Research Article Open Access

Chenliang Li*, Shiqian Chen, Yan Qi

Filtering and Classifying Relevant Short Text with a Few Seed Words

https://doi.org/10.2478/dim-2019-0011
received June 19, 2019; accepted August 1, 2019.

Abstract: Filtering out irrelevant documents and classifying the relevant ones into topical categories is a *de facto* task in many applications. However, supervised learning solutions require extravagant human efforts on document labeling. In this paper, we propose a novel **s**eed-guided topic model for dataless **s**hort text **c**lassification and **f**iltering, named SSCF. Without using any labeled documents, SSCF takes a few "seed words" for each category of interest, and conducts short text filtering and classification in a weakly supervised manner. To overcome the issues of data sparsity and imbalance, the short text collection is mapped to a collection of pseudo-documents, one for each word. SSCF infers two kinds of topics on pseudo-documents: *category-topics* and *general-topics*. Each category-topic is associated with one category of interest, covering the meaning of the latter. In SSCF, we devise a novel word relevance estimation process based on the seed words, for hidden topic inference. The dominating topic of a short text is identified through post inference and then used for filtering and classification. On two real-world datasets in two languages, experimental results show that our proposed SSCF consistently achieves better classification accuracy than state-of-the-art baselines. We also observe that SSCF can even achieve superior performance than the supervised classifiers supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) on some testing tasks.

Keywords: dataless text classification, short text, topic modeling, seed word, pseudo-document

***Corresponding author: Chenliang Li,** School of Cyber Science and Engineering, Wuhan University, Wuhan, China, E-mail: cllee@whu.edu.cn
Shiqian Chen, Yan Qi, School of Cyber Science and Engineering, Wuhan University, Wuhan, China

1 Introduction

A sheer volume of textual information is generated everyday. The information overload has already become a critical trouble for both individual person and organizations. Nowadays, it is common to perform a focused and deep analysis on the documents from some specified categories rather than the whole collection. That is, filtering irrelevant information and organizing relevant information into meaningful topical categories is a prerequisite for many applications. For example, a company may need to track an emerging event from a large text corpus for decision support (Wang et al., 2016). This emerging information need would dynamically change: a security issue (Ritter et al., 2015), an accident (Zhang et al., 2018), or a particular topic (Wang et al., 2016). A typical solution is text classification through supervised learning, which requires human annotation. When dealing with short texts, labeling documents become less effective because of the shortness of the documents and limited coverage of each short text in the category topical space. However, a large amount of information nowadays is indeed in the form of noisy short texts, including questions/answers from community question answer (QA) websites (Cao et al., 2010; Omari et al., 2016), web search snippets (Scaiella et al., 2012; Stein et al., 2012), tweets and comments (Naveed et al., 2011; Nishida et al., 2012; Efron et al., 2014), to name a few.

Recent studies on dataless text classification show promising results on reducing labeling effort (Liu et al., 2004; Druck et al., 2008; Chang et al., 2008; Hingmire et al., 2013; Hingmire and Chakraborti, 2014; Song and Roth, 2014; Chen et al., 2015; Li et al., 2016b; Li et al., 2018; Li et al., 2019a; Shalaby and Zadrozny, 2019). Without any labeled documents, a dataless classifier performs text classification by using a small set of relevant words for each category (called "seed words") or resorting to hidden topic labeling. The classification is mainly accomplished with seed words through word co-occurrences. However, *limited word co-occurrence information* is available in short texts compared to long documents because of the

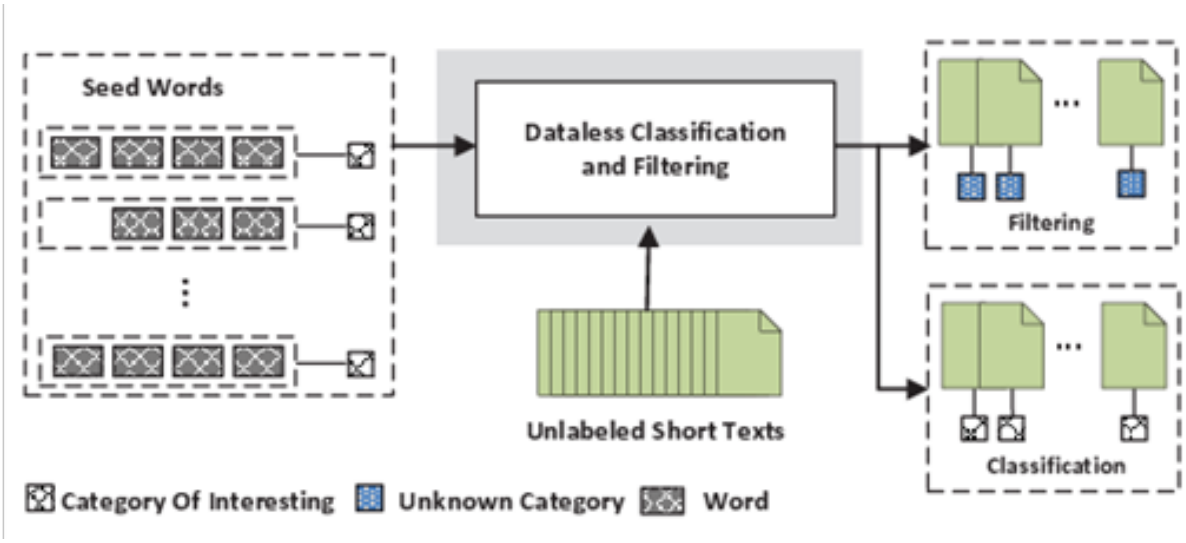


Figure 1. Illustration for dataless short text filtering and classification.

shortness. In addition to data sparsity, *existing dataless methods mainly do not consider document filtering*. That is, the existing dataless solutions perform classification over the whole collection. On the other hand, the demand to perform focused analysis on some particular topics are becoming increasingly prevalent. This is particularly useful when the information need (*i.e.*, the categories of interest) is dynamic and the short text collection is large. Focusing on a small number of relevant categories leads to another challenge: *data imbalance*. In other words, documents covered by the categories of interest could be a small proportion of the whole collection, since all other documents are considered irrelevant. This uneven distribution causes significant performance degradation for many classification algorithms (Abdi and Hashemi, 2016).

In this study, we aim to develop a dataless technique for the task of *filtering and classifying short texts* into categories of interest. An illustration for this dataless filtering and classification task for short texts is provided in Figure 1. More formally, given a short text collection of D documents and C categories of interest, where each category c is defined by a small set of seed words S_c, the goal is to filter out the documents irrelevant to any of the c categories of interest, and to classify the relevant documents to C categories of interest, without using any labeled documents.

In this paper, we propose a novel **s**eed-guided topic model for dataless **s**hort text **c**lassification and **f**iltering (SSCF). SSCF is built on the basis of a recent topic model for short texts named *word network topic model* (WNTM) (Zuo et al., 2016). As in word network topic model (WNTM), SSCF does not learn hidden topics from short texts directly, but from pseudo-documents created from the short texts. Specifically, a pseudo-document is created for each word by aggregating the words that appear in the contexts of that word. In this sense, the relevant words with respect to the semantics of a word would have high occurrences in its pseudo-document. SSCF achieves the goal of filtering and classification by modeling two kinds of topics over the pseudo-documents: *category-topics* and *general-topics*. Given C categories of interest, there are C corresponding category-topics, where each category-topic is associated with one category of interest. Through the contained relevant words, a category topic conveys the meaning of that category. The general-topics, on the other hand, absorb the semantics of the irrelevant short texts and noise. After inferring hidden topics, SSCF applies post inference to derive the dominating topic of each short document, which is used for document classification and filtering.

The challenge of SSCF is: how to precisely infer the category-topics and general-topics by using only a small number of seed words for each category of interest. In other words, seeking useful supervision from the seed words to infer relevant words becomes vital to the efficacy of SSCF. However, it is challenging to measure word relevance to a specific category directly via a small set of seed words, because of the sparse and imbalanced nature of short documents. In SSCF, we exploit the hidden topics learnt by the word network topic model (WNTM) over the same collection of pseudo-documents. Such hidden topics serve as the auxiliary knowledge to regulate the topic learning process in SSCF. On two real-world datasets

in two languages, experimental results show that the proposed SSCF consistently achieves better classification accuracy than state-of-the-art dataless baselines in terms of F_1. We also observe that SSCF can even achieve superior performance to supervised classifiers supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) on some specific tasks. To summarize, the main contributions of this paper are:

I. We propose and formalize a new task of dataless short texts classification and filtering. To the best of our knowledge, this is the first work to classify short documents into relevant categories of interest, and filter out irrelevant documents in a dataless manner.
II. We propose a novel seed-guided topic model for dataless short text classification and filtering (SSCF). In addition, we develop a novel word relevance estimation method solely based on seed words. With the estimated relevance as a prior knowledge, SSCF extracts category-topics and general-topics underlying the short document collection in a weakly supervised manner, enabling effective document classification and filtering via a post inference process.
III. We conduct extensive experiments to evaluate SSCF on two real-world short text collections in two languages. The experimental results show that SSCF achieves promising classification and filtering performance, and outperforms strong baselines. Moreover, compared with supervised learning techniques, SSCF even surpasses the supervised classifiers in many settings.

2 Related Work

Our research is related to the studies of both short text classification and dataless text classification. In the following, we briefly review works in these two areas.

Short Text Classification. The lack of information redundancy has rendered short text classification an active research topic in the past years. Instead of using bag-of-words (BOW) representation alone, Phan *et al*. proposed to augment each short text in terms of a hidden topic distribution from a knowledge base (Phan et al., 2008). They applied standard latent dirichlet allocation (LDA) to extract hidden topics from Wikipedia corpus. Then, these topics along with bag-of-words (BOW) were utilized to represent the high-level semantics of a short document for classification. Chen *et al*. further enhanced this topical representation approach by learning multi-granularity topics (Chen et al., 2011). Several works resorted to the external knowledge base to conduct short text enrichment for better classification. Long *et al*. utilized a transfer learning approach to exploit both labeled documents and unlabeled documents from Wikipedia (Long et al., 2012). Wang *et al*. proposed to represent a short document in terms of concepts covered by Probase (Wang et al., 2014). Then, a concept-based similarity measure is used to classify the given short text into the related category. Sun proposed a retrieval based approach for short text classification (Sun, 2012). Zhang *et al*. proposed a semi-supervised approach to classify the short texts (Zhang et al., 2013). Ghosh and Desarkar proposed a supervised *learning based term weighting* solution to address the shortness nature of short texts (Ghosh and Desarkar, 2018). The proposed solution achieves promising classification performance on different disaster related collections. Recently, several works are proposed to utilize neural networks for short text classification. Zeng *et al*. proposed a novel topic memory network to encode category relevant representation (Zeng et al., 2018). The topic inference and document classification are jointly performed in their model. Wang *et al*. proposed a knowledge enhanced neural network for short text classification (Wang et al., 2017). The conceptual information provided by an external knowledge base (*i.e*., Probase) is exploited to derive the representation of each short document. Very recently, Li *et al*. proposed a supervised technique by utilizing the word embeddings and a regularized word mover's distance (Li et al., 2019b). In their work, a representative semantic centroid for each category is built for document classification. Overall, these existing works on short text classification were developed in the paradigm of supervised learning. In the following, we describe the works of dataless classification techniques which require no labeled documents for training.

Dataless Text Classification. As requiring no training instances, dataless text classification has attracted much attention from the research community. In earlier days, dataless classifiers were mainly built by following semi-supervised methodology. Liu *et al*. proposed a semi-supervised Naive Bayes classifier based on expectation maximization algorithm (NB-EM) (Liu et al., 2004). The initial training instances are labeled by employing a few seed words relevant to each category. Similarly, Gliozzo *et al*. utilized a support vector machine (SVM) in a semi-supervised manner (Gliozzo et al., 2009). Downey and Etzioni provided a theoretical study confirming the possibility of achieving accurate classification in the absence of training data (Downey and Etzioni, 2008). Druck *et al*. proposed a seed-guided maximum entropy based dataless text classifier, named GE-FL (Druck et al., 2008). GE-FL assumes that the document containing the

seed words of a category is likely to be of that category. This constraint guides the parameter learning process of GE-FL.

Several works tried to exploit an external knowledge base to accomplish dataless classification (Chang et al., 2008; Song and Roth, 2014). However, external knowledge base like Wikipedia is not always available for many languages or domains. Such approaches may not be applicable in all cases. Recently, topic model based approaches for dataless text classification have been proposed (Hingmire et al., 2013; Hingmire and Chakraborti, 2014; Chen et al., 2015; Li et al., 2016b). Hingmire *et al*. proposed three topic annotation based dataless text classifiers based on standard Latent Dirichlet Allocation (LDA) (Hingmire et al., 2013; Hingmire and Chakraborti, 2014). The most effective classifier (named TLC++) allowed each topic to be associated with multiple relevant categories. The classification decisions from a mid-stage were utilized to derive more relevant words for each category. These relevant words are then incorporated with TLC++ for further classification refinement. Experimental results showed that TLC++ outperforms other two classifiers and GE-FL. Similarly, Chen *et al*. proposed a seed guided topic model for dataless classification, named DescLDA. In DescLDA, each category is associated with a fixed number of topics. The learning of these topics was then supervised based on the seed words of each category. Li *et al*. proposed a seed-guided topic model (named STM) for dataless text classification (Li et al., 2016b). Two separate sets of topics are modeled by STM: *category-topics* and *general-topics*. While general-topics are used to represent the general semantic information underlying the whole collection, category-topics are used in STM to determine the category of each document. Their experimental results demonstrated that STM outperforms DescLDA, TLC++, GE-FL, and a supervised topic model for classification (sLDA), and is more robust than these competitors. They also showed that STM even achieves very close or better performance than SVM in many tasks. This suggests that the setting of two separate sets of topics, *i.e.*, category-topics and general-topics, has potential to achieve good classification accuracy in dataless text classification. Further extension on the basis of STM was proposed in (Li et al. 2019a), named DFC. DFC was devised to support document filtering with respect to the categories of interest, by introducing another set of irrelevant-topics. The irrelevant documents are then modeled in DFC with the irrelevant-topics. Similar to STM, each irrelevant document is associated with a single irrelevant-topic and a set of general-topics. Li *et al*. further exploited modeling the document similarity for dataless document classification (Li et al., 2018). However, the similarity measure over short documents is itself a challenging research topic (Long et al., 2012). Shalaby and Zadrozny proposed a concept embedding solution for dataless document classification (Shalaby and Zadrozny, 2019). They first identified the concepts mentioned in a document with Wikipedia. Then, the concept embeddings are derived with the Skip-gram embedding model (Mikolov et al., 2013). The representations of bag-of-concepts are then built for documents and categories for dataless classification.

Although the above dataless classifiers achieved promising classification accuracy, all these techniques were developed for regular and long documents, *i.e.*, not short texts. Moreover, most solutions were developed only for document classification without filtering. That is, they do not provide the function for irrelevant document filtering.

To the best of our knowledge, the proposed SSCF is the first work to address short text filtering and classification in a dataless manner. Because of the good performance of STM and DFC due to the two separate sets of topics, SSCF also infers two sets of topics: *category-topics* and *general-topics*. Different from STM and DFC, the general-topics in SSCF are for filtering irrelevant documents, not for representing the general semantic information underlying the whole collection as in STM. Moreover, the two sets of topics are inferred directly from the text collection in STM. In SSCF, the two sets of topics are inferred from the collection of pseudo-documents created from the short texts.

3 Preliminary

The proposed SSCF model is built on the basis of the word network topic model (WNTM). To better understand SSCF, we now give a brief introduction to the word network topic model (WNTM).

The word network topic model (WNTM) is a probabilistic topic model to tackle the sparsity and imbalance problems for short texts (Zuo et al., 2016). Different from other topic modeling techniques for short texts that either adopt additional constraints or directly model word co-occurrence within the generative process, the word network topic model (WNTM) builds a pseudo-document collection from the short texts and adopts the standard latent dirichlet allocation (LDA) to infer hidden topic distributions.

Given a word w, it is intuitive that the topic distribution $p(z|w)$ strongly reflects the contexts that w appears in.

Figure 2. Pseudo-document generation for word "apple" in three short texts, with the sliding window size being 5.

Therefore, the word network topic model (WNTM) adopts a sliding window to scan each short document and takes all the words that appear together with w within the window to form a pseudo-document d_w. The frequency of word w' in the pseudo-document d_w is the co-occurrence count of w' with w within the sliding windows over the whole collection. Figure 2 illustrates the pseudo-document generation process with a window size of 5 over three short documents for the word "apple". Note that stop words are removed from the pseudo-document. As shown in Figure 2, there are 5, 1 and 3 sliding windows containing "apple" from the three short documents respectively. We can also observe that the frequent words (*e.g.*, "macbook", "iphone", "watch") in the pseudo-document d_{apple} show strong connection to the semantics of word "apple". Given that the pseudo-document d_w is relatively longer than the short documents and is comprised the semantically relevant words frequently co-occurring with the word w, any effective topic models over the regular documents can then be applied. After generating the pseudo-documents for all the words, the word network topic model (WNTM) adopts the standard latent dirichlet allocation (LDA) model to infer the topic word distribution $p(w|z)$ for each topic and the pseudo-document topic distribution $p(z|d_w)$. The topic distribution for a short document d can then be calculated[1] as follows:

$$p(z = k|d) \propto \sum_w p(z = k|w)p(w|d) \quad (1)$$

$$p(z = k|w) = \frac{p(z = k)p(w|z = k)}{\sum_i^K p(z = i)p(w|z = i)} \quad (2)$$

where $p(w|d)$ is estimated based on the relative frequency of w in the document.

Because the relevant words to a rare topic are expected to have relatively long pseudo-documents compared to the original short documents where they appear, the word network topic model (WNTM) has shown its superiority in detecting rare topics underlying the document collection (Zuo et al., 2016). In this sense, the word network topic model (WNTM) becomes a natural choice as a building block to derive a topic model for dataless short text classification and filtering.

4 The SSCF Model

In this section, we present the proposed **s**eed-guided topic model for **s**hort texts **c**lassification and **f**iltering, SSCF.

SSCF is a probabilistic topic model that infers hidden topics over the pseudo-documents. Illustrated in Figure 3, it consists of two components: *word relevance estimation* and *short text classification and filtering*. Given a set of seed words for each category of interest, we estimate the relevance score between a word and a category based on the inferred hidden topics by the word network topic model (WNTM) over the short text collection. The resultant relevance estimation serves as prior knowledge

1 In the original work (Zuo et al., 2016), the authors used $p(z|w)=p(w|z)$ to approximate $p(z|d)$ in Equation 1. Here, we rigidly derive $p(z|w)$ by following the Bayes rule.

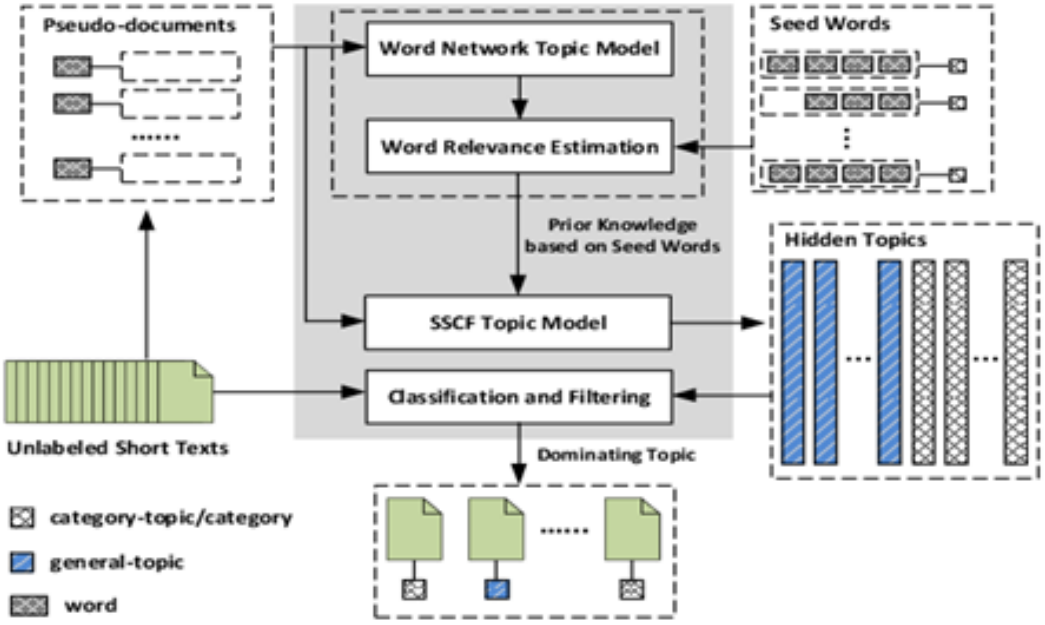


Figure 3. Overview of SSCF model.

for SSCF to supervise the topic inference over two kinds of topics: category-topics and general-topics. At last, the short documents are classified and filtered by the post inference of document topic distribution.

4.1 Model Generative Process and Inference

To identify the documents in categories of interest, SSCF needs to discover the category relevant topics and irrelevant topics based on the seed words. To support classification, it then becomes natural to represent each category as a category-topic. Here, we make a one-to-one correspondence between the categories and category-topics such that each category is associated with a single category topic and vice versa[2]. Each category-topic contains relevant words of the corresponding category. To filter out the documents irrelevant to any category of interest, we introduce T general-topics over the document collection to model the semantics covered by the irrelevant documents. In short, there are C category-topics and T general-topics in SSCF. A word can be generated by either a category-topic or a general-topic. We utilize a switch variable x to indicate which kind of topics the word is associated with. The classification and filtering are then achieved by inferring the dominating topic of each short document: *whether the dominating topic of a short document is a category topic and which category topic?*

Because of the way pseudo-documents are constructed, a pseudo-document typically has very few dominating category-topics. That is, the category-topic distribution should be highly skewed for most pesudo-documents. Motivated by the sparsity modeling strategy proposed in (Lin et al. 2014), we utilize a *weak smoothing prior* to decouple the sparsity and the smoothness of a category-topic distribution for a pseudo-document in SSCF. Given a normal word w (*i.e.*, not a seed word), let $\vec{\eta}_w$ denote the vector of length C such that $\vec{\eta}_w(c) \in [0,1)$ is the estimated relevance score between the word w and category c, the prior category-topic distribution of pseudo-document d_w is defined by $\delta\vec{\eta}_w + \vec{\epsilon}$, where δ works as the prior concentration parameter, ϵ is a weak smoothing prior, and $\vec{\epsilon}$ is the vector with all elements being ϵ. The impact of the weak smoothing prior ϵ is important in the probabilistic view. The value of $\vec{\eta}_w(c)$ would be zero, indicating that no word in the corresponding pesudo-document can be generated by the category topic c. This case would happen when we have an empty relevance estimated for word w and category c (ref. Equations 3 and 6). The introduction of ϵ avoids this zero probability, making the whole process mathematically correct.

As to the pseudo-document d_s corresponding to a seed word s, we set the category-topic prior distribution

2 Category and category-topic are considered equivalent and exchangeable in this work when the context is clear.

to be $\delta\vec{1}_c + \vec{\epsilon}$, where $\vec{1}_c$ is the indicator vector such that the element corresponding to category c is 1 and other elements are 0. By estimating the relevance scores $\vec{\eta}_w$, SSCF is expected to derive the category-topics correctly through word co-occurrence information within the pseudo-documents, for effective classification and filtering. Let s_c denote the seed word set of category c. The generative process is described as follows:

(1) For each general topic $t \in \{1...T\}$:
(a) draw a word distribution $\phi_t \sim \text{Dir}(\beta_1)$;
(2) For each category $c \in \{1...C\}$:
(a) draw a word distribution $\vartheta_c \sim \text{Dir}(\beta_0)$;
(b) for each seed word $s \in S_c$;
b.1 draw category-topic distribution $\varphi_s \sim \text{Dir}(\delta\vec{1}_c + \vec{\epsilon})$;
b.2 draw general-topic distribution $\psi_s \sim \text{Dir}(\alpha)$;
(3) For each category $c \in \{1...C\}$:
(a) for each seed word $s \in S_c$:
a.1 for each word $w \in d_s$;
a.1.1 draw a general-topic $z_w^0 \sim \text{Multi}(\psi_s)$;
a.1.2 draw a category-topic $z_w^1 \sim \text{Multi}(\varphi_s)$;
a.1.3 draw switch variable $x_w \sim \text{Bernoulli}(\pi_{w,z_w^1})$;
a.1.4 if x_w=0:
* $z_w \leftarrow z_w^0$;
* draw word $w \sim \text{Multi}(\phi_{z_w})$;
a.1.5 if x_w=1:
* $z_w \leftarrow z_w^1$;
* draw word $w \sim \text{Multi}(\vartheta_{z_w})$;
(4) For each normal word $w \in \mathbf{V} \setminus \mathbf{S}$;
(a) draw category-topic distribution $\varphi_w \sim \text{Dir}(\delta\vec{\eta}_w + \vec{\epsilon})$;
(b) draw general-topic distribution $\psi_w \sim \text{Dir}(\alpha)$;
(c) for each word $w' \in d_w$:
c.1 draw a general-topic $z_{w'}^0 \sim \text{Multi}(\psi_w)$;
c.2 draw a category-topic $z_{w'}^1 \sim \text{Multi}(\varphi_w)$;
c.3 draw switch variable $x_{w'} \sim \text{Bernoulli}(\pi_{w',z_{w'}^1})$;
c.4 if $x_{w'}$=0:
* $z_{w'} \leftarrow z_{w'}^0$;
* draw word $w' \sim \text{Multi}(\phi_{z_{w'}})$;;
c.5 if $x_{w'}$=1:
* $z_{w'} \leftarrow z_{w'}^1$;
* draw word $w' \sim \text{Multi}(\vartheta_{z_{w'}})$;

The switch variable x_w=1 indicates that word w is generated from a category-topic, otherwise w is generated from a general-topic. Parameter z_w then stores the specific hidden topic selected to generate word w according to the choice of x_w. The document collection is likely to be imbalanced, because the number of relevant pseudo-documents under each category could vary tremendously. Setting a constant prior to the choice of x_w for all words is not an appropriate mechanism. Because of the probabilistic nature, the relevant words with high frequency in the larger categories would become the dominating words in the general-topics. The resultant high similarity between the corresponding category-topics and general-topics will hurt the classification performance. As described in the generative process, a word and category dependent prior probability π_{w,z_w^1} is utilized in the SSCF to guide the decision of x_w. π_{w,z_w^1} works as the prior probability that word w is generated by category-topic z_w^1 against any general-topic. Similar to the setting of $\vec{\eta}_w$, we calculate the prior value π_{w,z_w^1} based on the relevance between a word and a category. The graphical representation of SSCF is shown in Figure 4.

Seed Words-based Relevance Estimation. Precise relevance estimation between a word and a category plays an important role in SSCF. Given the length of a short document is limited, the data sparsity problem obstructs the precise relevance estimation directly based on word co-occurrences. We propose to estimate the relevance based on the learned topics by using the word network topic model (WNTM). However, doing this leads to a new problem. The word network topic model (WNTM) infers hidden topics by using standard latent dirichlet allocation (LDA) over pseudo-documents. As the result, common words in the pseudo-document collection are usually ranked high in many topics. To address this issue, we adopt a topic keyword re-ranking approach proposed in (Song et al. 2009) to assign a weight for each word under a topic. Specifically, the weight $\omega(w,k)$ for word ω under topic k is calculated as follows:

$$\omega(w,k) = \frac{p_{wntm}(w|k)}{\sum_i^K p_{wntm}(w|z=i)} \tag{3}$$

where $p_{wntm}(w|k)$ is the word probability for w under topic k from the word network topic model (WNTM). We then calculate the relevance $rel(w,c)$ between word w and category c as follows:

$$rel(w,c) = \frac{1}{|\mathbf{S}_c|}\sum_{s\in\mathbf{S}_c}\sum_k^K \omega(w,k)p_{wntm}(k|d_s) \tag{4}$$

where $p_{wntm}(k|d_s)$ is the estimated topic distribution for pseudo-document d_s by the word network topic model (WNTM). However, the raw relevance score calculated in Equation 4 is category-dependent. That is, the scale of the relevance score depends on a specific category. It is necessary to normalize the relevant scores and then to calculate $\vec{\eta}_w$ as follows:

$$r(w,c) = \frac{rel(w,c)}{\sum_{w'} rel(w',c)} \tag{5}$$

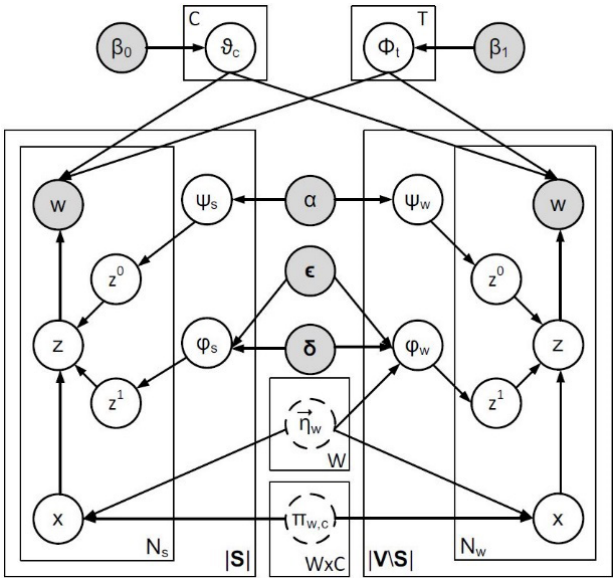


Figure 4. Graphical representation of SSCF.

$$\vec{\eta}_w(c) = \frac{r(w,c)}{\sum_{c'} r(w,c')} \tag{6}$$

In short, we first normalize the relevance scores of all words under each category in Equation 5. We then calculate $\vec{\eta}_w$ by applying *L*1 normalization in Equation 6.

Since a category relevant word *w* could have very few dominating category-topics in its pseudo-document, the corresponding category-dependent prior probability $\pi_{w,c}$ is expected to be highly biased. Therefore, we calculate $\pi_{w,c}$ by filtering out the less relevant categories based on $\vec{\eta}_w$:

$$\pi_{w,c} = \frac{\sigma \vec{v}_w(c)}{\sigma \vec{v}_w(c) + (1-\sigma)\tau_w} \tag{7}$$

$$\tau_w = 1 - \sum_c \vec{v}_w(c) \tag{8}$$

$$\vec{v}_w(c) = \max\left(\vec{\eta}_w(c) - \frac{1}{C}, \upsilon\right) \tag{9}$$

In Equation 9, we filter out the irrelevant categories that have $\vec{\eta}_w(c)$ below average, resulting in a sparse $\vec{\eta}_w(c)$ for each word. A much smaller constant υ is assigned to these irrelevant categories to avoid being zero. In Equation 8, $\tau_w \in (1/C,1)$ measures the category specificity of word *w*. Based on Equation 9, when word *w* is highly relevant to a single category, $\vec{v}_w$ has a single peak in that category. In this case, $\tau_w \to 1/C$.

In contrast, when *w* is relevant/irrelevant to all categories, $\vec{v}_w$ is close to a vector containing all zeros, leading to $\tau_w \to 1$. In the calculation of prior probability $\pi_{w,c}$ by Equation 7, σ serves as a tuning parameter in the range of [0,1] to control the importance of both category specificity and relevance. Note that τ_w is a word dependent measure. Given $\vec{v}_{w_1}(c) = \vec{v}_{w_2}(c)$ for two words w_1 and w_2, the word with higher category specificity (*i.e.*, a smaller τ_w) will have a larger $\pi_{w,c}$. When $\sigma = 1$, SSCF comprises only category-topics, which in turn loses the ability to filter out irrelevant documents. When $\sigma = 0$, SSCF is equivalent to the word network topic model (WNTM) such that the standard latent dirichlet allocation (LDA) is conducted over the pseudo-documents for hidden general-topic inference (*i.e.*, no classification). Because $\pi_{w,c}$ and $\vec{\eta}_w$ are both estimated based on seed words, the two sets of variables are then plotted as dotted circles in Figure 4.

Note that $\vec{\eta}_w(c)$ and $\vec{v}_w(c)$ in Equations 6 and 9 are calculated for $C > 1$. When there is only one category of interest (*i.e.*, $C = 1$), we may use the raw relevance score *rel(w, c)* instead: $\vec{\eta}_w(c) = \vec{v}_w(c) = rel(w,c)$. However, since no normalization can be applied based on Equations 5 and 6, we observe that most common words with high frequencies are ranked very high in terms of *rel(w, c)*. These common words obstruct the precise learning of category-topics and general-topics. Hence, we restrict the calculation of *rel(w, c)* based on the top-*N* topics in terms of $p_{wntm}(k|d_s)$. Then, $\vec{\eta}_w(c)$ is calculated by max-normalization over *rel(w,c)*: $\vec{\eta}_w(c) = rel(w,c)/\max_{w'} rel(w',c)$, and $\vec{v}_w(c)$ is set as $\vec{v}_w(c) = \vec{\eta}_w(c)$. In this work, we use $N = 3$.

Inference by Gibbs Sampling. We utilize Gibbs Sampling to perform approximate inference and parameter learning (Thomas and Mark, 2004). In SSCF, for a word d_w^i at the position *i* in pseudo-document d_w of a normal word *w*, we jointly sample $z_{d_w^i}$ and $x_{d_w^i}$ as follows ($z^0_{d_w^i}$ and $z^1_{d_w^i}$ are collapsed):

$$p(z_{d_w^i}, x_{d_w^i} | z_{\neg d_w^i}, x_{\neg d_w^i}, w) \propto$$

$$\begin{cases} \phi^{d_w^i}_{t,\neg d_w^i} \cdot \psi^t_{d_w,\neg i} \cdot \sum_c^C \varphi^c_{d_w,\neg i}\left(1 - \pi_{d_w^i,c}\right) & z_{d_w^i} = t, x_{d_w^i} = 0 \\ \vartheta^{d_w^i}_{c,\neg d_w^i} \cdot \varphi^c_{d_w,\neg i} \cdot \pi_{d_w^i,c} & z_{d_w^i} = c, x_{d_w^i} = 1 \end{cases} \tag{10}$$

where $\phi^{d_w^i}_{t,\neg d_w^i}$ is the probability of seeing d_w^i under the general-topic *t* excluding the current assignment, $\psi^t_{d_w,\neg i}$ is the probability of seeing the general-topic *t* in the document d_w excluding the current assignment, $\vartheta^{d_w^i}_{c,\neg d_w^i}$ is the probability of seeing the word d_w^i under the category-topic *c* excluding the current assignment, and $\varphi^c_{d_w,\neg i}$ is the probability of seeing the category-topic *c* in the document d_w excluding the current assignment. Similarly, for a word d_s^i at the position *i* in the pseudo-document d_s of the seed word *s*, we jointly sample $z_{d_s^i}$ and $x_{d_s^i}$ as follows:

$$p(z_{d_s^i}, x_{d_s^i} | z_{\neg d_s^i}, x_{\neg d_s^i}, w) \propto$$

$$\begin{cases} \phi^{d_s^i}_{t,\neg d_s^i} \cdot \psi^t_{d_s,\neg i} \cdot \sum_c^C \varphi^c_{d_s,\neg i} \left(1 - \pi_{d_s^i,c}\right) & z_{d_s^i} = t, x_{d_s^i} = 0 \\ \vartheta^{d_s^i}_{c,\neg d_s^i} \cdot \varphi^c_{d_s,\neg i} \cdot \pi_{d_s^i,c} & z_{d_s^i} = c, x_{d_s^i} = 1 \end{cases} \quad (11)$$

where $\varphi^c_{d_s,\neg i}$ is the probability of seeing the category-topic c in the document d_s excluding the current assignment.

Based on the point estimation, the word distributions $\{\vartheta_c, \phi_t\}$ and topic distributions $\{\varphi_{dw}, \varphi_{ds}, \psi_{dw}, \psi_{ds}\}$ can be computed as shown in Table 1, where n_c^w is the number of times word w being assigned to the category-topic c, n_t^w is the number of times the word w being assigned to the general-topic t, $n^c_{d_w}$ is the number of words that are assigned to the category-topic c within the pseudo-document d_w, $\vec{\eta}_w(c)$ is the element value corresponding to the category c in $\vec{\eta}_w$, $n^c_{d_s}$ is the number of words that are assigned to the category-topic c within the pseudo-document d_s, $n^t_{d_w}$ is the number of words that are assigned to the general-topic t within the pseudo-document d_w, $n^t_{d_s}$ is the number of words that are assigned to the general-topic t within the pseudo-document d_s, and $\mathrm{II}_{c,s}$ is an indicator such that 1 is returned when s is a seed word of category c (*i.e.*, $s \in \mathbf{S}_c$), otherwise 0 is returned.

Table 1
Point Estimation for Word Distributions and Topic Distributions.

$\vartheta_c^w = \frac{n_c^w + \beta_0}{\sum_{w'}^V (n_c^{w'} + \beta_0)}$	$\varphi^c_{d_w} = \frac{n^c_{d_w} + \vec{\eta}_w(c)\delta + \epsilon}{\sum_{c'}^C (n^{c'}_{d_w} + \vec{\eta}_w(c')\delta + \epsilon)}$
$\psi^t_{d_w} = \frac{n^t_{d_w} + \alpha}{\sum_{t'}^T (n^{t'}_{d_w} + \alpha)}$	$\varphi^c_{d_s} = \frac{n^c_{d_s} + \mathbb{I}_{c,s}\delta + \epsilon}{\sum_{c'}^C (n^{c'}_{d_s} + \mathbb{I}_{c',s}\delta + \epsilon)}$
$\phi_t^w = \frac{n_t^w + \beta_1}{\sum_{w'}^V (n_t^{w'} + \beta_1)}$	$\psi^t_{d_s} = \frac{n^t_{d_s} + \alpha}{\sum_{t'}^T (n^{t'}_{d_s} + \alpha)}$

4.2 Document Classification and Filtering

After applying SSCF over the pseudo-documents of all words (including seed words and normal words), we classify and filter short documents by inferring the topic distribution of each document over C category topics and T general topics. Specifically, the word topic distribution $p(z|w)$ over all category topics and general topics is calculated as follows:

$$p(k|w) = \frac{p(z=k)p(w|z=k)}{\sum_i^C p(z=i)p(w|z=i) + \sum_i^T p(z=i)p(w|z=i)} \quad (12)$$

The dominating topic $\hat{z}$ of document d is then used to indicate its category $\hat{z} = \mathrm{argmax}_z p(z|d)$, where $p(z|d)$ is calculated by using Equation 1. If $\hat{z}$ is a category-topic (*i.e.*, $\hat{z} \in C$), the corresponding category is assigned to document d. Otherwise, document d is considered as being irrelevant to any category of interest.

5 Experiment

In this section, we conduct experiments on two real-world short text collections to evaluate the filtering and classification performance of the proposed SSCF[3]. To simulate the filtering and classification task (*i.e.*, classification with filtering), we take the documents from one or several pre-selected categories as relevant documents, and the remaining documents in the collection as irrelevant documents. The task is to identify the categories of the relevant documents, and filter out the irrelevant ones. We also evaluate the classification performance of SSCF in terms of conventional classification without filtering.

5.1 Experimental Setting

Datasets. Here, we utilize two publicly accessible datasets, Web Snippet (WS) and Baidu question answer (BaiduQA), for performance evaluation of both classification with filtering and classification without filtering tasks.

Web Snippet (WS) is a widely used collection for short text classification study (Chen et al., 2011; Phan et al., 2008; Sun, 2012; Li et al., 2016a). It consists of 12,340 English Web search snippets, each of which belongs to one of 8 categories. The collection was constructed by performing Web search using different phrases for each category. For each search phrase, top 20 or 30 snippets were selected. The snippets were further divided into a training set and a test set such that the search phrases used for the two subsets are exclusive. That is, the snippets in the test set are considered difficult to classify, even for a supervised classifier (Phan et al., 2008). We perform the preprocessing by following steps used in (Li et al. 2016a).

3 The implementation will be released after paper acceptance.

Baidu question answer (BaiduQA) is a collection of 648,514 questions crawled from a popular Chinese question and answer (Q&A) website[4]. This collection has been previously used in the related works about short text topic modeling (Yan et al., 2013; Cheng et al., 2014; Li et al., 2016a). Based on the annotation made by its asker, each question belongs to one of 35 categories. The texts were already tokenized by using Chinese word segmentation, and the duplicated words are removed, *i.e.*, each word appears only once in a short document. In our experiments, we further remove the questions that contain only a single word. We build the training and test sets by randomly sampling short documents with a ratio of 80:20.

Table 3 summarizes the detailed statistics about the two datasets after preprocessing. Note that Baidu question answer (BaiduQA) collection covers more categories and contains much shorter documents compared to the WS dataset. We can also see that the pseudo-documents are much longer than the original short texts. In common, both datasets have a very skew category distribution. The sizes of the smallest and largest categories are 370/2,660 and 1,481/7,625 for a Web Snippet (WS) and Baidu question answer (BaiduQA) dataset respectively. We further check the length of the pseudo-document (*i.e.*, d_s) of each seed word s. The average pseudo-document length of all the seed words under a category is then calculated. We find that the average lengths of the smallest and largest categories are 59.8/72.2 and 67.4/83.1 for Web Snippet (WS) and Baidu question answer (BaiduQA) datasets respectively. Note that each short text has 18.9 and 4.1 words on average for the two datasets, respectively. It is clear that building pseudo-documents alleviates both the data sparsity and imbalance problems to some extent. In our experiments, eight categories with diverse topics from each dataset and their combinations are randomly chosen for performance evaluation. Specifically, six classification with filtering tasks are created for each dataset. Table 2 reports the number of relevant documents of each classification as well as filtering task and the percentage of these documents over the whole collection. We can see that data imbalance problem is a critical challenge for all classification with filtering tasks. Particularly, the documents relevant to the categories of interest are less than 10% on the Baidu question answer (BaiduQA) dataset.

Seed Word Selection. Following the seed word selection process in descriptive latent dirichlet allocation (DescLDA) (Chen et al., 2015), we manually select the seed words for each category based on the learnt topics from the word network topic model (WNTM): (i) extract hidden topics by using the word network topic model (WNTM) over the pseudo-documents; (ii) rank the words in descending order in terms of word-distribution $p(w|z)$ for each topic; and (iii) manually select 20 semantically relevant seed words for each category based on the most probable 50 words in each topic. Note that we run the word network topic model (WNTM) over all raw short texts in the corresponding training set without access to the category information. The number of hidden topics is set as 40 here. We need to point out here that no statistic and category information are used during the seed word selection. The seed words are selected solely based on their semantic relevance to the category. Tables 6 and 7 in Appendix A report the selected seed words of each category for the two datasets. We also include the percentage of the documents that contain any seed word for each category. On average, these documents only comprise less than 10% and 20% of all the relevant ones for each category on Web Snippet (WS) and Baidu question answer (BaiduQA) datasets, respectively.

Methods in Comparison. We compare the proposed SSCF against the following state-of-the-art *dataless text classification* methods and *supervised classification* methods.

TLC++ It takes the association between the topics and categories to classify documents (Hingmire and Chakraborti, 2014). We use the recommended settings by its authors.

DescLDA It classifies documents by applying document clustering over the learned hidden topics (Chen et al., 2015). We tune the number of topics and report the best results.

STM This method learns to identify a category-topic for each document, where the topic inference process is guided by the seed words (Li et al., 2016b). We use the implementation provided by the authors and use their recommended settings.

SVM A linear support vector machine (SVM) classifier with the default parameter settings and term frequency-inverse document frequency (TF–IDF) weighting scheme is used[5].

sLDA It is a supervised text classifier based on the latent dirichlet allocation (LDA) model (Blei and McAuliffe, 2007). The [i]implementation[6] provided by the authors is used. The best results are reported after parameter tuning.

4 http://zhidao.baidu.com.

5 www.csie.ntu.edu.tw/ cjlin/liblinear.

6 http://www.cs.cmu.edu/ chongw/slda/.

Table 2
Number of Relevant Documents in the Six Classification with Filtering Tasks for Web Snippet (WS) and Baidu Question Answer (BaiduQA). # Docs: The Number of Relevant Documents in the Whole Collection. Percentage (%): the Percentage of Relevant Documents in the Whole Collection.

Dataset	Classification task	#Docs	Percentage (%)
WS	Education Science	2,660	21.7
	Business	1,598	13.0
	Computers	1,500	12.2
	Education Science vs. Engineering	3,030	24.7
	Sports vs. Business vs. Computers	4,518	36.8
	Health vs. Politics Society vs. Culture Arts Ent.	4,860	39.6
BaiduQA	Employment	7,625	4.3
	Tax	6,667	3.7
	Economics Research	4,533	2.5
	Employment-Tax	14,292	8.0
	Real Estate-Database-Economics Research	15,804	8.8
	Poem-Tax-Western Popular Music	16,082	9.0

Table3
Statistics on the Two Datasets. #Docs: Total Number of Documents. Avg(|d|): Average Number of Words per Short Document. Avg($|d_w|$): Average Number of Words per Pseudo-document.

Dataset	#Docs	Avg(|d|)	Vocabulary	Avg($\|d_w\|$)
WS	12,265	18.89	5,581	201.68
BaiduQA	179,042	4.11	26,560	80.96

CRX It is the concept raw context model by building the bag-of-concept representations for dataless classification (Shalaby and Zadrozny, 2019).

DEC It is the seed-guided topic model for document filtering and classification (Li et al., 2019a). It was devised for performing classification and filtering for regular and long documents. Note that for performing only document filtering, DFC is equivalent to STM.

We further add a simple support vector machine (SVM) based baseline for classification with filtering as a reference. We choose to re-label each training document using the seed words. Specifically, a training document belongs to a category if it contains any seed word of that category. The training documents covering no seed word are removed from further consideration. We then train a one-class support vector machine (SVM) classifier based on this seed-guided training set. We denote this baseline as **SG-SVM**. We need to emphasize that a training document could belong to more than one category. In this case, we include the same document for all matched categories. This would inevitably introduce many noisy information and hurt the model optimization process.

Evaluation Protocol. Among the methods mentioned earlier, TLC++, DescLDA, STM, and CRX are state-of-the-art dataless techniques for document classification without filtering. However, only TLC++ can be adapted for filtering irrelevant documents. In detail, we add a pseudo category into the topic annotation process of TLC++ to indicate the irrelevance. As to support vector machine (SVM) and supervised latent dirichlet allocation (sLDA), they are supervised learning classifiers that require the labeled instances to train the model. For supervised latent dirichlet allocation (sLDA), we treat all irrelevant documents in the training set as a pseudo-category. For support vector machine (SVM), we adopt the one-class implementation provided in LIBSVM for classification with filtering. The hyper-parameters are tuned accordingly.

We train supervised classifiers using the training documents and conduct evaluation on the test set. For all dataless classifiers in comparison, all short documents in the collection are taken as input into the classifiers, as no labeled documents are needed for dataless classifiers. Note that the comparison between a dataless classifier and a supervised classifier is not *fair*, since the category information associated with the short documents in the training set are exploited by the supervised classifier.

Although a dataless classifier takes all short documents as input and performs classification and filtering for all the documents in a batch manner, no category information associated with each training document is available within the inference process. To enable a relative fair comparison, the performance of each classifier is evaluated on the test set. This evaluation protocol is also adopted in previous works (Hingmire and Chakraborti, 2014; Li et al., 2016b). Classification performance is evaluated by macro-averaged F_1 (Macro-F_1). Macro-F_1 is the averaged F_1 score of all categories. We report the average results over 5 runs for all the methods (excluding support vector machine (SVM)). The statistical significance is conducted by using the student *t-test*.

Parameter Setting. For hyper-parameter setting, we use $\alpha = \beta_0 = \beta_1 = 0.1$, $\delta = 10.0$, $\epsilon = 1.0\times10_{-2}$. For the tunable parameters σ and T in SSCF, we use the following settings. (1) For classification without filtering, we set $T = 1$. Parameter σ is set to 0.7 and 0.9 for Web Snippet (WS) and Baidu question answer (BaiduQA), respectively. (2) For classification with filtering, we set $\sigma = 0.1$, $T = 10$ for all the tasks. As to the word network topic model (WNTM) , we also set the topic number to be 40 for word relevance estimation. Regarding the pseudo-document generation, we directly take each short document as an individual sliding window. We conduct classification and filtering after running SSCF for 50 iterations.

5.2 Performance Comparison

Classification with Filtering. In this set of experiments, we need to filter out the documents irrelevant to any specified category. The Macro-F_1 scores are reported in Table 4. Here, we make the following observations.

First, SSCF achieves better filtering and classification performance on 4 out of 6 tasks on the Web Snippet (WS) dataset. The second performer is DFC, which delivers the best performance on 2 tasks and the second best performance on 1 task. Also, we note that support vector machine (SVM) achieves relative worse performance in most tasks. Given limited word overlaps between the training set and test set, support vector machine (SVM) cannot manage to filter out irrelevant documents and classify the relevant documents correctly. A similar observation is also made for seed-guided support vector machine (SG-SVM). Note that the support vector machine (SVM) performs significantly better than seed-guided support vector machine (SG-SVM) on 4 out of 6 tasks here. On the other hand, since we treat all irrelevant documents in the training set as being of a pseudo-category for supervised latent dirichlet allocation (sLDA), the supervision in terms of topic-level information used by supervised latent dirichlet allocation (sLDA) is more useful than counting on the word occurrences (*i.e.*, SVM) in this case.

Second, the support vector machine (SVM) achieves much better performance than other methods on the Baidu question answer (BaiduQA) dataset. SSCF, however, still manages to achieve the second best on 4 out of 6 tasks on this dataset. Note that Baidu question answer (BaiduQA) contains more documents and covers a much larger number of categories than the Web Snippet (WS) dataset. Therefore, data imbalance problem is more severe on the Baidu question answer (BaiduQA) dataset. All topic model based methods experience significant performance degradation here. Among all topic model based methods, SSCF obtains far better filtering and classification performance. The superiority of SSCF over other topic model based alternatives suggests that the seed-guided topic inference process over the pseudo-documents in SSCF is effective to distinguish the relevant documents from the irrelevant ones. Moreover, we observe that support vector machine (SVM) performs consistently better than seed-guided support vector machine (SG-SVM) across all six tasks. This is consistent with the observation we made for the Web Snippet (WS) dataset. The results suggest that building the training set by using seed words would introduce sufficient noise.

Overall, the experimental results demonstrate that SSCF is an effective algorithm to conduct document filtering and classification in a dataless manner.

Classification without Filtering. In this task, there is no need to filter out irrelevant documents. That is, we build a subset by selecting all documents from the specified categories for classification evaluation. This setting gives all other methods in comparison a small advantage, since only SSCF is furnished with an inherent filtering function. In this set of experiments, we include the tasks of classifying all 8 categories (All-8) and all 35 categories (All-35) for Web Snippet (WS) and Baidu question answer (BaiduQA) datasets, respectively. Since All-8 and All-35 are larger collections with more categories, we set $T = 10$ for these two tasks. Table 5 reports the Macro-F_1 scores of all the methods. We make several observations.

First, among all dataless classifiers, SSCF significantly outperforms other state-of-the-art competitors on 10 out of 14 classification tasks. STM obtains the best performance on 2 tasks and achieves the second best performance on 5 tasks. Then, descriptive latent dirichlet allocation (DescLDA) also achieves the best performance on one task and the second best performance on 5 tasks. Note that all

Table 4
The Best and Second Best Results are Highlighted in Boldface and Underlined Respectively, on Each Task. Among SSCF, TLC++ and sLDA, † Indicates That the Difference to the Best Result is Statistically Significant at 0.05 Level.

Dataset	Classification task	SSCF	DFC	TLC++	sLDA	SVM	SG-SVM
WS	Education Science	0.476†	**0.716**	0.131†	0.483	0.390	0.454
	Business	**0.485**	0.392†	0.337†	0.344†	0.380	0.271
	Computers	**0.593**	0.267†	0.126†	0.420†	0.454	0.383
	Education Science vs. Engineering	0.432†	**0.694**	0.134†	0.367†	0.269	0.297
	Sports vs. Business vs. Computers	**0.561**	0.458†	0.184†	0.516	0.439	0.375
	Health vs. Politics Society vs. Culture Arts Ent.	**0.619**	0.440†	0.450†	0.436†	0.366	0.335
BaiduQA	Employment	0.353	0.226†	0.182†	0.104†	**0.581**	0.453
	Tax	0.619	0.207†	0.067†	0.304†	**0.653**	0.455
	Economics Research	0.408	0.161†	0.015†	0.062†	**0.512**	0.271
	Employment-Tax	0.247†	0.258†	0.079†	0.285	**0.617**	0.454
	Real Estate-Database-Economics Research	0.538	0.288†	0.240†	0.288†	**0.654**	0.462
	Poem-Tax-Western Popular Music	0.460	0.218†	0.089†	0.248†	**0.570**	0.336

Table 5
The Best and Second Best Results by Dataless Classifiers are Highlighted in Boldface and Underlined Respectively, on Each Task. Among Dataless Classifiers, † Indicates That the Difference to the Best Result is Statistically Significant at 0.05 level.

Dataset	Classification task	SSCF	CRX	STM	DescLDA	TLC++	sLDA	SVM
WS	Education Science	**0.755**	0.494†	0.743	0.574†	0.705†	0.680	0.645
	Business	0.871‡	0.412†	**0.911**	0.537†	0.656†	0.745	0.824
	Computers	**0.809**	0.682†	0.787‡	0.747†	0.445†	0.900	0.771
	Education Science vs. Engineering	**0.881**	0.615†	0.875	0.804†	0.722†	0.854	0.851
	Sports vs. Business vs. Computers	**0.860**	0.681†	0.782‡	0.776†	0.671†	0.847	0.883
	Health vs. Politics Society vs. Culture Arts Ent.	0.700‡	0.472†	**0.782**	0.523†	0.565†	0.757	0.738
	All-8	**0.645**	0.385†	0.625‡	0.482†	0.456†	0.546	0.569
BaiduQA	Employment	0.690†	**0.754**	0.552†	0.725‡	0.425†	0.857	0.861
	Tax	**0.736**	0.652†	0.529†	0.690‡	0.600†	0.852	0.861
	Economics Research	**0.868**	0.765†	0.529†	0.823‡	0.763†	0.889	0.909
	Employment-Tax	0.856	0.811†	0.449†	**0.865**	0.703†	0.929	0.885
	Real Estate-Database-Economics Research	**0.832**	0.789†	0.601†	0.810†	0.565†	0.894	0.876
	Poem-Tax-Western Popular Music	**0.867**	0.770†	0.395†	0.786†	0.365†	0.913	0.953
	All-35	**0.463**	0.436†	0.108†	0.261†	0.316‡	0.364	0.128

CRX, STM, DescLDA, and TLC++ were developed based on long documents. The superior classification performance suggests that topic inference over the pseudo-documents is an effective strategy to handle the data sparsity problem.

Second, the supervised classification methods supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) obtain better performance on most tasks across two datasets. This is reasonable since a significant number of training instances are provided for them. However, the dataless counterparts SSCF and STM still manage to achieve better performance than these two supervised methods in some tasks. Specifically, both SSCF and STM deliver significantly better performance than supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) on 4 and 5 tasks, respectively on the Web Snippet (WS) dataset. Recall that the subtopics (search phrases) covered by the training set and test set of the Web Snippet (WS) dataset have no overlap (Phan et al., 2008). We believe that this observation can explain the inferior performance of supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) on the Web Snippet (WS) dataset. Because we randomly sample the training and test documents on the Baidu question answer (BaiduQA) dataset, hence supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) obtain the best performance instead. It is surprising that SSCF still achieves better performance than supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) on both datasets, when all categories are considered for classification (*i.e.*, All-8 and All-35). On All-35, we can see that both supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) achieve relatively poorer performance. Note that some categories in Baidu question answer (BaiduQA) are very similar to each other, making the multi-class classification very difficult. For example, there are several programming language related categories in Baidu question answer (BaiduQA): Visual C++, Visual Basic, Java, and Assembler Language. These 4 categories share a lot of words in common. Based on word occurrences alone, supervised latent dirichlet allocation (sLDA) and support vector machine (SVM) cannot perform well on this difficult task. The superiority of SSCF on All-8 and All-35 suggests that deriving word-level discriminative signals based on the seed words of each category is beneficial to difficult classification tasks such as All-8 and All-35. This observation shed light on many related tasks that require a focused retrieval and analysis on some particular aspects from a specific domain corpus (Wang et al., 2016).

Third, we observe that STM experiences a significant performance degradation on the Baidu question answer (BaiduQA) dataset. The word relevance regarding each category in STM is estimated based on the word co-occurrence statistics between a word and a seed word. Since the documents in Baidu question answer (BaiduQA) are much shorter, word relevance estimated by word co-occurrences may not be accurate. In addition, STM infers the hidden topics directly from the short documents themselves. Based on these issues, it is reasonable that STM obtains much poorer classification performance on the Baidu question answer (BaiduQA) dataset. On the other hand, SSCF consistently achieves much better performance than STM on Baidu question answer (BaiduQA).

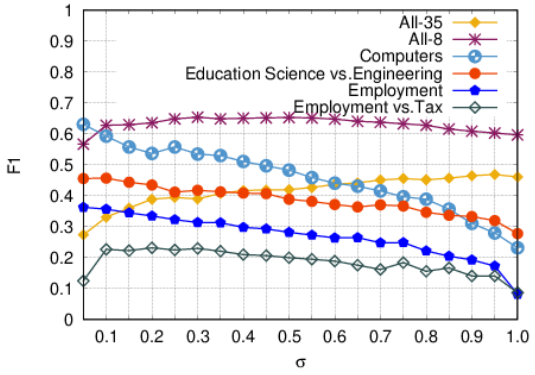
5.3 Analysis of SSCF

We now study the impact of σ, T, and ϵ values, and number of seed words, on these 6 tasks: All-8, All-35, Computers, Education Science *vs.* Engineering, Employment, and Employment *vs.* Tax[7]. While the first two tasks are for classification without filtering, the rest are for classification with filtering. When studying a particular parameter, we fix the other parameters to the values used in Section 5.2.

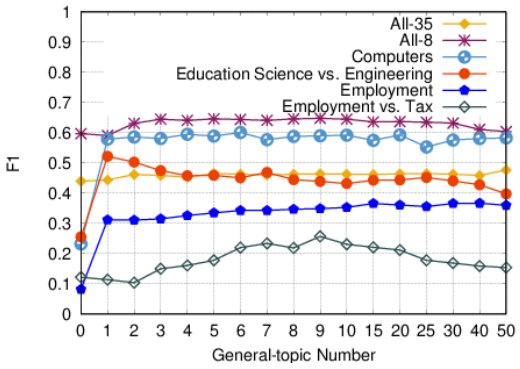
Impact of σ value. Figure 5(a) plots the performance by varying σ value in [0, 1] with a step of 0.05. For All-8 and All-35, we can observe that better performance is achieved by SSCF with larger σ values. While the optimal range of σ values is [0.25, 0.8] on All-8, the optimal range is [0.8, 0.95] on All-35. As discussed above in Section 5.2, many categories in Baidu question answer (BaiduQA) are highly similar to each other. Therefore, a word relevant to several similar categories will have a higher τ_w (ref. Equation 8). In this sense, we need a larger σ to enhance the prior bias in this case. When $\sigma = 1.0$, SSCF has no filtering function. However, performance deterioration is observed when σ is set to 1 on the two tasks. This suggests that general-topics can absorb the noise or background semantics underlying the documents. This finding is consistent with STM (Li et al., 2016b). Here, we choose $\sigma = 0.7/0.9$ for all classification without filtering tasks on Web Snippet (WS) and Baidu question answer (BaiduQA) datasets respectively.

As for the classification with filtering tasks, the optimal performance is delivered by a much smaller σ. We observe that the optimal σ value increases as more categories are specified. That is, the choice of σ is positively proportional to the number of specified categories. Accordingly, we choose to use $\sigma = 0.1$ for all classification with filtering tasks.

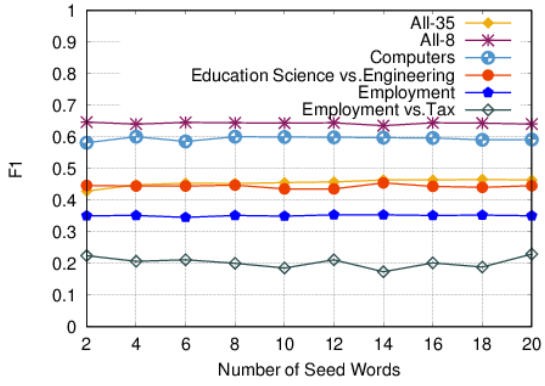
7 Similar performance patterns are observed for other tasks.



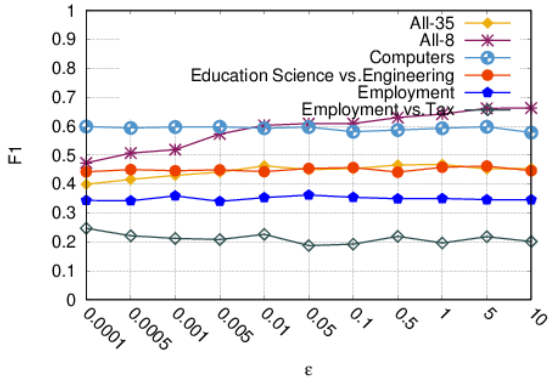
(a) σ values



(b) T values



(c) Number of seed words



(d) ϵ values

Figure 5. Performance of SSCF with different parameter settings.

Impact of T value. The T value indicates number of general topics underlying the short documents. When T = 0, SSCF loses the ability to filter out irrelevant documents. In contrast, more semantics could be modeled by using a larger T value. Figure 5(b) plots the performance of SSCF with different T values. We observe that little performance fluctuation is experienced by SSCF in a large range of T values (*i.e.*, [5, 40]) on almost all tasks, except for the task of Employment-Tax. Note that All-8 and All-35 are tasks of classification without filtering. It makes sense that a larger T would hurt the classification accuracy significantly. However, we observe that T = 0 delivers much worse classification accuracy instead. This observation is consistent with the study of parameter σ. Since more categories are covered by All-8 and All-35, there would involve noisier or background semantics. We indeed observe that a better accuracy is achieved when T is relatively larger for these two tasks, especially for All-8. We set T to 1 and 10 for classification without filtering and with filtering respectively, except for All-8 and All-35 (T = 10 instead).

Impact of ϵ value. Recall that parameter ϵ works as a weak smoothing prior in the prior category-topic distribution $\delta\vec{\eta}_w + \vec{\epsilon}$ of the pseudo-document d_w. Given a fixed $\vec{\eta}_w$, the ratio of δ to ϵ controls the sparsity of dominating category-topics in a pseudo-document. Here, Figure 5(d) plots the performance pattern of SSCF by varying ϵ value in the range of [0.0001, 10]. We can observe that little performance fluctuation is experienced by SSCF in a large range of ϵ values (*i.e.*, [0.005, 5]), except for All-8 on the Web Snippet (WS) dataset. As to All-8, a much larger ϵ could lead to better classification performance. Generally, SSCF is not sensitive to the choice of ϵ value in most tasks across the two datasets. Accordingly, we choose to use ϵ=0.01 in our experiments.

Impact of seed words. The information covered by seed words is critical for the classification performance of SSCF. Here, we investigate performance of SSCF by varying the number of seed words available. For each category, we vary the number of the seed words from 2 to 20. For example, when only 2 seed words are allowed, we randomly sample these two words from total 20 seed words of the category. For each number setting (excluding 20), we report the average classification score of SSCF over 5 runs. Note that there is no randomness when all 20 seed words are used. In this case, we choose to report the average results over 5 runs instead. Figure 5(c) plots the performance of SSCF with different numbers of seed words per category. It is surprising that SSCF delivers little performance fluctuation when using various numbers of seed words per category. Especially, when only two seed words are available for each category, SSCF can still achieve almost the same classification accuracy as using 20 seed words. We further calculate the proportions of the documents that contain any selected seed words under the categories of interest. These numbers are 17.5%, 20.1%, 22.0%, 18.3%, and 15.3% on average by using 2 randomly selected seed words per category. We can see that the document coverage by using merely 2 seed words per category is just small. That is, applying a simple seed word matching could be an inferior strategy for both dataless classification with filtering and classification without filtering. In contrast, the proposed SSCF can deliver robust classification accuracy when seed words are scarce.

6 Conclusion

In this paper, we propose a novel seed-guided topic model for dataless short text classification and filtering, named SSCF. A novel word relevance estimation method is proposed to grasp the supervision of the provided seed words for better category-topics and general-topics learning. The experimental results show that SSCF outperforms the existing state-of-the-art dataless alternatives in either classification with filtering task or classification without filtering task. Also, we found that SSCF even surpasses the supervised classifiers in many settings. Given increasing demand in focused retrieval and fine-grained analysis of a particular aspect or facet from diverse short texts, extracting relevant information by utilizing just few seed words is a cost-effective mechanism. We believe that the proposed SSCF can provide useful implications for many related studies. In future, we plan to apply SSCF in many other interesting scenarios with short texts of more diverse topics, such as large-scale tweet classification[8] (Zubiaga and Ji, 2013) and crisis analysis[9] (Olteanu et al., 2014; Olteanu et al., 2015b; Olteanu et al., 2015a; Olteanu et al., 2016). In addition, we intend to devise an automatic procedure to derive the optimal σ value in an iterative manner. Another possible extension is to use the word embedding techniques to exploit the general word semantic relatedness from a large external corpus.

Acknowledgment: This research was supported by the National Natural Science Foundation of China (No. 61872278) and the Natural Science Foundation of Hubei Province (No. 2017CFB502).

References

Abdi, L., & Hashemi, S. (2016). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering, 28*(1), 238–251.

Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. *Neural Inforamtion Processing Systems Conference*, 121-128.

Cao, X., Cong, G., Cui, B., & Jensen, C. S. (2010). A generalized framework of exploring category information for question retrieval in community question answer archives. *Proceedings of the 19th International Conference on World Wide Web*, 201–210. doi: 10.1145/1772690.1772712

Chang, M., Ratinov, L., Roth, D., & Srikumar, V. (2008). Importance of semantic representation: Dataless classification. *Proceedings of the the 23rd AAAI Conference on Artificial Intelligence,* 830–835.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12), 2928–2941.

Chen, M., Jin, X., & Shen, D. (2011). Short text classification improved by learning multi-granularity topics. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 1776-1781.

Chen, X., Xia, Y., Jin, P., & Carroll, J. A. (2015). Dataless text classification with descriptive LDA. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2224–2231.

Downey, D., & Etzioni, O. (2008). Look ma, no hands: Analyzing the monotonic feature abstraction for text classification. *Neural Information Processing Systems*, 393–400.

Druck, G., Mann, G. S., & McCallum, A. (2008). Learning from labeled features using generalized expectation criteria. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 595–602. doi: 10.1145/1390334.1390436

8 The dataset is available at http://www.zubiaga.org/datasets/odpt-weets/.

9 The datasets are available at http://crisislex.org/data-collections.html.

Efron, M., Lin, J. J., He, J., & de Vries, A. P. (2014). Temporal feedback for tweet search with non-parametric density estimation. *International Conference on Research and Development in Information Retrieval*, 33–42.

Ghosh, S., & Desarkar, M. S. (2018). Class specific TF-IDF boosting for short-text classification: Application to short-texts generated during disasters. *Companion of the The Web Conference 2018 on The Web Conference 2018*, 1629-1637. doi: 10.1145/3184558.3191621

Gliozzo, A., Strapparava, C., & Dagan, I. (2009). Improving text categorization bootstrapping via unsupervised learning. *ACM Transactions on Speech and Language Processing*, *6*(1):1, 1:24.

Hingmire, S., & Chakraborti, S. (2014). Topic labeled text classification: A weakly supervised approach. *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 385–394. doi: 10.1145/2600428.2609565

Hingmire, S., Chougule, S., Palshikar, G. K., & Chakraborti, S. (2013). Document classification by topic labeling. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 877–880. doi: 10.1145/2484028.2484140

Lin, T., Tian, W., Mei, Q., & Cheng, H. (2014). The dual-sparse topic model: Mining focused topics and focused terms in short text. *Proceedings of the 23rd International Conference on World Wide Web,* 539–550. doi:10.1145/2566486.2567980.

Liu, B., Li, X., Lee, W. S., & Yu, P. S. (2004). Text classification by labeling words. *Proceedings of the National Conference on Artificial Intelligence*, 425–430.

Li, C., Chen, S., Xing, J., Sun, A., & Ma, Z. (2019a). Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems*. *37*(1), 9:1-37 . doi:10.1145/3238250

Li, C., Ouyang, J., & Li, X. (2019b). Classifying extremely short texts by exploiting semantic centroids in word mover's distance space. *The World Wide Web Conference*, 939-949 doi: 10.1145/3308558.3313397

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016a). Topic modeling for short texts with auxiliary word embeddings. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*,165-174. doi:10.1145/2911451.2911499

Li, C., Xing, J., Sun, A., & Ma, Z. (2016b). Effective document labeling with very few seed words: A topic model approach. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 85-94. doi:10.1145/2983323.2983721

Li, X., Li, C., Chi, J., Ouyang, J., & Li, C. (2018). Dataless text classification: A topic modeling approach with document manifold. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 973-982. doi: 10.1145/3269206.3271671

Long, G., Chen, L., Zhu, X., & Zhang, C. (2012). Tcsst: Transfer classification of short & sparse text using external data. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 764–772. doi: 10.1145/2396761.2396859

Mikolov, T., Chen, K., Corrada, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 1-12.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011). Searching microblogs: coping with sparsity and document quality. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 183–188. doi:10.1145/2063576.2063607

Nishida, K., Hoshide, T., & Fujimura, K. (2012). Improving tweet stream classification by detecting changes in word probability. *Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval,* 971–980. doi:10.1145/2348283.2348412

Olteanu, A., Castillo, C., Diakopoulos, N., & Aberer, K. (2015a). Comparing events coverage in online news and social media: The case of climate change. *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, 288–297.

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, 376-385.*

Olteanu, A., Vieweg, S., & Castillo, C. (2015b). What to expect when the unexpected happens: Social media communications across crises. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 994–1009. doi:10.1145/2675133.2675242

Olteanu, A., Weber, I., & Gatica-Perez, D. (2016, March). *Characterizing the demographics behind the #blacklivesmatter movement*. 2016 AAAI Spring Symposia, Symposium conducted at the meeting of Stanford University, Palo Alto, California, USA.

Omari, A., Carmel, D., Rokhlenko, O., & Szpektor, I. (2016). Novelty based ranking of human answers for community questions. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval,* 215–224. doi:10.1145/2911451.2911506

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. *Proceedings of the 17th International Conference on World Wide Web,* 91–100. doi:10.1145/1367497.1367510

Ritter, A., Wright, E., Casey, W., &Mitchell, T. M. (2015). Weakly supervised extraction of computer security events from twitter. *Proceedings of the 24th International Conference on World Wide Web,* 896–905. doi:10.1145/2736277.2741083

Scaiella, U., Ferragina, P., Marino, A., & Ciaramita, M. (2012). Topical clustering of search results. *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012*, 223–232.

Shalaby, W., & Zadrozny, W. (2019). Learning concept embeddings for dataless classification via efficient bag-of-concepts densification. *Knowledge and Information Systems, 61*(2), 1047–1070.

Song, Y., Pan, S., Liu, S., Zhou, M. X., &Qian, W. (2009). Topic and keyword re-ranking for lda-based topic modeling. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 1757–1760. doi:10.1145/1645953.1646223

Song, Y., & Roth, D. (2014). On dataless hierarchical text classification. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1579–1585.

Stein, B., Gollub, T., & Hoppe, D. (2012). Search result presentation based on faceted clustering. *Proceedings of the 21st ACM*

International Conference on Information and Knowledge Management, 1940–1944. doi: 10.1145/2396761.2398548

Sun, A. (2012). Short text classification using very few words. *Proceedings of the 35th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1145-1146. doi:10.1145/2348283.2348511

Thomas, G., & Mark, S. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101(SUPPL. 1),* 5228-5235.

Wang, F., Wang, Z., Li, Z., & Wen, J.-R. (2014). Concept-based short text classification and ranking. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1069-1078. doi: 10.1145/2661829.2662067.

Wang, J., Wang, Z., Zhang, D., & Yan, J. (2017). Combining knowledge with deep convolutional neural networks for short text classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017,* 2915–2921. doi: 10.24963/ijcai.2017/406

Wang, S., Chen, Z., Fei, G., Liu, B., & Emery, S. (2016). Targeted topic modeling for focused analysis. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1235–1244. doi:10.1145/2939672.2939743

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. *Proceedings of the 22nd international conference on World Wide Web*, 1445-1456. doi:10.1145/2488388.2488514

Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., & King, I. (2018). Topic memory networks for short text classification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 3120–3131. doi: 10.18653/v1/D18-1351

Zhang, S., Jin, X., Shen, D., Cao, B., Ding, X., & Zhang, X. (2013). Short text classification by detecting information path. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 727-732. doi: 10.1145/2505515.2505638.

Zhang, Y., Szabo, C., Sheng, Q. Z., & Fang, X. S. (2018). SNAF: Observation filtering and location inference for event monitoring on twitter. *World Wide Web (Bussum), 21*(2), 311–343.

Zubiaga, A., & Ji, H. (2013). Harnessing web page directories for large-scale classification of tweets. *Proceedings of the 22nd international conference on World Wide Web companion,* 225–226.

Zuo, Y., Zhao, J., & Xu, K. (2016). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2), 379–398.

Appendix

A. Seed Words for Evaluation

We manually selected 20 semantically relevant seed words for each category based on the learnt topics from word network topic model (WNTM). The seed words for the Web Snippet (WS) dataset are listed as below.

Table 6
Seed Words for the Web Snippet (WS) Dataset. P-Train/P-Test Refers to the Percentage of the Documents That Contain Any Seed Word Under Each Category for Training/Testing Respectively.

Category	Seed Words	P-Train (%)	P-Test (%)
Business	business, market, trade, economic, marketing, finance, management, tax, investment, international, income, export, company, import, jobs, products, sales, markets, tax, budget	16.1	4.5
Computers	computer, web, software, programming, internet, systems, computing, digital, linux, network, java, apple, hardware, server, computers, cpu, processor, games, chip, security	16.4	8.9
Culture Arts Ent	movie, music, culture, art, movies, arts, american, history, video, imdb, books, oscar, band, reviews, comedy, romantic, classical, literature, academy, pop	14.4	5.9
Education Science	research, education, science, school, theoretical, physics, journal, university, department, natural, graduate, biology, mathematical, scientific, library, theorem, mathematics, technology, faculty, edu	15.7	11.4
Engineering	engine, diesel, fuel, turbine, cylinder, jet, automobile, automation, piston, engines, gas, technology, auto, industrial, power, products, motor, manufacturer, engineering, gasoline	28.2	4.6
Health	health, medical, cancer, disease, healthy, nutrition, diet, hospital, physical, therapy, diagnosis, treatment, care, heart, fitness, medicine, food, symptoms, hepatitis, clinical	21.5	6.4
Politics Society	political, democracy, party, government, republic, military, democratic, politics, united, president, parliamentary, freedom, national, congress, social, communist, revolution, parties, civil, presidential	18.7	4.2
Sports	sports, football, games, game, tennis, match, league, espn, tickets, team, players, hockey, sport, olympic, basketball, golf, cup, volleyball, club, scores	19.6	10.4

And the seed word from Baidu question answer (BaiduQA) dataset are listed as below:

Table 7
Seed Words for the Baidu Question Answer (BaiduQA) Dataset. P-Train/P-Test Refers to the Percentage of the Documents That Contain Any Seed Word Under Each Category for Training/Testing Respectively.

Category	Seed Words	P-Train (%)	P-Test (%)
JAVA	java, JAVA, Java, 程序(program), 文件(file), 代码(code), jsp, javascript, 方法(method), 手机(cellphone), 编程(programming), 数据库(database)), 编写(write), JSP, 软件(software),页面(page), 对象(object), 学习(learn), 项目(project), 运行(run)	23.2	23.5
VB	VB, vb, 程序(program), 数据(data), 代码(code), 控件(control), 文件(file), 编程(programming), 窗体(form), 文本(text), 数据库(database), VBA, 字符(char), 函数(function), 错误(error), 编写(write), 判断(judge), 运行(run), 网页(webpage), 窗口(window)	26.5	25.8

Continued Table 7

Seed Words for the Baidu Question Answer (BaiduQA) Dataset. P-Train/P-Test Refers to the Percentage of the Documents That Contain Any Seed Word Under Each Category for Training/Testing Respectively.

Category	Seed Words	P-Train (%)	P-Test (%)
VC++	VC, 程序(program), vc, 对话框(dialog box), 函数(function), 文件(file), MFC, 编程(programming), 运行(run), 窗口(window), 代码(code), 编译(compiling), 调用(call),安装(install), 数据(data), VC6, mfc, 字符(char), 成员(member), 指针(point)	26.5	25.8
百度空间 (baidu space)	空间(space), 百度(Baidu), 图片(picture), 音乐(music), 背景(background), QQ, 文章(article), 博客(blog), 上传(upload), 好友(friend), qq, 模板(template), 相册(photoalbum), 照片(photo), 视频(video), 开心农场(name of a game), 贴吧(post bar), 日志(journal), 留言(message), 播放(play)	43.3	44.5
财务税务 (tax)	会计(accountant), 发票(invoice), 企业(enterprise), 公司(company), 现金(cash), 增值税(added-value tax), 资产负债表(balance sheet), 流量表(flowmeter), 财务(finance), 科目(subject), 会计分录(accounting entry), 财务报表(financial statement), 费用(charge)),纳税人(taxpayer), 利润表(profit statement), 成本(cost), 固定资产(fixed assert), 所得税(income tax), 收入(income), 利润(profit)	17.7	17.8
产业信息 (bizinfo)	销售(sell), 行业(industry), 公司(company), 产品(product), 价格(price), 发展(development), 前景(prospect), 市场(market), 品牌(brand), 产业(industrial), 厂家(manufacturer), 设备(equipment), 生产(production), 企业(enterprise), 汽车(automobile), 信息(information), 酒店(hotel), 服装(clothing), 旅游(travel), 广告(advertising)	15.7	15.0
法律 (law)	起诉(prosecute), 赔偿(compensation), 合同(contract), 公司(company), 户口(registered permanent residence), 劳动合同(labor contract), 律师(lawyer), 法院(court), 财产(property), 离婚(divorce), 工资(salary), 责任(responsibility), 规定(stipulate), 劳动(labor), 劳动法(labor law), 咨询(consult), 工伤(occupational injury), 偿(compensation), 纠纷(dispute), 法律(law)	14.1	13.4
肝胆外科 (hepatological surgery department)	乙肝(hepatitis B), 治疗(cure), 胆囊(cholecyst), 肝功能(liver function), 胆结石(gallstone), 转氨酶(transaminase), 体检(medical examination), 检查(examination), 肝硬化(liver cirrhosis), 胆红素(bilirubin), 结石(lithiasis), 手术(operation), 胆囊炎(cholecystitis), 医院(hospital), 囊肿(cyst), 肝脏(liver), 大三(liver disease), 阳性(positive reaction), 患者(patient), 医生(doctor)	28.9	28.7
工程技术科学 (engineering technology science)	电机(electrical machinery), 电压(voltage), 原理(theory), 材料(material), 设计(design), 电流(electricity), 变压器(transformer), 电路(circuit), 建筑(building), 型号(type), 卫星(statellite), 设备(machine), 控制(control), 信号(sign), 功率(power), 电容(capaciance), 压力(pressure), 技术(technology), 检测(detection), 混凝土(concrete)	9.4	9.6
购房置业 (real estate)	房子(house), 贷款(loans), 房屋(building), 二手房(second-hand house), 买房(buy a house), 房产(house property), 房产证(property ownership certificate), 公积金(accumulation fund), 住房(housing), 房价(housing price), 过户(transfer ownership), 拆迁(remove), 购房(buy a house), 合同(contract), 租房(renting), 办理(transaction), 中介(intermediary agent), 购买(buy), 套房(suite), 按揭(loans)	22.8	22.9
管理学 (management)	管理(management), 人力资源(human resource), 管理学(management), 事业(career), 领导(leader), 工作(job), 事业单位(public institution), 公司(company), 专业(major), 编制(formation), 组织(organization), 大学(university), 企业管理(business management), 理论(theory), 行政管理(administrative management), 公共管理(public management), 社会(society), 论文(paper), 发展(development), 干部(cadre)	19.3	18.8
韩语 (korean)	翻译(translate), 韩文(korean), 翻译成(translate), 韩国(korea), 韩国语(korean), 学习(learn), 中文(Chinese), 发音(pronounce), 词语(word), 培训(train), 考试(exam), 韩语歌(korean song), 首尔(seoul), 标准(standard), 专业(major), 单词(word), 英语(english), 翻译器(translator), 韩国人(korean), 语法(grammar)	22.1	23.3
汇编语言 (assembly language)	汇编(assembly), 汇编语言(assembly language), 单片机(singlechip), 程序(program), 语言(language), 显示(display), 汇编程序(assembly program), 指令(order), 编程(programming), 网页(webpage), asp, ASP, 51, 文件(file), 地址(address), 字符(char), 输出(output), 编写(writing), 设计(design), 代码(code)	24.1	23.7

Continued Table 7

Seed Words for the Baidu Question Answer (BaiduQA) Dataset. P-Train/P-Test Refers to the Percentage of the Documents That Contain Any Seed Word Under Each Category for Training/Testing Respectively.

Category	Seed Words	P-Train (%)	P-Test (%)
经济研究 (economic research)	经济(economics), 中国(china), 市场(market), 经济学(economics), 影响(influence)), 货币(currency), GDP, 政策(policy), 金融危机(financial crisis), 价格(price), 利率(interest rate), 通货膨胀(currency inflation), 企业(enterprise), 金融(financial), 需求(demand), 商品(commodity), 公司(company), 人民币(China Yuan CNY), 政府(government), 消费(consumption)	15.8	15.9
贸易 (trade)	快递(expressage), 外贸(foreign trade), 价格(price), 生产(production), 公司(company), 有限公司(limited company), 出口(export), 贸易(trade), 产品(product), 包裹(package), 批发(wholesale), 淘宝(taobao), 国际(international), 生产厂家(manufacturer), 销售(sell), 国际贸易(international trade), 报关(customs clearance), 海关(china customs), 经营(manage), 物流(physical distribution)	15.9	15.8
欧美流行乐 (western popular music)	歌词(lyric), 英文歌(english song), 音乐(music), 歌曲(song), 歌名(song name), 英文歌曲(english song), 英文(English), 女声(female voice), 伴奏(accompany), 歌手(singer), 欧美(western), 插曲(episode), MV, 说唱(rap), 铃声(ringtones), 女生(female), 英语(English), 乐队(band), 好听(melody), DJ	20.0	19.4
皮肤科 (dermatology department)	治疗(cure), 白癜风(leucoderma), 过敏(allergy), 皮肤(skin), 医院(hospital), 疙瘩(pimple), 湿疹(eczema), 身上(on one's body), 脸上(on one's face), 疤痕(scar), 牛皮癣(psoriasis), 皮肤病(skin disease), 尖锐湿疣(verruca acuminata), 脱皮(decrustation), 症状(symptom), 疱疹(herpes), 荨麻疹(urticaria), 治愈(cure), 狐臭(bromhidrosis), 医生(doctor)	21.9	22.6
其他编程语言 (other programming languages)	程序(program), 代码(code), 语言(language), 文件(file), 数据库(database), 数据(data), 函数(function), 错误(error), matlab, VB, php, asp, ASP, 单片机(singlechip), 运行(run), vb, delphi, PHP, 系统(system), 编程(programming)	15.3	15.2
其他社会话题 (othersocialtopics)	中国(China), 身份证(identity card), 日本(Japan), 世博会(world expo), 户口(registered permanent residence), 文化(culture), 国家(country), 档案(record), 公司(company), 世界(world), 征兵(conscription), 城市(city), 活动(activaty), 毕业生(graduate), 计划生育(family planning), 改革(reform), 社会(society), 简介(abstract), 政府(government), 大学生(college student)	10.4	20.5
企业管理 (business management)	企业(enterprise), 管理(management), 公司(company), 软件(software), 企业管理(business management), 员工(staff), 营销(marketing), 采购(purchase), 管理制度(management system), 销售(sell), 仓库(warehouse), 酒店(hotel), 制度(institution), 认证(identification), 生产(production), 物流(physical distribution), 集团(group), 质量管理(quality control), 质量(quality), 人力资源(human resource)	17.5	17.8
求职就业 (employment)	工作(job), 待遇(treatment), 公司(company), 毕业(graduate), 面试(interview), 上班(work), 就业(employment), 员工(staff), 工资(salary), 辞职(resign), 专业(career), 简历(resume), 兼职(part-time job)), 招聘(recruitment), 毕业生(graduate), 职业(occupation), 单位(company), 老板(boss), 档案(record), 户口(registered permanent residence)	21.7	21.5
日韩流行乐 (k-pop)	歌词(lyric), 歌曲(song), MP3, 音乐(music), 下载(download), mp3, 罗马(rome), 韩文歌(korean song), 插曲(episode), 韩国(korea), 伴奏(accompany), 韩文(korean), 铃声(ringtones), 时代(epoch), 日文歌(japanese song), MV, 歌名(song name), 日本(japan), 日文(japanese), 日语(japanese)	23.7	23.4
日韩明星 (japanese and south korean star)	东方神起(name of a pop group), 韩国(korean), 演唱会(vocal concert), SJ, 豆花(name of a fans), super, junior, 少女(maiden), 时代(epoch), sj, 韩庚(name of a singer), 专辑(album), 节目(program), superjunior, 明星(star), 综艺节目(variety show), 跳舞(dance), 滨崎步(name of a singer), shinee, 张根锡(name of a singer)	14.7	15.0
软件共享 (software share)	下载(download), 注册(register), 激活(activate), 序列(sequence), 授权(authorization), 软件(software), 在线(online), 电子书(e-book), 地址(address), 申请(apply), 手机(cellphone), 免费(free), 视频(video), 破解(crack), 屏幕(screen), 格式(format), 录像(video), 电脑(computer), Adobe, QQ	35.8	37.2

Continued Table 7

Seed Words for the Baidu Question Answer (BaiduQA) Dataset. P-Train/P-Test Refers to the Percentage of the Documents That Contain Any Seed Word Under Each Category for Training/Testing Respectively.

Category	Seed Words	P-Train (%)	P-Test (%)
散文 (prose)	描写(describe), 作文(composition), 散文(prose), 鲁迅(name of an author), 修辞(rhetoric), 作者(author), 作家(writer), 美文(beautiful article), 读后感(reaction), 文学(literature), 表达(expression), 作品(production), 读书(reading), 写作(writing), 比喻(metaphor), 朱自清(name of an author), 感情(emotion), 冰心(name of an author), 小说(novel), 古文(ancient chinese prose)	11.5	11.2
诗歌 (poem)	诗歌(poetry), 诗句(poem), 古诗(ancient poetry), 描写(describe), 诗词(poem), 赏析(appreciate), 诗人(poet), 歌词(lyric), 作者(author), 句子(sentence), 朗诵(recite), 爱情诗(love poem), 爱情(love), 古诗词(ancient poetry), 写诗 (poetize), 思念(miss), 李白(name of a poet), 情诗(love poem), 赞美(praise), 全诗(whole poem)	14.9	13.7
实况足球 (pro evolution soccer)	补丁(patch), 转会(transfer), 实况(scene), 实况足球(pro evolution soccer), 球员(player), 足球(soccer), 联赛(league), 解说(narrate), 实况足球10(pro evolution soccer 10), 存档(on file), 球衣(poloshirt), 传奇(legend), 卡卡(ricardo), 中超(chinesetootballassociationsuper league), 版本(version), PES2009, pes2010, pes6, 球队(team), 游戏(game)	36.9	36.8
视频共享 (video share)	视频(video), 下载(download), 地址(address), 在线(online), 观看(watch), 土豆(name of a video website), 网站(website), 全集(complete works), 免费(free), 播放(play), 音乐(music), 上传(upload), 电脑(computer), 网址(url), 迅雷(name of a video player), 格式(format), 网上(online), 清晰(sharpness), 电影(movie), MV	38.0	36.3
数据库 (database)	SQL, sql, 数据库(database), 查询(select), 数据(data), 语句(statement), server, oracle, 字段(field), 安装(install), mysql, 记录(record), Server, 创建(create), 服务器(server), Oracle, 删除(delete), 文件(file), access, 代码(code)	31.8	32.3
戏剧 (theater)	京剧(peking opera), 剧本(scenario), 曲子(melody), 话剧(drama), 戏剧(theater), 戏曲(traditional chinese opera), 曲谱(music score of chinese operas), 越剧(shaoxing opera), 歌剧(opera), 表演(perform), 相声(cross talk), 脸谱(facial makeup), 豫剧(yu opera), 黄梅戏(huangmei opera), 段子(joke), 二人转(the couple dance opera), 音乐剧(musical), 演员(actor), 台词(actor's lines), 评书(storytelling)	18.3	16.9
小说 (novel)	小说(novel), TXT, 全集(complete work), txt, 穿越(pass through), 完结(finish), 耽美(boy's love), 言情小说(romantic fiction), 暮光之城(the twilight saga), 同人(homosexual, 主角(leading actor), 章节(chapter), 风流(romantic), 作者(author), 都市(metropolis), 古代(ancient times), 阅读(read), 穿越小说(time travel fiction), 网游(webgame), 玄幻小说(fantasy novel)	21.7	21.5
语言学 (linguistics)	句子(sentence), 翻译(translate), 解释(explain), 全文(full text), 读音(pronunciation), 成语(idiom), 词语(word), 阅读(read), 语文(chinese language), 古文(ancient chinese prose), 汉语(chinese), 汉字(chinese character), 英文(english), 文字(character), 故事(story), 繁体字(complex chinese character), 文章(article), 修辞(thetoric), 英语(english), 歇后语(two-part allegorical saying)	19.6	19.3
院校信息 (colleges and universities information)	大学(university), 专业(major), 学院(college), 学校(school), 设计(design), 职业(career), 校区(campus), 科技(science), 技术学院(technical college), 录取(admit), 研究生(master), 工程(engineering), 师范大学(normal university, 学生(student), 招生(recruit student), 就业(employment), 理工大学(university of science and engineering), 分数线(cutoff score), 本科(bachelor), 毕业(graduate)	24.6	24.8
职业培训 (professional training)	培训(train), 专业(major), 学校(school), 学习(learn), 培训学校(training school)), 考试(exam), 就业(employment), 培训班(training course), 电脑(computer), 工程师(engineer), 计算机(computer), 软件(software), 课程(course), 中专(technical secondary school), 技术(technical), 职业(career), 电工(electrician), 网络(network), 会计(accountant), 大学(university)	21.5	21.0
中考 (senior high school entrance examination)	中考(senior high school entrance examination), 作文(composition), 年级(grade), 中学(middle school), 语文(chinese language), 数学(math), 答案(answer), 初中(middle school), 考试(exam), 初三(third year of middle school), 上册(volume one), 成绩(performance), 复习(review), 高中(high school), 学生(student), 物理(physics), 附中(affiliated middle school), 学校(school), 试题(test), 录取(admit)	25.8	26.2