yield topics that are of interest to users in addition to guaranteeing accuracy. To some extent, these concerns expressed by users also reflect the readability and normativity of the posted content. Furthermore, research has shown that titles and bodies play different roles in the rendering of content [33-34]. For this reason, we propose an Innovation-LDA model based on two reasonable assumptions for identifying valuable knowledge topics within the innovation community. As the first point, the user's attention to each thread simultaneously reflects the significance of UGC and the quality of the text to some extent. Second, thread titles are generally shorter and tend to focus on the main idea, whereas thread bodies usually provide more details but contain more irrelevant information.

3 Model

In this section, we propose a variant of the LDA model that can be used to discover valuable knowledge topics in the innovation community. The purpose of our proposed model is described first, followed by the original LDA model and the parameter learning algorithm. The Innovation-LDA model is then formally presented with its variables and notations.

3.1 Preliminaries

According to Table 1, an online thread contains primarily two parts: thread title and thread body. Our observations indicate that the title is shorter and more focused on its main point, while the thread body often varies in length and includes some unnecessary content. Consequently, there is a difference in the importance of each part of the thread in determining the topic categories. However, it is important to note that the original LDA is a bag-of-words model in which the boundaries between title and body are ignored and it samples each word in the same document independently. To resolve this problem, we take the boundaries between the title and the thread body into consideration in our proposed model to make the words in the various parts of the thread no longer interchangeable. That is, different weights are assigned to texts in different parts of threads to obtain topic categories for each thread. Specifically, we follow the work of Li et al. (2021) [6] and set the weights of the title and thread body to 0.75 and 0.25, respectively. Remarkably, it is necessary to take both the title and thread body as our corpus in topic discovery in case one of them is lacking, irrelevant, or too short to process.

Furthermore, it is undeniable that many irrelevant threads that go off the target subject exist in the innovation community despite the founder having clarified its vision. Fortunately, as mentioned previously, users' attention provides an important clue as to the importance of knowledge and the quality of threads in the innovation community. The reasoning behind this is that users are more likely to view and respond to threads that relate to their own interests and to express them more clearly. As a way of maximizing the benefits of users' wisdom, it is advisable choice to assess the users' interests in various threads as part of the topic identification process. To achieve this, we assign varying weights to different threads in the topic generation based on threads' *pageviews* and *replies*. Specifically, the weight of each thread is taken as the normalized value of the sum of *page views* and *replies* after log transformation.

3.2 LDA Model

The original LDA model is briefly described before we introduce our new model. The LDA model is considered one of the most mature models in the field of unsupervised topic mining. The basic principle of this model is to represent each document through a discrete probability distribution of potential topics and each topic through a discrete probability distribution of words. It is possible to consider the generation of topics and the distributions of probability above as a random process of document creation determined by a probabilistic generative process. Figure 1 illustrates the graphical model of the LDA model. A description of the corresponding generative process can be found in Table 2. In Table 3. We summarize the notation we used.

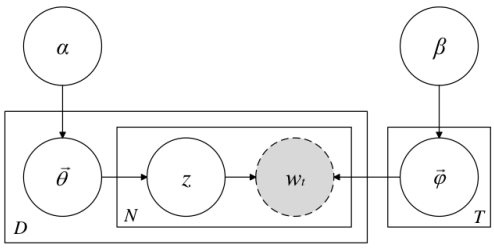


Figure 1. Graphical model of LDA

Table 2. Generative process of the LDA model

Input and Output		
Input:	$\vec{\alpha},\vec{\beta}$	/* Hyperparameters of Dirichlet distribution
	T	/* Number of topics
	D	/* Number of documents
Output:	$\vec{\theta}$	/* Topic distribution of the documents
	$\vec{\varphi}$	/* Word distribution of the topics

Main Procedure

for each topic $t \in \{1,2,...,T\}$ **do**

Draw a topic-word distribution: $\vec{\varphi}_t \sim Dirichlet\left(\vec{\beta}\right)$

end for

for each document $d \in \{1,2,...,D\}$ **do**

Draw a document-topic distribution: $\vec{\theta}_d \sim Dirichlet\left(\vec{\alpha}\right)$

for each word $w_{d,i} \in d, i \in \{1,2,...,N_d\}$ **do**

Draw a topic assignment: $z_{d,i} \sim Multinomial\left(\vec{\theta}_d\right)$

Draw a word: $w_{d,i} \sim Multinomial\left(\vec{\varphi}_{z_{d,i}}\right)$

end for

end for

According to Figure 1 and Table 2, $\vec{\theta}_d$ represents the T-dimensional topic distribution of document d, which is derived from a symmetric Dirichlet distribution with prior $\vec{\alpha}$. $\vec{\varphi}_t$ represents the V-dimensional word distribution of topic t, which is derived from a symmetric Dirichlet distribution with prior $\vec{\beta}$. *Dirichlet* $(\cdot)$ and *Multinomial* $(\cdot)$ indicate the Dirichlet distribution and multinomial distribution of parameter $(\cdot)$, respectively. N_d is the number of words in document d. $z_{d,i}$ is a topic extracted from a particular document distribution.

Based on Table 2, the key variables that need to be set are the topic number T and the Dirichlet prior parameters $\vec{\alpha}$ and $\vec{\beta}$. There are two target variables that need to be found: the document-topic distribution $\vec{\theta}$ and the topic-word distribution $\vec{\varphi}$. Unfortunately, the computation of related distributions is, in general, an intractable problem. For this purpose, two main types of estimation algorithms are available: variational algorithms and sampling-based algorithms. In the case of the LDA model that was just being proposed, the variational algorithms, especially the variational EM algorithm [10], are most commonly used. Later, Gibbs sampling [34], a sampling-based algorithm, became the most popular solution due to its simplicity and rapid convergence. Here, the target parameters are estimated by using Gibbs sampling.

Table 3. Notations

Notation	Detailed description
D	Number of documents (constant scalar)
N	Number of words in all documents (constant scalar)
T	Number of topics (constant scalar)
V	Number of words in vocabulary (constant scalar)
$\vec{\alpha}$	Hyperparameters of Dirichlet distribution (T-dimensional vector)
$\vec{\beta}$	Hyperparameters of Dirichlet distribution (V-dimensional vector)
$\vec{\theta}$	Topic distribution for documents ($D \times T$ matrix)
$\vec{\varphi}$	Word component for topics ($T \times V$ matrix)
$z_{d,i}$	Mixture indicator that chooses the topic for the ith word in document d
$w_{d,i}$	Term indicator for the ith word in document d

3.3 Parameter Estimation

Gibbs sampling aims at obtaining the posterior distribution over topic assignments of words, $p\left(z_i = t \mid \vec{w}, \vec{z}_{\neg i}\right)$, rather than object parameters $\vec{\theta}$ and $\vec{\varphi}$ as mentioned earlier. To accomplish this, the sampling probability for assigning the ith word in document d to a particular topic t is shown as below [35].

$$p\left(z_i = t \mid \vec{w}, \vec{z}_{\neg i}\right) \propto \frac{n_{t,\neg i}^{(v)} + \beta_v}{\sum_{v=1}^{V} n_{t,\neg i}^{(v)} + \beta_v} \bullet \frac{n_{d,\neg i}^{(t)} + \alpha_t}{\sum_{t=1}^{T} n_d^{(t)} + \alpha_t}, \quad \textbf{(1)}$$

where $\vec{w} = \left\{w_i = v, \vec{w}_{\neg i}\right\}$ is word vector in the document d, $\vec{z}_{\neg i}$ denotes the topic assignments for all words except word w_i, $n_{t,\neg i}^{(v)}$ represents the number of times that word w_i has been assigned to topic t except the current assignment of z_i, and $n_{d,\neg i}^{(t)}$ indicates that whether document d has been assigned to topic t when $t \neq z_i$, if so, $n_{d,\neg i}^{(t)} = 1$, otherwise 0.

In equation (1), the first ratio represents the probability of w_i in topic t, and the second ratio represents the probability of topic t in document d. Following Gibbs sampling, given a set of samples from posterior distribution $p(z_i = t \mid \vec{w}, \vec{z}_{\neg i})$, we can estimate the object parameters $\hat{\theta}$ and $\hat{\varphi}$ by using the following computational equations:

$$\hat{\theta}_{d,t} = \frac{n_{d,\neg i}^{(t)} + \alpha_t}{\sum_{t=1}^{T} n_d^{(t)} + \alpha_t}. \quad (2)$$

$$\hat{\varphi}_{v,t} = \frac{n_{t,\neg i}^{(v)} + \beta_v}{\sum_{v=1}^{V} n_{t,\neg i}^{(v)} + \beta_v}. \quad (3)$$

3.4 Innovation-LDA Model

To effectively apply the LDA model to informal texts in our case, we develop a model called the Innovation-LDA model that combines the target of topic discovery with the particular features of the innovation community's texts. Table 4 illustrates the complete solution process for Innovation-LDA. Specifically, we extend the original LDA model based on two reasonable assumptions. First, threads with more user attention are more innovative and have a higher level of text quality (i.e., accuracy, credibility, objectivity, and usefulness) [36]. Second, titles are shorter and focus more on the main idea than dose the body of the thread.

Therefore, to achieve a valuable and smooth result, we differentiate the priority of each text in the topic generation process. That is, different weight coefficients, $weight_d$ (i.e., product of weights determined by the structure text and weights derived by the interest of the user) are assigned to different texts in calculating the probability density function, $p(z_i = t \mid \vec{w}, \vec{z}_{\neg i})$. In particular, when word $w_{d,i}$ is assigned to topic t by using Gibbs sampling, the value of $n_d^{(t)}$ and $n_t^{(v)}$ in Equation (1) should be plus $weight_d$ rather than 1, as opposed to the approach of the original LDA model. We propose that weight coefficients could be used to guide the topic model to generate topics that have a higher innovation value, a higher text quality, and a higher generalization value. In addition, it is important to emphasize that we consider both titles and text bodies as our corpus when determining topics since the problem of data sparsity is prominent when using only the titles, whereas noise interference is serious when using only the text bodies.

Similar to the original LDA model, the parameters including the topic number T and the Dirichlet prior parameters $\vec{\alpha}$ and $\vec{\beta}$ should be set in advance in order to run Gibbs sampling on our corpus. When convergence has been achieved, target distributions $\vec{\theta}_d$ and $\vec{\varphi}_t$ can be calculated by using Equations (2) and (3) above. In the end, each thread's final topic assignment is jointly determined by the topic distribution of the title and body of the thread.

Table 4. Gibbs sampling algorithm for innovation-LDA

Input and Output		
Input	$\vec{\alpha}$, $\vec{\beta}$	/* Hyperparameters of Dirichlet distribution
	T	/* Number of topics
	D	/* Number of documents (including both header and text body)
	W	/* Weight sets of documents
Output	$\vec{\theta}$	/* Topic distribution for documents ($D \times T$ matrix)
	$\vec{\varphi}$	/* Word component for topics ($T \times V$ matrix)

Main Procedure

Initialization: $n_d^{(t)}, n_d, n_t^{(v)}, n_t = 0$

for each document $d \in \{1,2,...,D\}$ **do**

  for each word $w_{d,i} \in d, i \in \{1,2,...,N_d\}$ **do**

    sample topic index: $z_{d,i} = t \sim Multinomial(1/T)$

    increment document-topic count: $n_d^{(t)} + weight_d$

    increment document-topic sum: $n_d + weight_d$

    increment topic-word count: $n_t^{(v)} + weight_d$

    increment topic-word sum: $n_t + weight_d$

  end for

end for

while not **finished do**

  for each **document** $d \in \{1,2,...,D\}$ **do**

    for each word $w_{d,i} \in d, i \in \{1,2,...,N_d\}$ **do**

      /* for current assignment of t to a term v for word $w_{d,i}$:

      decrement counts and sums: $n_d^{(t)} - weight_d$; $n_d - weight_d$;

      $n_t^{(v)} - weight_d$; $n_t - weight_d$

      /* multinomial sampling according to Equation (3-1)

      sample topic index: $\tilde{t} \sim p(z_i \mid \vec{w}, \vec{z}_{\neg i})$

      /* use the new assignment of $z_{d,i}$ to the term v for $w_{d,i}$:

      increment counts and sums: $n_d^{(t)} + weight_d$; $n_d + weight_d$;

      $n_t^{(v)} + weight_d$; $n_t + weight_d$

    end for

  end for

  if converged or reached the sampling iterations L **then**

    read out document-topic distribution $\hat{\theta}$ according to Eq. 2

    read out topic-word distribution $\hat{\varphi}$ according to Eq. 3

  end if

end while

4 Experiments

The topic model is designed to uncover meaningful topics in data. However, not all arbitrarily generated topics reflect such a purposeful framework. To this end, our proposed model was evaluated by using the actual data and was compared with three competing unsupervised models in three aspects: predictive power, clustering quality, and topic quality.

4.1 Data and Experiment Setup

4.1.1 Data Preparation

Our proposed model is validated by using the threads posted in the MIUI Forum. The MIUI Forum is a typical example of an open innovation community in China. It is a professional information exchange platform developed by Xiaomi Tech. for users of the MIUI operation system or potential users. We prepared the data by crawling a sample of threads posted in the New Function Suggestion Section between Aug. 26th, 2017, and May. 11th, 2018. Following the removal of redundant and missing data, the number of valid threads was 23243. In each thread, only text information was retained, such as the category, title, text body, page views, comments, and active user information. In total, there were 76 classes and 15289 active users.

It is important to note that users are prompted to tag a category (i.e., class) when they post a thread in the MIUI community. The arbitrary nature of user-generated behavior, however, makes these categories often ambiguous or even inaccurate. Specifically, as illustrated in Figure 2, a significant percentage of threads (approximately 17.3%) are tagged with "Others" by users, yet most can be labeled into a particular category. Additionally, we observed that a significant number of threads are incorrectly categorized as other irrelevant categories. The use of incorrect tags not only negatively affects user search results but also greatly hampers the discovery of innovation inspiration by the enterprises. As a consequence, it is necessary to investigate an efficient method for identifying topics of UGC in the MIUI Forum.

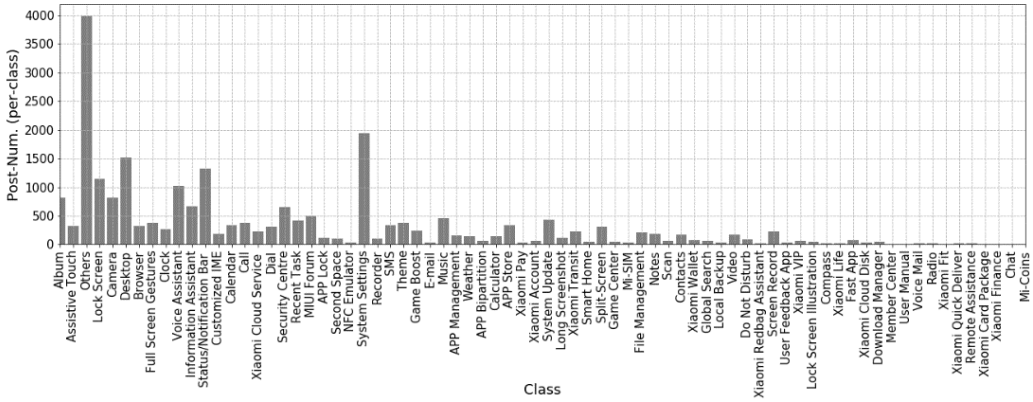


Figure 2. Distribution of threads' classes in MIUI community

In data preprocessing, we use Jieba, a Python library for Chinese word segmentation, to segment Chinese texts into single words. Specifically, we use an HMM-based model with the Viterbi algorithm to detect the new words in user-generated content, which could be emerging product features. On this basis, we construct a custom lexicon and use it to increase the efficiencies of word segmentation. After Chinese tokenizing, we remove all stop words and the top 10 most frequent words and finally obtain 26921 keywords as our lexicon.

4.1.2 Benchmark Methods

To demonstrate the performance of our proposed model, a comparison was made between the proposed model and the original LDA model [10], the local-LDA model [32], and the Sent-LDA model [11] in terms of its predictive power, clustering quality, and topic quality. Our model, along with the last two methods, is designed to enhance the original LDA model by taking advantage of the structure of the text. In contrast to the original LDA model, the local-LDA model emphasizes the boundary between sentences and determines the topic distribution within each sentence rather than the topic distribution within each document. In the Sent-LDA model, the boundary between sentences is also emphasized, and the difference is that it assumes 'one topic per sentence' in the generation of topics. It has been demonstrated that both the Local-LDA model and the Sent-LDA model perform significantly better than other methods when summarizing reviews or discovering aspects [11]. For our evaluation, therefore, we select the original LDA model, the Local-LDA model, and the Sent-LDA model as our benchmarks.

As a baseline, we use Gibbs sampling as a learning algorithm for all methods so that we can fairly compare the performance of our proposed model with that of the competing methods. For Gibbs sampling, the Dirichlet prior parameters α and β are set to 0.1 and 0.01, respectively. In addition, Gibbs sampling is iterated 2000 times to guarantee convergence.

4.2 Predictive Power

Predictive power refers to the ability of an algorithm to generate testable predictions. It differs from explanatory power in that it allows a prospective assessment of methodological validity other than examination of results since this phenomenon cannot be explained retrospectively by a given theory or established fact. Therefore, predictive power is a typical quality indicator for unsupervised algorithms such as classification methods or regression predictive models. In the topic model, predictive power is proposed to measure how well the model predicts potential semantic structures. Specifically, when assigning words to a particular 'topic', a good model should gather similar words in higher proportions for each topic. To assess the predictive

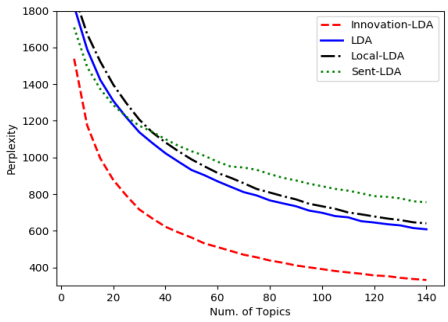
power of topic models, we use a metric called perplexity [35], which is a standard way of measuring the predicted performance of a probabilistic model.

The perplexity refers to the log-averaged inverse probability on unseen documents [11], and a lower perplexity indicates that the probability model performs better at predicting the words on new unseen documents. The perplexity value of a set of documents D containing the word w_d is defined as follows:

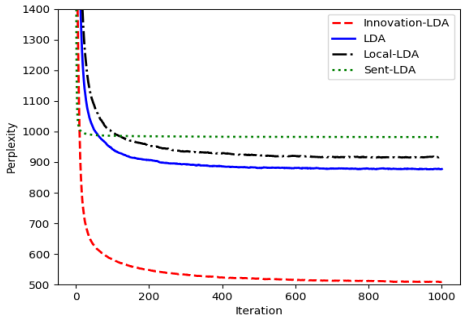
$$perplexity(D) = \exp\left\{-\frac{\sum_{d=1}^{D}\log p(w_d)}{\sum_{d=1}^{D} N_d}\right\}, \quad (4)$$

where N_d is the number of words in document d.

We evaluated the performance of competing models by varying the number of topics to ensure that a fair comparison could be made. Figure 3(a) illustrates four models of perplexity over various topic numbers, which were obtained at the 140th iteration. Figure 3(a) shows that the perplexity of all methods decreases monotonically with increasing of the topic number and tends to converge when the topic number exceeds 60. Additionally, the Sent-LDA model has much worse perplexity, which may result from inaccurate segmentation of informal texts. While the Innovation-LDA model, our proposed model, has a lower perplexity score than that of the other competing methods, which implies that it has greater predictive power and can predict words in new untested documents well.



(a) Perplexity as a function of the number of topics



(b) Perplexity as a function of iterations

Figure 3. Perplexity of competing models

Figure 3(b) shows the perplexity against the iterations of the four models. According to the figure, most of the models tend to converge within 100 iterations. In this experiment, our model achieved a convergence rate second only to that of the Sent-LDA model, with the number of effective iterations below ten. Therefore, the topic model is more effective when considering the hierarchical structure of the corpus of text.

4.3 Clustering Quality

Clustering quality measures determine the level of "goodness" or "cognancy" of clustering results by quantifying the intercluster and intracluster similarity [37]. Many cluster validity indices have been developed for the purposes of assessing the cluster quality by using similarities, including external (requiring ground truths) and internal (with inherent data) methods [38]. Considering that the ground truth of our dataset is not available, we use an intrinsic method, referred to as the silhouette coefficient [39], to determine whether the topics generated by competing models are appropriate. The silhouette coefficient of a given document d is defined as follows:

$$s(d) = \frac{b(d) - a(d)}{\max\{a(d), b(d)\}}, \quad (5)$$

where $a(d)$ is the average distance between d and all other documents in the topic to which d belongs. $b(d)$ is the minimum average distance from d to all topics to which d does not belong. Therefore, $a(d)$ reflects the compactness, while $b(d)$ captures the separation of topic membership. In general, the silhouette coefficient ranges from -1 (for poor clustering quality) to 1 (for good clustering quality).

As shown in Table 5, we can compute an average silhouette coefficient value for all documents in the dataset to determine the quality of clustering in a model. In Table 5, the silhouette coefficient values are presented for all competing methods when the topic number is set to 60. To calculate the distance between various documents, the Euclidean metric is used to measure the straight-line distance between two objects in Euclidean space. According to Table 5, our proposed model obtains the highest average silhouette coefficient value, which indicates that our proposed model performs better than the other competing models. Following that are the Original-LDA model and the Sent-LDA model.

Table 5. Model comparison in terms of silhouette coefficient

	Original-LDA	Sent-LDA	Local-LDA	Innovation-LDA
Mean	-0.075	-0.115	-0.116	-0.044
Std. Dev.	(±0.026)	(±0.018)	(±0.016)	(±0.001)
p-value	0.077	0.037	0.023	-

4.4 Topic Quality

The evaluation metrics we used above are important to help us understand the computer-based performance of our model. However, it is more important to conduct a further evaluation to determine whether generated topics are sufficiently informative and meaningful for practical applications. To this end, we utilize artificial judgment to evaluate all topics generated by all the competing models. Prior to that, we present some of the topics that have been

generated by the Innovation-LDA model in the MIUI community, as shown in Figure 4.

Figure 4 shows that each topic is represented by a word cloud map with the font size indicating the likelihood of the words being in the target topic and each topic is defined by a maximum of 50 keywords. According to Figure 4, the words with a higher probability of occurrence in a specific topic are semantically similar. For example, in terms of "camera", high-frequency words include *photographing, optimization,* and *watermarking*, which are all relevant features of camera functionality. Most of the generated topics we obtain from Innovation-LDA echo user-labeled categories, such as "Full Screen Gestures" and "Game Boost". Moreover, according to the results, some highly similar topics have been merged; for example, "Lock Screen" and "Lock Screen Illustration" have been integrated into "Unlock". Additionally, certain original categories, such as "Voice Assistant", have been replaced by more logical and user-friendly topics, i.e., "Xiaoai Classmate". More importantly, we have discovered several new topics of knowledge that have not previously been explored by users, such as "Mobile Traffic" and "Advertising". We could, therefore, identify more useful topic structures beyond those user-labeled categories by employing the proposed Innovation-LDA model.

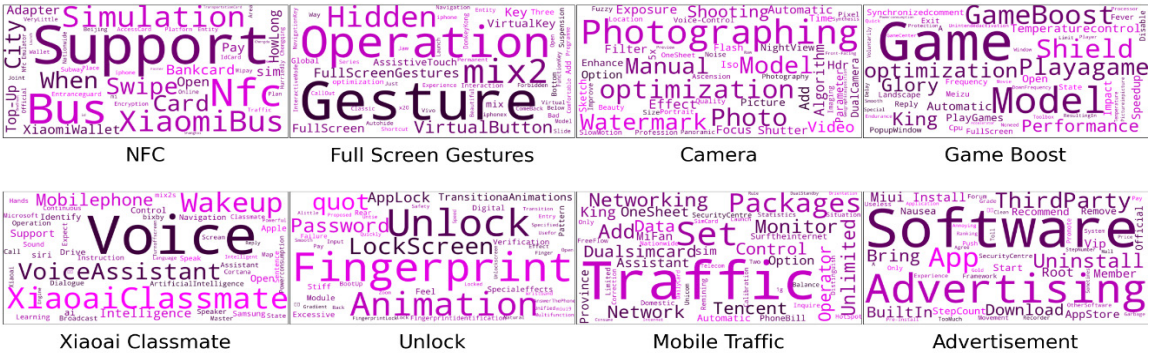


Figure 4. Several topics generated by innovation-LDA

5 Conclusion

Researchers and practitioners have long realized that user-generated content in the innovation community is enlightening for product improvement and innovation. To this end, researchers have turned to work on automatic text analysis in knowledge-mining tasks. Despite this, the majority of existing research has focused on general knowledge mining rather than innovation-oriented knowledge extraction. The purpose of this paper is to introduce a novel topic model called the Innovation-LDA model, which can be used to automatically identify innovation-oriented knowledge categories, i.e., topics, from unstructured content in the innovation community. In particular, we comprehensively consider the users' interests (characterized by *pageviews* and *replies*) as well as the structure of the thread (e.g., *title* or *body*) and then, we assign weight coefficients accordingly when modeling the generative process of the topic model. By incorporating these elements, our Innovation-LDA model can not only obtain more rigorous topics for each thread, but can also learn smoother topics that contain more semantic information. Extensive experiments conducted on real data demonstrate that the proposed model yields more coherent and valuable results and obtains better performance than that of the other competing approaches.

Several theoretical implications can be drawn from our research. First, our work is guided by the knowledge-based theory of the firm, which holds that organizations are integrators rather than creators of knowledge, and that knowledge resides within individuals. By extracting innovation-oriented topics from individuals, our work can expand the firm's ability to integrate their users' knowledge and can enrich the literature on the knowledge-based theory of the firm. Second, our study is the first to identify topics from user-generated content according to text structure and user interests, which provides a certain reference for the improvement in other relevant approaches.

Several managerial implications can be drawn from our research. In particular, our model infers more representative topics than manually labeled categories and the results generated by other competing algorithms. These topics provide managers with a more precise understanding of aspects of user-generated content, which could inspire further product innovation. In addition, we found that our topic assignment for each thread is more informative and effective, thereby providing more efficient tools for information retrieval and content management within the innovation community.

Despite our proposed model receiving considerable advancement in topic assignment, this work has much room for future research. First, this work uses metrics of clustering methods for the evaluation of the unsupervised topic model, and more sophisticated methods are anticipated to be developed to obtain more reliable results. Second, the topic number coefficients are determined by experiments that waste computing power. In future work, a method that could scientifically specify the number of topics needs to be further studied. Finally, our proposed model was verified for only a single community, The generalization of our proposed model should be investigated in the future.

Acknowledgment

We gratefully acknowledge the funding support from the National Natural Science Foundation of China (Grant 72101060, 72272039) and Guangzhou Basic and Applied Basic Research Foundation (Grant 202201010333).

References

[1] M. Seraj, We create, we connect, we respect, therefore we are: intellectual, social, and cultural value in online communities, *Journal of Interactive Marketing*, Vol. 26, No. 4, pp. 209-222, November, 2012.

[2] M. Sheng, R. Hartono, An exploratory study of knowledge creation and sharing in online community: A social capital perspective, *Total Quality Management & Business Excellence*, Vol. 26, No. 1-2, pp. 93-107, February, 2015.

[3] C. Rupietta, U. Backes-Gellner, Combining knowledge stock and knowledge flow to generate superior incremental innovation performance — Evidence from Swiss manufacturing, *Journal of Business Research*, Vol. 94, pp. 209-222, January, 2019.

[4] H. H. Do, P. W. C. Prasad, A. Maag, A. Alsadoon, Deep learning for aspect-based sentiment analysis: a comparative review, *Expert Systems with Applications*, Vol. 118, pp. 272-299, March, 2019.

[5] X. Liu, G. A. Wang, W. Fan, Z. Zhang, Finding Useful Solutions in Online Knowledge Communities: A Theory-Driven Design and Multilevel Analysis, *Information Systems Research*, Vol. 31, No. 3, pp. 731-752, September, 2020.

[6] Z. Li, H. Tang, X. Xu, Q. Chen, Knowledge Topic-Structure Exploration for Online Innovative Knowledge Acquisition, *IEEE Transactions on Engineering Management*, Vol. 68, No. 6, pp. 1880-1894, December, 2021.

[7] J. Hartmann, J. Huppertz, C. Schamp, M. Heitmann, Comparing automated text classification methods, *International Journal of Research in Marketing*, Vol. 36, No. 1, pp. 20-38, March, 2019.

[8] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania, USA, 2006, pp. 161-168.

[9] V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, J. M. Gambetta, Supervised learning with quantum-enhanced feature spaces, *Nature*, Vol. 567, No. 7747, pp. 209-212, March, 2019.

[10] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, *the Journal of machine Learning research*, Vol. 3, pp. 993-1022, March, 2003.

[11] Y. Bao, A. Datta, Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science*, Vol. 60, No. 6, pp. 1371-1391, June, 2014.

[12] A. Suominen, H. Toivanen, Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification, *Journal of the Association for Information Science and Technology*, Vol. 67, No. 10, pp. 2464-2476, October, 2016.

[13] A. Onan, S. Korukoğlu, A feature selection model based on genetic rank aggregation for text sentiment classification, *Journal of Information Science*, Vol. 43, No. 1, pp. 25-38, February, 2017.

[14] A. Onan, Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach, *Computer Applications in Engineering Education*, Vol. 29, No. 3, pp. 572-589, May, 2021.

[15] A. Onan, Biomedical text categorization based on ensemble pruning and optimized topic modelling, *Computational and Mathematical Methods in Medicine*, Vol. 2018, pp. 1-22, July, 2018.

[16] K. Li, W. Ai, Z. Tang, F. Zhang, L. Jiang, K. Li, K. Hwang, Hadoop recognition of biomedical named entity using conditional random fields, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 26, No. 11, pp. 3040-3051, November, 2015.

[17] J. Grimmer, B. M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, *Political analysis*, Vol. 21, No. 3, pp. 267-297, Summer, 2013.

[18] N. Kumar, D. Venugopal, L. Qiu, S. Kumar, Detecting review manipulation on online platforms with hierarchical supervised learning, *Journal of Management Information Systems*, Vol. 35, No. 1, pp. 350-380, 2018.

[19] A. Onan, M. A. Toçoğlu, A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification, *IEEE Access*, Vol. 9, pp. 7701-7722, January, 2021.

[20] X. Mao, H. Yang, S. Huang, Y. Liu, R. Li, Extractive summarization using supervised and unsupervised learning, *Expert Systems with Applications*, Vol. 133, pp. 173-181, November, 2019.

[21] L. Guo, C. J. Vargo, Z. Pan, W. Ding, P. Ishwar, Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling, *Journalism & Mass Communication Quarterly*, Vol. 93, No. 2, pp. 332-359, June, 2016.

[22] A. Onan, Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering, *IEEE Access*, Vol. 7, pp. 145614-145633, October, 2019.

[23] K. Deng, P. K. Bol, K. J. Li, J. S. Liu, On the unsupervised analysis of domain-specific Chinese texts, *Proceedings of the National Academy of Sciences*, Vol. 113, No. 22, pp. 6154-6159, May, 2016.

[24] C. H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent semantic indexing: A probabilistic analysis, *Journal of Computer and System Sciences*, Vol. 61, No. 2, pp. 217-235, October, 2000.

[25] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, *Machine learning*, Vol. 42, No. 1-2, pp. 177-196, January, 2001.

[26] W. S. Lee, S. Y. Sohn, Identifying emerging trends of financial business method patents, *Sustainability*, Vol. 9,

No. 9, Article No. 1670, September, 2017.
[27] J. Tang, M. Zhang, Q. Mei, One theme in all views: modeling consensus topics in multiple contexts, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, Chicago, Illinois, USA, 2013, pp. 5-13.
[28] T. Kruft, C. Tilsner, A. Schindler, A. Kock, Persuasion in Corporate Idea Contests: The Moderating Role of Content Scarcity on Decision-Making, *Journal of Product Innovation Management*, Vol. 36, No. 5, pp. 560-585, September, 2019.
[29] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, Washington, USA, 2004, pp. 306-315.
[30] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, X. Li, Comparing twitter and traditional media using topic models, *European conference on information retrieval*, Dublin, Ireland, 2011, pp. 338-349.
[31] C. Chen, J. Ren, Forum latent Dirichlet allocation for user interest discovery, *Knowledge-based systems*, Vol. 126, pp. 1-7, June, 2017.
[32] S. Brody, N. Elhadad, An unsupervised aspect-sentiment model for online reviews, *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, Los Angeles, California, 2010, pp. 804-812.
[33] Y. Zhao, S. Yang, Y. Li, Y. Chen, J. Yao, A. Qazi, Does the review deserve more helpfulness when its title resembles the content? Locating helpful reviews by text mining, Information Processing & Management, Vol. 57, No. 2, Article No. 102179, March, 2020.
[34] T. L. Griffiths, M. Steyvers, Finding scientific topics, *Proceedings of the National academy of Sciences*, Vol. 101, No. suppl 1, pp. 5228-5235, April, 2004.
[35] S. Burkhardt, S. Kramer, Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model, *Journal of Machine Learning Research*, Vol. 20, No. 131, pp. 1-27, 2019.
[36] H. Tang, X. Xu, Z. Li, R. Qin, Identifying contributory domain experts in online innovation communities, *Electronic Commerce Research*, Vol. 23, No. 4, pp. 2759-2787, December, 2023.
[37] A. Nazari, A. Dehghan, S. Nejatian, V. Rezaie, H. Parvin, A comprehensive study of clustering ensemble weighting based on cluster quality and diversity, *Pattern Analysis and Applications*, Vol. 22, No. 1, pp. 133-145, February, 2019.
[38] J. C. Lamirel, N. Dugué, P. Cuxac, New efficient clustering quality indexes, *2016 International joint conference on neural networks (IJCNN)*, Vancouver, BC, Canada, 2016, pp. 3649-3657.
[39] P. Gunarathne, H. Rui, A. Seidmann, When social media delivers customer service: Differential customer treatment in the airline industry, *MIS Quarterly*, Vol. 42, No. 2, pp. 489-520, June, 2018.

Biographies

Hongting Tang received her Ph.D. from South China University of Technology, Guangzhou, China. She is currently an Assistant Professor in School of Management, Guangdong University of Technology, Guangzhou, China. Her research has been published in leading journals, such as IEEE TEM, IEEE TCSS, among others.

Yanlin Zhang is an associate professor in the School of Management, Guangdong University of Technology, Guangzhou, China. He received his Ph.D. from Sun Yat-sen University, China. His research has appeared in several leading information technologies/information systems journals, such as Information Systems Research, EEE TCSS, ICIS, among others.

Xianyun Lin received the B. Eng degree in information security from Guangdong University of Technology, Guangzhou, China, in 2021. His research interests include network security, data mining, and cloud computing.

Lanteng Wu received his Master degree in management science and engineering from South China University of Technology (SCUT). He is currently a Ph.D. candidate with School of Management, Huazhong University of Science and Technology (HUST). His current research interests include data-driven decision-making, advertising decision and reinforcement learning.