Exploring Topic Discriminating Power of Words in Latent Dirichlet Allocation

Kai Yang, Yi Cai,[*] Zhenhong Chen
School Of Software Engineering, South China University of Technology, China

Ho-fung Leung
Department of Computer Science and Engineering,
The Chinese University of Hong Kong,
Hong Kong

Raymond LAU
College of Business,
City University of Hong Kong,
Hong Kong

Abstract

Latent Dirichlet Allocation (LDA) and its variants have been widely used to discover latent topics in textual documents. However, some of topics generated by LDA may be noisy with irrelevant words scattering across these topics. We name this kind of words as topic-indiscriminate words, which tend to make topics more ambiguous and less interpretable by humans. In our work, we propose a new topic model named TWLDA, which assigns low weights to words with low topic discriminating power (ability). Our experimental results show that the proposed approach, which effectively reduces the number of topic-indiscriminate words in discovered topics, improves the effectiveness of LDA.

1 Introduction

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its variants are generative statistical topic models providing a powerful framework for finding topics in text documents. In generative process, each document is a mixture of several topics, and the generation of each word belongs to one of the document's topics (Heinrich, 2009).

Mimno et al. have found that LDA often produces topics that are not interpretable or meaningful (Mimno et al., 2011). According to our observation, most of topics (especially those considered uninterpretable) contain some words which are common in the corpus. For example, words like 'science', 'academic' or 'abstract' in a corpus about scientific publications will appear in most of topics. To explain this kind of words more clearly, we prepare another example showed in Table 1(a). The table shows the top 5 words for 5 topics generated by standard LDA from a corpus of reviews about smart phones. Word 'phone' can be easily recognized as a common word in the corpus about phones, and we can find that all the topics contain this word. For words which are likely to scatter across many topics are difficult to discriminate different topics, we denote this kind of words, such as 'phone', as **topic-indiscriminate words**. We use the term **topic discriminating power** to denote the ability of a word discriminating different topics. Topic-indiscriminate words have low topic discriminating power.

Table 1: An example about a result of standard LDA

(a) Result of standard LDA

Topic	Word
1	sound, headphones, **phone**, bass, card
2	screen, iphone, **phone**, display, ear
3	picture, **phone**, photo, video, gb
4	memory, sd, gb, **phone**, battery
5	android, **phone**, nexus, Samsung, google

(b) Document frequency (DF) of words

Word	DF	Word	DF
phone	2041	screen	1900
memory	1553	picture	1451
sound	1221	sd	928
android	915	iphone	837
headphones	428	card	389

[1] Corresponding author, e-mail: ycai@scut.edu.cn

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2238–2247, Osaka, Japan, December 11-17 2016.

These topic-indiscriminate words tend to bring some irrelevant words into topics, which make these topics less interpretable. We explain the cause of this negative effect using the following example. Subjectively, in Table 1(a), we can see that Topics 1, 2, 3, 4, 5 can be easily interpreted to topics about sound, screens, pictures, memory cards and Android systems respectively. Words such as ‘card’ in Topic 1, ‘ear’ in Topic 2 and ‘battery’ in Topic 4 seem to be irrelevant to other words in their topics, which make these topic hard to be understood. We consider that these words are brought into topics by topic-indiscriminate words. Figure 1 shows the word co-occurrence relationship about words in Topic 1. Each node represents a word, while two words are linked together if they co-occur in the same document. We can find that word ‘card’ only co-occur with ‘phone’ and never co-occur with other words. According to (Heinrich, 2005), if two words co-occur in the same document, these two words are more likely to be assigned at the same topic in LDA. Plausibly, word ‘card’ is assigned to Topic 1 because of the co-occurrence with word ‘phone’. Hence, it is reasonable for us to consider that topic-indiscriminate words will result in worse performance of LDA.

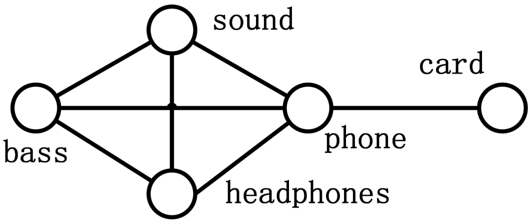


Figure 1: Graph about word co-occurrence

Wilson et al. claim that LDA should take weights of words in documents into consideration (Wilson and Chew, 2010). They consider that words which scatter across more documents are less important and should be given lower weights. We call this kind of word as document-indiscriminate words in our paper. Generally, stop words (e.g. ‘the’, ‘of’ and ‘is’) or common words are document-indiscriminate words, for the reason that these words appear in most of documents. Topic-indiscriminate words are a bit different from document-indiscriminate words, which is illustrated in the following example. Table 1(b) shows us top 10 most frequent words in the corpus of Table 1(a). We can find that words ‘screen’ and ‘memory’, which are kernel words for Topics 2 and 4, have high document frequency. On the other hand, they just appear in Topics 2 and 4 respectively. Therefore, they are document-indiscriminate words instead of topic-indiscriminate words, i.e. words with high topic discriminating power. In Wilson’s approach, word ‘screen’ and ‘memory’ will be assigned lower weights to decrease their rankings in Topics 2 and 4. This will make topics less interpretable, as these words are important for people to understand topics. Hence, Wilson’s approach has low ability to find out topic-indiscriminate words accurately, although it does well in finding document-indiscriminate words.

In this paper, we explore the topic discriminating power of LDA, and propose a new LDA model called Term Weighting LDA (TWLDA), which provides a way to measure this power according to supervised term weighting schemes. With our model, topic-indiscriminate words will be given lower weights and have less negative effect on the results of LDA. The reason why we apply supervised term weighting schemes to measure topic discriminating power is that they have been used to measure the discriminating power of words among categories in text categorization tasks (Lan et al., 2009). Words which concentrate on one topic can better discriminate that topic, and the topic discriminating power of these words are stronger. Hence, topic-indiscriminate words, whose topic discriminating power are weak, will be considered less important and be given lower weights in our proposed model. In summary, we conclude our contributions as follows: (a) We explore the topic discriminating power of words in LDA, and find that these words will make the generated topics less interpretable; (b) To solve the problem caused by topic-indiscriminate words, we propose a new model called TWLDA, which can measure the topic discriminating power of words and assign low weights to topic-indiscriminate words in order to reduce the negative effect caused by these words; (c) We explore our proposed TWLDA with different term weighting schemes, and find that supervised schemes, especially entropy-based supervised schemes, have better performance than others; (d) We also conduct several experiments to demonstrate the effectiveness of TWLDA with different evaluation metrics.

2 Related Work

2.1 Latent Dirichlet Allocation

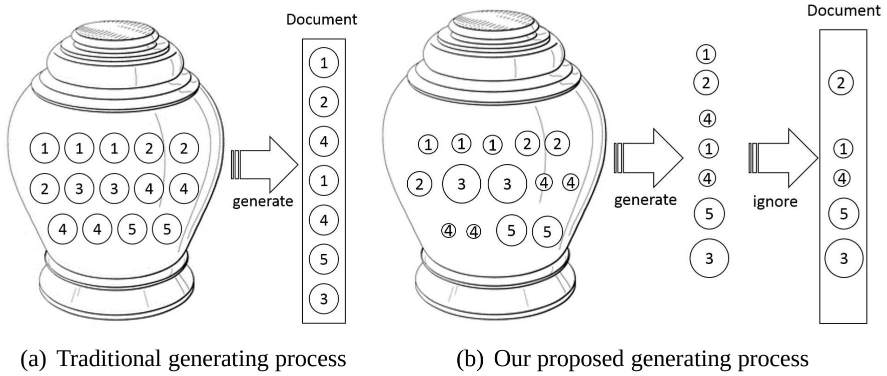
Latent Dirichlet Allocation (Blei et al., 2003) is a generative topic model. It assumes that the words in a document are drawn from a set of latent variables called topics which are distributions over words in the vocabulary.

However, some of the generated topics may mix unrelated or loosely-related words (Mimno et al., 2011). To tackle this problem, some knowledge-based topic models have been proposed in (Andrzejewski et al., 2009; Chen et al., 2013; Chen et al., 2014). These models use expert domain knowledge to guide LDA. For example, DF-LDA (Andrzejewski et al., 2009) takes domain knowledge in the form of must-links and cannot-links given by users. A must-links means that two words should be assigned to the same topic while a cannot-links means that two words should not. Besides, there are several models utilizing seed words provided by users (Burns et al., 2012; Jagarlamudi et al., 2012; Mukherjee and Liu, 2012). In some recent works, for example, GKLDA model (Chen et al., 2013) utilizes the general knowledge such as lexical knowledge to boost the performance.

2.2 Term Weighting Schemes and its Usage in LDA

Term weighting schemes are widely used to measure the importance of words in documents. They can be classified into supervised schemes and unsupervised schemes (Lan et al., 2009). The supervised schemes exploit category information of training documents while unsupervised schemes do not. There are many unsupervised schemes widely used in Information Retrieval (IR) tasks, such as tf, $tf \cdot idf$ (Sparck Jones, 1972) and some variants (Leopold and Kindermann, 2002; Paik, 2013). However, these schemes ignore the categories labels of each document. On the contrast, supervised schemes use the documents labeled with category information. Some supervised schemes are proposed recently, e.g., $iqf \cdot qf \cdot icf$ (Quan et al., 2011), rf (Lan et al., 2009) and some variants (Ko, 2012). Wang et al. propose some entropy-based term weighting schemes such as bdc which are based on the entropy of terms in categories (Wang et al., 2015). Wang et al. declare that bdc outperforms the state-of-the-art schemes, e.g. $tf \cdot idf$, $iqf \cdot qf \cdot icf$ and rf, in text categorization tasks.

Wilson et al. propose a model called WLDA, which applies term weighting schemes to weight terms in LDA. In their model, term weighting schemes are applied to measure the document discriminating power of words. Words which scatter across more documents are given relatively low weights. However, topic-indiscriminate words may not scatter across almost all the documents. Instead, they scatter across most of topics. Hence, the model proposed by Wilson et al. cannot give topic-indiscriminate words relatively low weights. To overcome this problem, we propose a new LDA model, which can give topic-indiscriminate words relatively low weights.



(a) Traditional generating process (b) Our proposed generating process

Figure 2: Word generating process

3 Generative Process Considering Weights of Words

Some of topics generated by LDA may be uninterpretable which contain irrelevant words. According to our observation, these words tend to scatter across many topics. We denote these words as **topic-indiscriminate words** due to the fact that they cannot discriminate different topics. We use the term **topic discriminating power** to denote the ability of words discriminating topics. Topics of LDA will be less interpretable if they mix with these topic-indiscriminate words. Hence, these words have much negative effect on the results of LDA.

To eliminate or alleviate the negative effect caused by topic-indiscriminate words, a possible way is to reduce the number of them occurring in documents. If the number of topic-indiscriminate words occurring in documents is discounted, the negative effect of these words to the results of LDA will also be alleviated (Heinrich, 2005). Inspired by this way, we propose a new generative process, which take weights of words into consideration. If a word gets a lower weight, the number of this word will be discounted more strongly in documents. Therefore, words with lower weights will have less negative effect on the results of LDA.

To explain our generative process, we describe the procedure of generating words from one topic in Figure 2. The urn represents the word distribution of a topic. Each ball has a mark number, which corresponds to a word in the vocabulary. The number of the balls is proportional to the number of words in the topic, while the size of balls represents its weights. In traditional LDA, each ball is considered having the same size (shown in Figure 2 (a)). In our proposed process, the sizes of balls are varied according to their weights (shown in Figure 2 (b)). The process of generating a word from a topic is as follows. Firstly, a ball is selected from the urn with the same process as traditional model. Secondly, we conduct a random choice to decide whether to put this ball into the document or not. Balls with large size are more likely to be put into the document. As the example shown in Figure 2 (b), balls '1' and '4' are smaller and are less likely to be put into the document.

4 Term Weighting LDA

According to the proposed generative process, we propose a new topic model called Term Weighting LDA (TWLDA). Section 3 has shown that words with lower weights generally have weaker negative effect on results of LDA. Since topic-indiscriminate words negatively affect the results of LDA, we expect to find out a way to give these words relatively low weights. In our work, we use supervised term weighting schemes to calculate weights of words. Supervised term weighting schemes are widely applied to measure the the importance of words in different categories in text categorization (Wang et al., 2015). We regard the topics as categories in documents. Topic-indiscriminate words, which scatter across many topics, will be considered unimportant and get relatively low weights by supervised schemes. However, the topics of words are unknown in the beginning. In order to obtain topics of words, an additional step is conducted before we calculate weights of words. In this step, we execute a topic model. This topic model can be standard LDA or other topic models, such as PLSI (Hofmann, 1999) and so on, which can find out the topics of words in documents. In summary, the proposed TWLDA consists of four main processes, which are shown as follows:

- Step 1: $\overrightarrow{\varphi\prime} \longleftarrow TopicModel()$
- Step 2: $\overrightarrow{\sigma} \longleftarrow Calculate(\overrightarrow{\varphi\prime})$
- Step 3: Discounting the number of words
- Step 4: Executing *xLDA* with the discounted values

In Step 1, a topic model is executed. Then a topic-word distribution $\overrightarrow{\varphi\prime}$ is generated by a this topic model.

In Step 2, according to the $\overrightarrow{\varphi\prime}$, we apply a supervised term weighting scheme to calculate weights of words $\overrightarrow{\sigma}$. Since supervised schemes have the ability to measure the topic discriminating power of words, in principle, all the supervised term weighting schemes can be applied here.

Step 3 is to discount the number of words by their weights. The number of words is diminished proportionally according to weights of words. Hence, the total discounted number of words in document m under topic k is calculated as follows:

$$n'^{(k)}_{m} = \sum_{t=1}^{t=V} \sigma_t n_{mkt} \tag{1}$$

where σ_t denotes the weight of word t, which is ranging from 0 to 1. n_{mkt} is the number of word t belonging to topic k in document m. Similarly, the total discounted number of word t under topic k is calculated as follows:

$$n'^{(t)}_{k} = \sum_{m=1}^{m=M} \sigma_t n_{mkt} \tag{2}$$

Step 4 is to execute the standard LDA or its variants, denoted as xLDA, using the discounted values calculated in Equations 1 and 2. Generally, xLDA can be standard LDA or its variants, such as GKLDA (Chen et al., 2013). The main procedures are the same as xLDA. We take standard LDA for example. In Gibbs Sampling process (Chatterji and Pachter, 2004), conditional probability of word t in document m under topic k is calculated using the following formula:

$$p(z_i = k|\overrightarrow{z}_{k,\neg i}, \overrightarrow{w}, \overrightarrow{\alpha}, \overrightarrow{\beta}) = \frac{n'^{(k)}_{m,\neg i} + \alpha_k}{\sum_{k=1}^{k=K}(n'^{(k)}_{m,\neg i} + \alpha_t)} \frac{n'^{(t)}_{k,\neg i} + \beta_k}{\sum_{t=1}^{t=V}(n'^{(t)}_{k,\neg i} + \beta_t)} \tag{3}$$

where $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$ are hyperparameters of the model. Equation 3 is mostly the same as the formula in the Gibbs Sampling process of traditional LDA (Geman and Geman, 1984). The difference is that those counting variables are replaced with the discounted values, such as $n'^{(k)}_{m,\neg i}$ and $n'^{(t)}_{k,\neg i}$, calculated in Step 3. Equation 3 shows that word t will have less probability to be assigned in topic k if the weight of word t is lower. As a result, words with lower weights will get lower ranking in topics. Hence, the negative effect on results of LDA caused by topic-indiscriminate words will be alleviated if their weights are relatively low. By replacing with the discounted values, other variants of LDA can also be executed in Step 4.

5 Experiment

In this section, we conduct experiments to verify the effectiveness of TWLDA. In the first experiment, we apply the following supervised term weighting schemes in TWLDA: $iqf \cdot qf \cdot icf$ and bdc. We also use unsupervised term weighting schemes $tf \cdot idf$ for comparison. Besides, we will compare the performance of TWLDA with standard LDA and WLDA, which is proposed in (Wilson and Chew, 2010). In our second experiments, we test the performance of TWLDA, WLDA and standard LDA if we do not delete stop words in the pre-processing step.

5.1 Datasets and Pre-processing

Datasets: We use two datasets in our experiments. ‘dataset1’ consists of online reviews from Amazon. There are totally 39,554 reviews mixed together. ‘dataset2’ has been used in (Chen et al., 2013), which consists of 8,958 reviews about camera and phone. We obtain it in the website [1].

Pre-processing: Reviews in ‘dataset1’ are preprocessed as follow. Firstly, words are converted into lower case, and the words with upper or lower case are treated as the same words. Secondly, all punctuations in documents are eliminated and only those alphabetic and numeric characters can be retained. Thirdly, we perform stemming and remove the stop words. In this work, we only use nouns,

[1] https://github.com/czyuan/GKLDA

adjective, verb and adverb. Besides, we do not preprocess $dataset2$ for it has been pre-processed in (Chen et al., 2013).

Parameter Setting: The iteration of TWLDA is set to 2500, which consists of 1000 iterations for the topic model in Step 1 and 1500 iterations for xLDA in Step 4 in Section 4. We set the iterations of preceding LDA model to 1000 due to the reason that most of topic models will converge within 1000 iterations in both two datasets which are used in our experiments. For the reason that small changes of α and β will not affect the results much (Jo and Oh, 2011; Titov and McDonald, 2008), we set $\alpha = 1$ and $\beta = 0.1$ as the setting in (Chen et al., 2013). The value of topic number K is fixed to 20.

5.2 Evaluation Metrics

In this section, we use two ways to evaluate the performance of our proposed model, one is quantitative evaluation and the other is qualitative evaluation. In quantitative evaluation, we use the Topic Coherence (Mimno et al., 2011) and $Precision@n$ as our evaluation metric. Topical Coherence (Mimno et al., 2011) is a metric commonly used to evaluate the performance of LDA, since it shows a well consistence with the judgement of human beings. In (Arora et al., 2012; Brody and Elhadad, 2010; Chen et al., 2014), Topical Coherence is used to compare the performance of different topic models. The better performance of topic model will get higher score in Topic Coherence. We also use $Precision@n$ (or $p@n$), a commonly used metric in information retrieval (Mukherjee and Liu, 2012; Zhao et al., 2010), for evaluation. Top words are more important in topic models, and we set *n* to 5, 10, 15 and 20. We ask two judges to label top 20 words in topics. Each topic is labeled as *correct* if it had more than half of its words related to each other; otherwise *incorrect*. Then, we asked these two judges to label each word of the top 20 words in topics which are labeled good. Since judges already had the conception of each topic in mind, each word was labeled *correct* if it consisted with the concept of the topic; otherwise *incorrect*. We use $p@n$ in two experiments shows in Section 5.3 and Section 5.4. Cohen's Kappa score for word labeling is showed in Table 2, which indicates high agreements between two judges with all the scores larger than 0.8 according to scale in (Landis and Koch, 1977).

Table 2: Cohen's Kappa for agreements of judges

	Topic Labeling	**Word Labeling**			
		p@5	p@10	p@15	p@20
Dataset1	0.858	0.806	0.830	0.879	0.859
Dataset2	0.832	0.842	0.875	0.892	0.872

Table 3: The number of correct topics

	dataset1	dataset2
bdc-TWLDA	15	14
WLDA	9	12
tf-idf-TWLDA	10	12
standard LDA	9	10
iqf-TWLDA	13	13

5.3 Comparison of Exiting LDA & TWLDA with Different Term Weighting Schemes

Different term weighting schemes in TWLDA can result in different performance. In our experiment, we firstly compare the performance of TWLDA using different state-of-the-art supervised term weighting schemes, such as $iqf \cdot qf \cdot icf$ and bdc. We also use $tf \cdot idf$ for comparision. We denote TWLDA using these schemes as tf-idf-TWLDA, iqf-TWLDA and bdc-TWLDA. Furthermore, we compare TWLDA with standard LDA and WLDA.

Quantitative Evaluation: For the reason that the process of Gibbs Sampling is random, we will get different results each time we run the model. We executed each model for 10 times, and calculate the average Topic Coherence value in each iteration. The results of the standard LDA using different term weighting schemes are shown in Figure 3. Figure 4 shows the average $precision@n$ of all good topics over two datasets, while Table 5.2 shows the number of correct topics. We find that:

- From the Topic Coherence results results, bdc-TWLDA and iqf-TWLDA outperform standard LDA and WLDA in both two datasets. bdc-TWLDA, which is an entropy-based scheme, performs the best. On the contrary, the results of TWLDA get worse when it apply tf-idf.

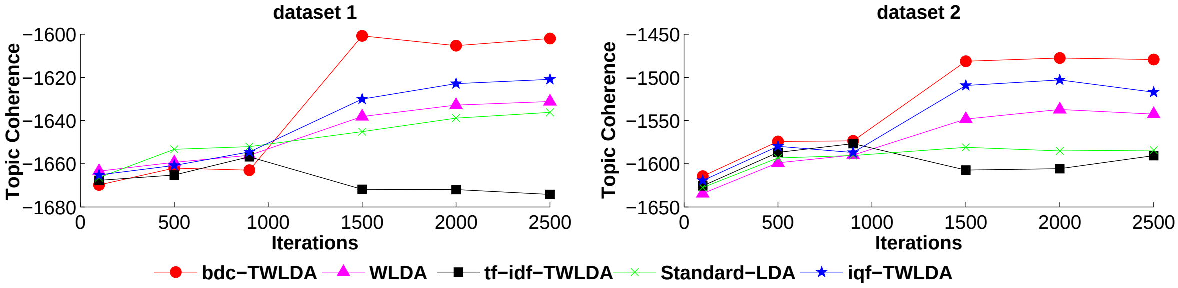


Figure 3: Comparison of TWLDA (xLDA is standard LDA) with different term weighting schemes on Topic Coherence Evaluation

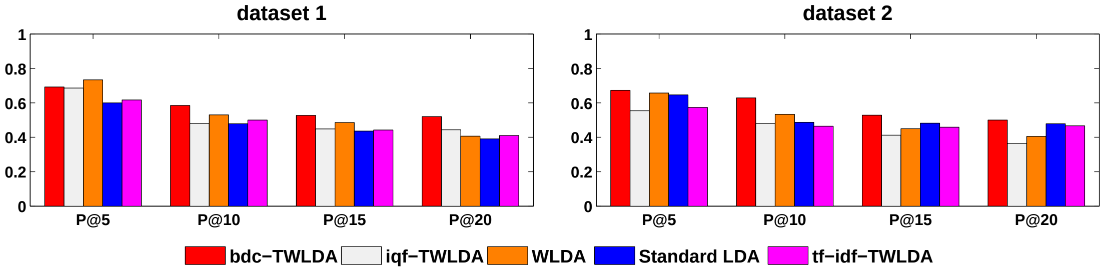


Figure 4: Average Precision@n (P@n) of coherent topics from two datasets

- From the $Precision@n$ results in Figure 4 and correct topic number in Table 5.2, bdc-TWLDA performs best and improve standard LDA by more than 10%. bdc-TWLDA generates most correct topics in both datasets, while iqf-TWLDA ranks second. Although iqf-TWLDA performs worse than WLDA in dataset2 in $Precision@n$ score, it performs better than WLDA in dataset1.

- In general, entropy-based term weighting scheme bdc performs best in both datasets. It corresponds to the experimental result of Wang et al. which shows that the bdc performs better than $iqf \cdot qf \cdot icf$. Supervised scheme $iqf \cdot qf \cdot icf$ also performs better than standard LDA at most of cases. However, tf-idf-LDA gets the worst results in both datasets.

Qualitative Results: Table 4 shows the qualitative results of LDA and bdc-TWLDA in two datasets. We choose the top 5 words of each topic generated respectively by LDA and bdc-TWLDA. We ask two judges to mark those ‘bad’ topics which are un-interpretable by human into red color. Although the labeling of topics may be subjective, we tried our best to have the consensus between two human judges. As the results shown in Table 4, standard LDA has 11 un-interpretable topics, while there are only 6 interpretable topics in the result of bdc-TWLDA. Furthermore, there are topic-indiscriminate words scattering across several topics, such as ‘phone’, ‘time’ and ‘word’. In the results of dataset2, there are 10 uninterpretable topics in standard LDA and only 6 topics in bdc-TWLDA. We do not present the results of dataset2 for the limitation of space in our paper. We also do not show the results of WLDA here, which have 11 and 9 un-interpretable topics in dataset1 and dataset2 respectively. Overall, we can see that TWLDA shows higher performance than the standard LDA.

5.4 Performance of TWLDA without Eliminating Stop Words

To demonstrate that our approach also has good performance even though we do not eliminate stop words in the preprocessing step, we execute bdc-TWLDA, WLDA and standard LDA in the following situation: eliminating stop words and retain stop words. In this experiment, we only use dataset1, since dataset2 has been pre-processed and all the stop words have been deleted. We asked two judges to label correct topics (the labeling criteria are introduced in Section 5.2). The Cohen’s Kappa value of these two judges are 0.891, which indicates they achieve high agreements. Figure 5 shows the number of correct topics in

Table 4: Quality comparison between standard LDA and bdc-TWLDA

(a) Standard-LDA

topic	word
0	**canon,well,digital,nikon,point**
1	**read,reading,games,video,videos**
2	ipad,mini,size,screen,display
3	ipad,apple,mini,love,product
4	phone,samsung,galaxy,nexus,android
5	battery,life,phone,long,time
6	**screen,phone,back,case,glass**
7	**easy,user,set,features,settings**
8	headphones,ear,sound,quality,buds
9	**amazon,price,google,buy,well**
10	**phone,buy,apple,know,back**
11	**ipad,mini,purchase,happy,product**
12	**phone,recommend,android,best,highly**
13	video,focus,mode,pictures,auto
14	bought,love,gift,loves,old
15	**time,easy,love,size,small**
16	sound,bass,headphones,price,quality
17	pictures,lens,quality,canon,zoom
18	**apps,ipad,apple,touch,free**
19	**month,plan,storage,working,work**

(b) $bdc - TWLDA$

topic	word
0	battery,life,memory,storage,gb
1	**recommend,product,highly,arrived,wifi**
2	sound,bass,music,headphones,hear
3	**ipad,mini,kindle,fire,set**
4	lens,canon,mm,picture,zoom
5	apps,wifi,internet,download,email
6	display,retina,muy,responsive,deal
7	size,small,carry,weight,hand
8	pictures,takes,quality,shots,zoom
9	nexus,google,phone,android,version
10	**money,amazon,wait,return,months**
11	ear,sony,buds,headphones,pair
12	happy,choice,glad,purchase,satisfied
13	manual,mode,video,settings,auto
14	**charge,half,phone,charging,search**
15	apple,products,ios,system,devices
16	canon,nikon,dslr,shoot,lens
17	gift,bought,card,loves,christmas
18	**reviews,front,know,piece,mind**
19	wifi,internet,data,home,web

different models when they eliminate and retain stop words. The number of correct topic in all the three models experiences a fall if they retain stop words, especially standard LDA which decreases from 9 to 2. We can also find that bdc-TWLDA still has high performance when it retain stop words.

Table 5: The number of correct topics in different models

Models	Eliminate stop words	Retain stop words	percentage of decrease
bdc-TWLDA	15	13	13%
WLDA	9	5	44%
Standard LDA	9	2	78%

5.5 Experimental Results Discussion

Our experiments show that the performance of TWLDA depends on the term weighting schemes we choose. The reason is that the capacities of different schemes measuring topic discriminating power are different. Entropy-based schemes like bdc perform the best. In information theory, words which are scattered in most of topics have larger entropy. The entropy of a word can well indicate those topic-indiscriminate words. We get the conclusion that entropy-based term weighting schemes are effective in TWLDA. In the experiments, supervised term weighting schemes outperform unsupervised term weighting schemes in TWLDA. Both bdc and $iqf \cdot qf \cdot icf$ perform better than the standard LDA, while $tf \cdot idf$ perform worse than standard LDA. The reason is that unsupervised term weighting schemes can just measure the document discriminating power of words, other than topic discriminating power.

6 Conclusions

In this paper, we firstly explore topic discriminating power of words in LDA. We observe that topics perform worse if they contain words with low topic discriminating power. These topic-indiscriminate words have negative effects on the results of LDA. In order to solve these problems, we proposed a new model called TWLDA. TWLDA can apply different supervised term weighting schemes to give topic discriminating words relatively low weights in LDA or variants of LDA. The results show that TWLDA has a significant performance while applying supervised term weighting schemes like *bdc*. The number of topic-indiscriminate words is reduced in topics generated by TWLDA with *bdc*.

7 Acknowledgements

This work is supported by National Natural Science Foundation of China (project no. 61300137, Tip-top Scientific and Technical Innovative Youth Talents of Guangdong special support program (No. 2015TQ01X633), Science and Technology Planning Project of Guangdong Province (No. 2013B010406004).

References

David Andrzejewski, Xiaojin Zhu, and Mark Craven. 2009. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2012. A practical algorithm for topic modeling with provable guarantees. *arXiv preprint arXiv:1212.4777*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812. Association for Computational Linguistics.

Nicola Burns, Yaxin Bi, Hui Wang, and Terry Anderson. 2012. Extended twofold-lda model for two aspects in one sentence. In *Advances in Computational Intelligence*, pages 265–275. Springer.

Sourav Chatterji and Lior Pachter. 2004. Multiple organism gene finding by collapsed gibbs sampling. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 187–193. ACM.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Discovering coherent topics using general knowledge. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 209–218. ACM.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In *Proceedings of ACL*, pages 347–358.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Gregor Heinrich. 2005. Parameter estimation for text analysis. Technical report, Technical report.

Gregor Heinrich. 2009. A generic approach to topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 517–532. Springer.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 815–824. ACM.

Youngjoong Ko. 2012. A study of term weighting schemes using class information for text classification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1029–1030. ACM.

Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. 2009. Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):721–735.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Edda Leopold and Jörg Kindermann. 2002. Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3):423–444.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 339–348. Association for Computational Linguistics.

Jiaul H Paik. 2013. A novel tf-idf weighting scheme for effective ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 343–352. ACM.

Xiaojun Quan, Liu Wenyin, and Bite Qiu. 2011. Term weighting schemes for question categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):1009–1021.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, pages 111–120. ACM.

Tao Wang, Yi Cai, Ho-fung Leung, Zhiwei Cai, and Huaqing Min. 2015. Entropy-based term weighting schemes for text categorization in VSM. In *27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, Vietri sul Mare, Italy, November 9-11, 2015*, pages 325–332.

Andrew T Wilson and Peter A Chew. 2010. Term weighting schemes for latent dirichlet allocation. In *human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 465–473. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65. Association for Computational Linguistics.