Applied Intelligence
https://doi.org/10.1007/s10489-020-01672-w

Targeted aspects oriented topic modeling for short texts

Jin He[1,2] · Lei Li[1,2] · Yan Wang[3] · Xindong Wu[4]

Abstract
Topic modeling has demonstrated its value in short text topic discovery. For this task, a common way adopted by many topic models is to perform a full analysis to find all the possible topics. However, these topic models overlook the importance of deeper topics, leading to confusing topics discovered. In practice, people always tend to find more focused topics on some special aspects (or events), rather than a set of coarse topics. Therefore, in this paper, we propose a novel method, Targeted Aspects Oriented Topic Modeling (TATM), to discover more focused topics on specific aspects in short texts. Specifically, each short text is assigned to only one targeted aspect derived from an enhanced Dirichlet Multinomial Mixture process (E-DMM). This process helps group similar words as many as possible, which achieves topic homogeneity. In addition, TATM discovers the topics for each targeted aspect from as many angles as possible by performing target-level modeling, which achieves topic completeness. Thus, TATM can make a balance between the two conflicting properties without employing any additional information or pre-trained knowledge. The extensive experiments conducted on five real-world datasets demonstrate that our proposed model can effectively discover more focused and complete topics, and it outperforms the state-of-the-art baselines.

Keywords Topic modeling · Text mining · Short text clustering · Focused analysis

1 Introduction

One of the important text mining tasks is to discover topics from a vast amount of documents (also called a corpus). Along with the prevalence of social media, like Twitter, short texts have become a popular type of corpus [15, 17]. Typically, a short text contains dozens of words while a normal text often contains more than hundreds of words [38]. For example, the average length of a tweet is usually less than 20 words, but a news article usually contains more than hundreds of words. This smaller text length naturally leads to the sparsity problem of short texts, which makes the conventional topic models, like Latent Dirichlet Allocation (LDA) [4], very hard to achieve as good performance as on normal texts. Thus, researchers have proposed many topic models for short texts. However, they mainly perform full analysis, which tends to find all the topics from broad aspects, thus overlooking the importance of focused topics on some specific aspects.

In contrast, in practice, people always tend to find more focused topics on some specific aspects, rather than the general topics discovered by full analysis. These aspects are associated with some popular events, each of which covers several very related topics, and these aspects do not overlap each other, meaning that they are highly discriminative. We refer to these aspects as *targeted aspects* or *targets* in this paper. For example, *Superbowl* (the annual championship game of the National Football League in the United States) and the *Oscars Ceremony* (an internationally well-known film award) are two popular aspects/events discussed on Twitter. *Superbowl* covers several interesting topics, such as the *commerce* of the game and its opening *show*. And the *Oscars Ceremony* is associated with the topics about films, such as an awarded film

✉ Lei Li
lilei@hfut.edu.cn

Jin He
jinhe@mail.hfut.edu.cn

Yan Wang
yan.wang@mq.edu.au

Xindong Wu
wuxindong@mininglamp.com

[1] Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230009, China

[2] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

[3] Department of Computing, Macquarie University, Sydney NSW 2109, Australia

[4] Mininglamp Academy of Sciences, Mininglamp Technologies, Beijing 100084, China

Published online: 07 March 2020

King's Speech. There is no direct relationship between the two aspects, and the topics covered by the two aspects have no overlap. When a person wants to find the topics about *Superbowl*, the topics about *Oscars Ceremony* are not expected.

In the above scenario, the existing full analysis-based approaches cannot assign each topic to a highly related aspect, leading to their captured topics coarse and confused for a person who wants to gain deeper insights for a specific aspect. Thus, there is a need for targeted aspects oriented methods. Targeting this problem, researchers have proposed several new models, which can be generally classified into two categories: (1) identifying targets with pre-trained knowledge [8, 33, 37], and (2) identifying targets with additional information [2, 11, 30, 31, 36].

The models in the first category internally employ pre-trained knowledge, such as learning topics from the lexical semantics encoded by word embedding [8]. This is useful as the acquired knowledge is strongly related to the latent topics. However, the widely used representation of the knowledge, like Word2Vec [19], cannot discover the exact semantic coherence of the words, leading to poor topic homogeneity. The models in the second category employ a jointly modeling process with external information, such as the specified background words [11, 30], temporal information [2], and sentiments [31, 36]. This is certainly useful as the additional information can help capture some words specific to a targeted aspect. However, these models aggravate the sparsity problem when simply discarding the words that cannot be directly identified by the additional information. Therefore, they cannot discover the topics from as many angles as possible, leading to poor topic completeness.

To address the above-mentioned problems, a natural way is to achieve the two important properties, i.e., topic homogeneity and topic completeness. The former means the topics about the same aspect should be grouped. The latter means the topics of a targeted aspect should cover as many angles as possible. However, the existing models cannot simultaneously take the two properties into account. This is because there exists conflict when achieving both properties, leading to the difficulty to obtain the best performance for both of them. In addition, due to the sparsity problem, this task becomes more complex in short texts.

In this paper, we propose a novel method, Targeted Aspects Oriented Topic Modeling (TATM), to discover more focused topics on targeted aspects in short texts. Targeting the conflict between the two properties, TATM identifies the targeted aspects with similar words and discovers topics for each targeted aspect from as many angles as possible. Specifically, each short text is assigned to only one aspect rather than one topic, and the topics are derived from the target level rather than the document level. This helps to discover more focused and comprehensive topics.

Overall, the major characteristics and contributions of our work can be summarized as follows:

- We propose a novel Targeted Aspects Oriented Topic Modeling (TATM) method for short texts. Specifically, the proposed model can effectively identify the targeted aspects via an enhanced Dirichlet Multinomial Mixture process (E-DMM) and perform target-level topic modeling without employing any pre-trained knowledge or additional information;
- The proposed model can discover focused topics from as many angles as possible on targeted aspects consisting of similar words, which makes it able to achieve both topic homogeneity and topic completeness for short text topic discovery;
- The extensive experiments conducted on five real-world datasets demonstrate the effectiveness of TATM to discover more focused and complete topics in short texts, and it outperforms the state-of-the-art baselines.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the proposed TATM model in detail. Section 4 reports the experimental results and discussions. The conclusion is shown in Section 5.

2 Related work

In this section, we briefly discuss the related work from the following two perspectives: (1) topic models with full analysis, and (2) topic models on targeted aspects.

2.1 Topic models with full analysis

Based on conventional method LDA [4], many full analysis-based have been proposed for dealing with the sparsity problem in short texts [9, 22]. These models adopt one of the following three popular strategies: (1) Dirichlet Multinomial Mixture (DMM), (2) global word co-occurrence, and (3) self-aggregation.

Based on Strategy (1), some proposed models follow a simple assumption that each short text covers only one topic [21, 33, 34]. For example, Nigam et al. [21] use an Expectation Maximization (EM)-based algorithm as a supervised procedure of DMM. Based on Strategy (2), some studies try to enrich the original short texts with global word co-occurrence patterns to infer the latent topics [6, 7, 39]. For instance, Cheng et al. [7] propose BTM to directly extract the co-occurrence word pairs on the whole corpus, rather than modeling on the document level. Zuo et al. [39] enrich short texts by using a word co-occurrence network. Based on Strategy (3), some approaches employing the aggregation

algorithm to enhance word-topic correlation, such as incorporating the aggregation at the topic level, or the document level, to enhance the word-topic correlation [3, 13, 18, 24, 29]. For example, Blair et al. [3] aggregate the topics captured by a group of topic models according to the Jensen-Shannon divergence between them to improve topic coherence.

However, these full analysis-based topic models tend to find all the possible topics from the whole corpus with broad aspects, leading to the captured topics confusing and too general. Moreover, the simple one-topic assumption that is widely adopted by these models does not fit each short text.

2.2 Topic models on targeted aspects

Because the scale of aspects that people talk on social media is greatly expanding, targeted topic modeling has emerged. Such targeted topic modeling approaches gain deeper insights into a large-scale short text collection than that of the full analysis-based models. As discussed in Section 1, the existing methods for targeted topic modeling can be generally classified into two categories: (1) modeling with internally pre-trained knowledge, and (2) modeling with additional information.

The models in the first category identify the targeted aspects based on the mutual reinforcement between topics and words. One of the widely used methods is word embedding [23, 26]. Clustering is another effective way [16, 33]. For example, Yin et al. [33] propose a quickly converged Gibbs Sampling algorithm GSDMM for inferring the targeted aspects. Yu et al. [35] propose the DPMFS model to select appropriate features for clustering although it is slow to converge. And they further propose the DMAFP model to improve DPMFS [14]. The models in the second category identify the targeted aspects based on the additional constraints over the topic-word distribution. For example, some models extend BTM [7] by employing the statistics of words as a constraint to enhance the topic-word connection [11, 12, 32]. Another type of constraint is to filter unrelated words by using external information, such as keywords indicator [28, 30] or sentiment [1, 25, 31].

However, these models cannot achieve both two important properties, i.e., topic homogeneity and topic completeness. This is because the prevalent representation of pre-trained knowledge cannot capture the exact lexical semantics, leading to poor topic homogeneity. In addition, the word-topic constraint from additional information aggravates the sparsity problem as it overlooks the words that cannot be directly identified, leading to poor topic completeness.

3 Targeted aspects oriented topic modeling

In this section, we present a novel method, Targeted Aspects Oriented Topic Modeling (TATM), to capture more focused and complete topics in short texts. As mentioned in the introduction part, people always tend to find more focused topics on some special aspects (referred to as *targeted aspects* or *targets* in this paper), rather than the coarse topics generated by full analysis-based methods. For this issue, given a short text corpus of a broad range of aspects, our proposed model needs to achieve a balance between the two properties, i.e., topic homogeneity and topic completeness.

Since our proposed model needs to identify the targeted aspects from the entire corpus, the intuitive idea of our model is that a targeted aspect tends to group similar words as many as possible, which can help achieve topic homogeneity. For short texts, our basic hypothesis is that each document focuses on only one aspect but can cover several different topics. Although it might not always be correct, this one-target hypothesis relaxes the one-topic assumption from DMM-based methods [21, 33, 34] and can achieve good performance on short texts (to be shown in Section 4). Following this hypothesis, an enhanced Dirichlet Multinomial Mixture (E-DMM) process is proposed to identify the targeted aspects, then TATM can perform a target-level topic modeling to find the topics from as many angles as possible, which can help achieve topic completeness.

In the following subsections, firstly, we describe the proposed TATM model with its generative process. Then, we introduce the design of the identification of targeted aspects with the E-DMM process. At last, we present the inference of the hidden variables. The notations used in this paper are listed in Table 1.

Table 1 Notation description

D	# of documents in the corpus
V	# of words in the vocabulary
T	# of targeted aspects in the corpus
K	# of topics in each targeted aspect
B	# co-occurrence word pairs in each targeted aspect
π	multinomial distribution over document-target
θ	multinomial distribution over target-topic
ϕ	multinomial distribution over topic-word
γ,α,β	Dirichlet priors for π, θ, ϕ
z	the topic indicator
d	the document indicator
t	the targeted aspect indicator
b	the word pair indicator
N_d	# of words in document d
D_t	# of documents in targeted aspect t
N_t	# of words in targeted aspect t
n_t^w	# of word w occurring in targeted aspect t
n_k^w	# of word w assigned to topic k

3.1 Model description

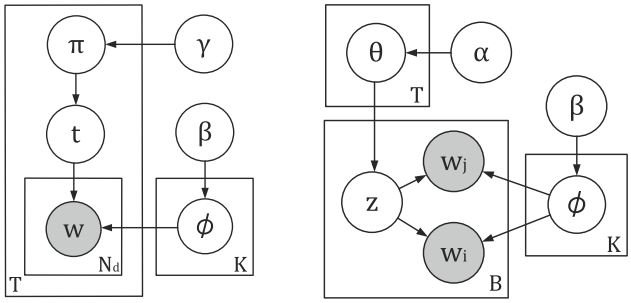
Formally, we assume that the corpus contains D short texts with a vocabulary of size V. And the corpus contains T popular targeted aspects. D_t denotes the number of documents assigned to a targeted aspect t and N_t denotes the number of words in t. For a document d, the targeted aspect it focuses on can be denoted as t_d. N_d denotes the number of words in document d.

Figure 1 presents the graphical representation of our TATM model. As shown in Fig. 1a, Phase 1 identifies T targeted aspects via an enhanced Dirichlet Multinomial Mixture (E-DMM) process. In this phase, each document d is assigned to a targeted aspect t_d. As shown in Fig. 1b, Phase 2 performs a target-level topic modeling to capture K topics for each targeted aspect. These two phases correspond to topic homogeneity and topic completeness, respectively. This means, for each short text, there is only one targeted aspect responsible for its generation, and the K topics captured for the targeted aspect are highly coherent. Algorithm 1 describes the generative process of TATM, where Lines 1-4 correspond to Phase 1 while Lines 5-13 correspond to Phase 2. Here, γ, α, and β are three hyper-parameters for Dirichlet distribution, and π, θ, and ϕ denote the multinomial distributions over document-target, target-topic, and topic-word level, respectively.

Algorithm 1 Generative process for TATM

Input:
D documents of short texts;
The number of targeted aspects, T;
The number of topics in each targeted aspect, K;
The hyper-parameters, α, β, γ.
Output:
The topic words collection.
1: Draw $\pi \sim \text{Dir}(\gamma)$
2: **for** $d = 1$ to D **do**
3: Sample a targeted aspect $t_d \sim \text{Multi}(\pi)$
4: **end for**
5: **for** Targeted aspect $t = 1$ to T **do**
6: Draw $\theta_t \sim \text{Dir}(\alpha)$
7: **for** Topic $z = 1$ to K **do**
8: Draw $\phi \sim \text{Dir}(\beta)$
9: Sample a co-occurrence word pair b_i $(i \in [1, \text{B}])$
10: Sample a topic $z_k \sim \text{Multi}(\theta_t)$ $(k \in [1, \text{K}])$
11: Sample words $w_{i,1}$, $w_{i,2}$ for $b_i \sim \text{Multi}(\phi_{z_k})$
12: **end for**
13: **end for**

In addition to the commonly used hidden variables z, α, and β, TATM also involves t, γ, and π as new hidden variables to identify the targeted aspects. When generating a document d, TATM firstly samples a targeted aspect t_d in Phase 1 via E-DMM. Then, TATM performs target-level topic modeling by extracting co-occurrence word pairs for each targeted aspect. Thus, following the above generative process, we can denote the conditional probability of a word pair b on the three multinomial distributions θ, β, and π as follows:



(a) Phase 1: E-DMM (b) Phase 2: Target-level topic modeling

Fig. 1 Graphical representation of TATM

$$p(b|\theta, \phi, \pi) = p(b|\alpha, \beta, \pi_t)p(T = t, \pi_t|\gamma) \quad (1)$$

where $T = t$ denotes the current co-occurrence word pair b is sampled from targeted aspect t. The computation of this conditional probabilistic distribution can be solved by the variational inference, such as Gibbs Sampling [10], which is shown in Section 3.3.

3.2 Identifying targeted aspects

Since our TATM model aims to identify the targeted aspects and find the topics over them, one of the key problems is how to identify the targeted aspects from the entire corpus denoted as Phase 1 of TATM. In this subsection, we present how our TATM model solves this issue via an enhanced Dirichlet Multinomial Mixture (E-DMM) process, so as to achieve topic homogeneity.

For identifying targeted aspects, our intuitive idea is that each targeted aspect tends to cover similar words as many as possible. Thus, Phase 1 can be regarded as a generative process over the words belonging to each corresponding target. Based on the Naive Bayes (NB) theory that each word is generated independently, given a document d with N_d words, the probability of the occurrence of a targeted aspect t can be denoted as follows:

$$\begin{aligned} p(T = t|d, \pi) &= \frac{p(d|T = t)p(T = t|\pi)}{p(d)} \\ &= \frac{\prod_{t=1}^{T} p(w|T = t)p(T = t|\pi)}{\prod_{w \in N_d} p(w)} \end{aligned} \quad (2)$$

To implement this process, we proposed an enhanced Dirichlet Multinomial Mixture (E-DMM) process. It relaxes the one-topic assumption of standard DMM by employing

one-target assumption for each short text. Since E-DMM aims to identify the targeted aspects with similar words as much as possible, it achieves topic homogeneity. The pseudo-codes of E-DMM is shown in Algorithm 2.

As shown in Algorithm 2, given a document d, firstly, E-DMM samples a new aspect label t_{new} as a candidate (Line 6). Secondly, E-DMM computes cosine similarity between d and the documents already included in t_{new}, and get an average value $AvgSim(d,t_{new})$ (Line 7). This value is to evaluate the similarity between the current document d and the candidate aspect t_{new}, which helps group similar words into the same targeted aspect as many as possible. Thirdly, a similarity threshold s is adopted to decide whether to assign the candidate t_{new} to the current document d or randomly sample a targeted aspect t for d (Lines 8-11). Finally, after enough times of iterations, we obtain the top T targeted aspects with the most number of documents, which are used to capture topics in Phase 2.

Algorithm 2 Enhanced DMM for TATM

Input:
D documents of short texts;
The number of targeted aspects, T;
The hyper-parameter, γ;
Similarity threshold, s;
The number of iterations, $iterNum$.
Output:
T targeted aspects with most short texts.
1: Initialize targeted aspect assignments randomly
2: **for** $iteration = 1$ to $iterNum$ **do**
3: **for** $d = 1$ to D **do**
4: Record the current targeted aspect of d: t_{old}
5: Update $D_{t_{old}} = D_{t_{old}} - 1$, $N_{t_{old}} = N_{t_{old}} - N_d$
6: Sample a targeted aspect t_{new} with Eq. 2
7: Compute $AvgSim(d, t_{new}) = \frac{\sum_{c=1}^{D_{t_{new}}} CosSim(d,c)}{D_{t_{new}}}$
8: **if** $AvgSim(d, t_{new}) < s$ **then**
9: Assign a targeted aspect t randomly as t_{new}
10: **end if**
11: Assign targeted aspect t_{new} to d
12: Update $D_{t_{new}} = D_{t_{new}} + 1$, $N_{t_{new}} = N_{t_{new}} + N_d$
13: **end for**
14: **end for**

It is worth mentioning that, to reduce the impact of the targeted aspects with a small number of documents on the sampling, we discard the targeted aspects with less than 2 documents before the next iteration. We set the number of iterations for E-DMM (i.e., *iterNum* in Algorithm 2) to 100 and the similarity threshold s to 0.025.

3.3 Gibbs sampling for TATM

In this subsection, we introduce the collapsed Gibbs Sampling process in TATM to infer the hidden variables, i.e., document-target assignment (t), target-topic assignment (z), and word-topic assignment (word pair-topic assignment b). It can be described with two essential steps according to the above-mentioned two phases.

First, suppose that each word w can appear more than one time, Phase 1 identifies T targeted aspects for the given D short texts via E-DMM process. For each document d, the probability of the targeted aspect assignment in (2) can be described with a conditional probability distribution as follows:

$$p(T = t|d, \pi) = \frac{p(T = t|\pi)\prod_{w\in N_d} p(w|T = t)}{\sum_{t=1}^{T} p(T = t)\prod_{w\in N_t} p(w|T = t)} \tag{3}$$

Here, N_t is the number of words in targeted aspect t, and $p(w|T = t)$ denotes the probability of word w conditioned on targeted aspect t.

According to [27], Dirichlet distribution can be used as the prior of multinomial distribution, namely, $\pi\sim$ Dir(γ) and $\phi\sim$ Dir(β). Thus, given the hyper-parameters γ and β, the conditional probability distribution in (3) can be derived as follows:

$$\begin{aligned} & p\left(T = t\ |T_{t,\neg d}, \gamma, \beta\right) \\ \propto\ & \frac{D_{t,\neg d} + \gamma}{D - 1 + \gamma T} \frac{\prod_{w\in N_d}\prod_{i=1}^{n_d^w}\left(n_{t,\neg d}^{w} + \beta + i - 1\right)}{\prod_{j=1}^{N_d}\left(n_{t,\neg d} + N_t\beta + j - 1\right)} \end{aligned} \tag{4}$$

Here, $\neg d$ means short text d is excluded, $n_{t,\neg d}^{w}$ is the number of word w occurring in targeted aspect t excluding d.

Second, Phase 2 captures K topics for each targeted aspects from as many angles as possible. There are three hidden variables to be estimated in this phase, i.e., the topic assignment indicator z, the target-topic distribution θ, and the topic-word distribution ϕ. Instead of modeling at the document level, TATM extracts B co-occurrence word pairs at the target level, which achieves topic completeness. We adopt the representation of co-occurrence word pairs proposed by BTM [7]. Therefore, given a targeted aspect t, i.e., $p(T = t|\pi_t)$, the conditional probability distribution of a co-occurrence word pair b_i can be derived as follows:

$$p(b_i|\pi, \theta, \phi) = \sum_{k=1}^{K} C\left(w_{i|k}\right) p(z_i = k|\theta_k) p(T = t|\pi_t) \tag{5}$$

Here, $C(w_{i|k})$ is the product of $p\left(w_{i,1}|z_i = k, \phi_{k,w_{i,1}}\right)$ and $p\left(w_{i,2}|z_i = k, \phi_{k,w_{i,2}}\right)$. $C(w_{i|k})$ and $p(z_i = k|\theta_k)$ of (5) make TATM can model at the target level, which achieves topic completeness. And $p(T = t|\pi_t)$ of (5) helps narrow the range of topics, which achieves topic homogeneity.

Similar to (4), given the conjugate Dirichlet priors α and β, the multinomial distributions θ and ϕ can be integrated. Thus,

for a word pair b_i extracted from targeted aspect t, we only need to sample its topic assignment z_i, and the conditional probability distribution is derived as follows:

$$p(T=t, z_i=k|z_{\neg i}, B, T) \propto \eta_t^{w_d} \frac{(n_{k,\neg i}+\alpha)\left(\prod_{j\in b_i} n_{w_j,\neg i|k}+\beta\right)}{\left(n_{\cdot,\neg i|k}+N_t\beta+1\right)\left(n_{\cdot,\neg i|k}+N_t\beta\right)} \quad (6)$$

Here, $\eta_t^{w_d}$ is the abbreviation of (4), $n_{\cdot,\neg i|k}$ is the summation of the number of words assigned to topic k excluding b_i, and $n_{\cdot,\neg i|k}=\sum_{w=1}^{N_d} n_{w,\neg i|k}$.

4 Experiments

In this section, we present the experiments conducted on five real-world short text datasets, all of which are publicly available. Specifically, our experiments wish to answer the following six key questions:

Q1: How do the parameters, such as similarity threshold s in the E-DMM process, affect the effectiveness of our model?
Q2: How effective is our model for identifying the targeted aspects?
Q3: How are the captured topics semantically coherent and do the topics describe multiple angles of a targeted aspect?
Q4: How are the captured topics consistent with the corresponding targeted aspects?
Q5: How effective is our model for short text classification?
Q6: How efficient is our model for topic modeling?

4.1 Datasets description and setup

Dataset. We use five real-world short text datasets for evaluating the performance of TATM, all of which are publicly available. Four of them are divided from GoogleNewsAggregator[1] (NewsTitle for short) according to category labels and one is Tweet dataset[2] (Tweet for short).

- *NewsTitle* consists of 422,937 online news articles, including news titles, corresponding web pages, category labels, and cluster labels generated from news events[1]. We further divide the original dataset into four datasets according to the four category labels: *Business* (BSet for short), *Technology* (TSet for short), *Entertainment* (ESet for short), and *Medicine* (MSet for short). And we only use the news titles as short texts. Each text has a cluster label according to its news events, which is regarded as the targeted aspect label. We use these four datasets to evaluate the performance of different methods on short texts with different targeted aspects.
- *Tweet* consists of 2,472 tweets with 89 highly relevant queries. Each tweet is labeled in the 2011 and 2012 microblog tracks at Text REtrieval Conference (TREC)[2]. We regard the queries as targeted aspects, then the highly-relevant tweets of the same query can be grouped into the same targeted aspect.

For each dataset, we conduct the following preprocessing: (1) convert letters into lowercase, (2) remove meaningless words and the characters not in Latin as stop words, (3) remove the words whose frequency is less than 2, and (4) remove the documents whose length is smaller than 5 words. The statistics of these five datasets are summarized in Table 2.

Baselines We compare the performance of TATM with the following seven methods in two groups: (1) five topic models selected for comparing the performance of topic discovery, and (2) two clustering methods selected for comparing the performance of targeted aspects identification.

(1) Topic discovery:

- Latent Dirichlet Allocation (LDA) [4] is a well-known topic model, which is based on the full analysis. We compare TATM with it to demonstrate the effectiveness of targeted aspects oriented topic modeling.
- Biterm Topic Model (BTM) [7] is a typical topic model for short texts, which performs full analysis by employing global co-occurrence word pairs, i.e., the mechanism of biterms.
- Targeted Topic Model (TTM) [30] is a state-of-the-art topic model, which learns the topics for a particular targeted aspect with a set of

Table 2 Statistics of the datasets (D: the number of documents, T: the number of targeted aspects, V: the size of vocabulary, AvgLen: the average number of words in a document)

Dataset	D	T	V	AvgLen
BSet	115,967	2,019	32,829	9.68
TSet	108,503	1,789	29,975	9.51
ESet	152,828	2,076	36,035	10.45
MSet	45,639	1,347	17,897	9.09
Tweet	2,472	89	5,061	8.04

[1] https://data.world/uci/news-aggregator
[2] http://trec.nist.gov/data/microblog.html

pre-given keywords. We use the words with the highest frequency as the initial keywords.

- Attentional Segmentation Topic Model (ASTM) [29] is a state-of-the-art topic model for short texts, which employs the self-aggregation strategy with pre-trained word embedding.
- Aggregated Topic Model (ATM) [3] is a state-of-the-art topic model, which employs Jensen-Shannon divergence to aggregate the captured topics.

(2) Targeted aspects identification:

- Dirichlet Multinomial Mixture with collapsed Gibbs Sampling (GSDMM) [33] is a state-of-the-art model for short text clustering, which adopts the standard one-topic assumption.
- K-means [16] is a well-known unsupervised clustering method, which computes the distance of each document to the centroid.

Parameter settings. For fair comparison, we optimize the parameters as follows:

- The settings of Dirichlet priors (i.e., α and β), the number of topics K, and the number of iterations for Gibbs Sampling. First, we set $\alpha = 0.1$ for all the compared methods except that $\alpha = 50/K$ for BTM, and $\alpha = 0.05$ for LDA, so as to achieve their best performance. Second, we set $\beta = 0.01$ for all the compared methods except that $\beta = 0.1$ for GSDMM to achieve the best performance. Third, for the number of topics K, we set K changing from 10 to 70 with a step of 10. In addition, the Gibbs Sampling runs 500 iterations for all the compared methods, which is sufficient for convergence.
- The settings of enhanced Dirichlet Multinomial Mixture (E-DMM) process. Since our E-DMM process is similar to a clustering method, there are three critical parameters, i.e., the initial number of targeted aspects T_{init}, the number of iterations *iterNum*, and the hyper-parameter γ for document-target distribution. For T_{init}, we set it is close to the true number for TATM and K-means. In addition, we set T_{init} is larger than the true number for GSDMM as suggested by its paper, so as to achieve its best performance. In terms of *iterNum*, we set it to 100 and report the number of targeted aspects found in each iteration (to be shown in Section 4.3.2). And we optimize the hyper-parameter γ by using a grid search with cross-validations in {0,0.01,0.1,1,10} and eventually set it to 0.1.

4.2 Experimental design and evaluation metrics

Experimental design. To demonstrate the effectiveness of our TATM model, we design the following six tasks associated with the above-mentioned questions:

- **Task 1 (Influence of parameters, results to be shown in Section 4.3.1):** Evaluating the performance of our model under different parameter settings. This task aims to investigate the influence of parameters on our model.
- **Task 2 (Target-Iteration, results to be shown in Section 4.3.2):** Analyzing the number of targeted aspects found in each iteration. This task aims to validate the effectiveness of our model for identifying targeted aspects and its convergence speed.
- **Task 3 (Topic quality, results to be shown in Section 4.3.3):** Evaluating the quality of the captured topics in terms of topic homogeneity and topic completeness, respectively. This task aims to validate that TATM can capture coherent and complete topics for targeted aspects.
- **Task 4 (Topic purity, results to be shown in Section 4.3.4):** Evaluating the consistency between discovered topics and the corresponding targeted aspect. This task aims to validate that TATM can capture the topics for a

specific target with as little noise as possible.

- **Task 5 (Short text classification, results to be shown in Section 4.3.5):** Evaluating the performance of short text classification. This task aims to validate that TATM can learn high-quality document features.
- **Task 6 (Time complexity, results to be shown in Section 4.3.6):** Comparing the time complexity and time consumption of TATM with the baselines. This task aims to validate the time efficiency of TATM.

Evaluation metrics. To evaluate the experimental results, we adopt the following metrics.

For Task 1, we investigate how the performance of our TATM model is affected by the parameters, including the priors α, β, and γ, and the similarity threshold s in the E-DMM process. The effectiveness of TATM can be automatically evaluated by the metric *Pointwise Mutual Information* (PMI), which reports the degree of the coherence between captured topics and is widely used in the literature [7, 12, 20, 29]. Given a topic k, the PMI score is computed as follows:

$$PMI(k) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{7}$$

Here, N is the number of most probable words in topic k, $p(w_i,w_j)$ is the probability of words w_i and w_j co-occurring, and $p(w_i)$ and $p(w_j)$ are the probabilities that words w_i and w_j appear, respectively. Typically, the averaged PMI scores of all captured topics can be used to measure the overall coherence.

For Task 2, we investigate the number of targeted aspects found in each iteration. It reports the convergence speed of each model. In addition, comparing with the true number, the experimental results can validate the effectiveness of our model in identifying targeted aspects.

For Task 3, we conduct quantitative and qualitative evaluations, respectively. First, for quantitative evaluation, many conventional metrics try to estimate the likelihood of testing data on the parameters inferred from training data [24]. However, this likelihood relationship is not a straightforward indicator of topic coherence [5]. Thus, we also adopt the PMI metric for quantitative evaluation. Second, for qualitative evaluation, we visualize the captured topical words to investigate the completeness of topics on some special aspects, namely, the different angles the topics can cover. In addition, we also check the correctness according to their coherence to the aspect.

For Task 4, we adopt the *purity* metric proposed by clustering methods [24]. This is consistent with the objective goal of targeted aspects oriented topic modeling, which aims to distill the topics that are in maximum alignment with the content covered by the corresponding targeted aspects. The *purity* score can be computed by comparing the N most probable words in each topic with the words discovered from all the documents within the corresponding targeted aspect as follows:

$$purity = \frac{1}{KN} \sum_i \max_j |\tau_{z_i} \cap \tau_{s_j}| \tag{8}$$

Here, z_i and s_j denote the testing topic and the standard global topic, respectively. We view the topics discovered by LDA from each targeted aspect as the standard global topics. And τ denotes the set of N most probable words from topics.

For Task 5, we adopt the *accuracy* of classification as the evaluation metric according to the cluster labels. It is reasonable to evaluate each method by applying the captured topics to classification as an external task. We conduct this task on the Naive Bayes Classifier (NBC)[3] as it has a strong probability foundation.

For Task 6, we adopt *running time* as the evaluation metric, which only includes the time of iterations for topic modeling. It does not include the random initialization part at the beginning and the part of saving topics at the ending. It is worth mentioning that the results are averaged after running three times for each method.

4.3 Experimental results and analysis

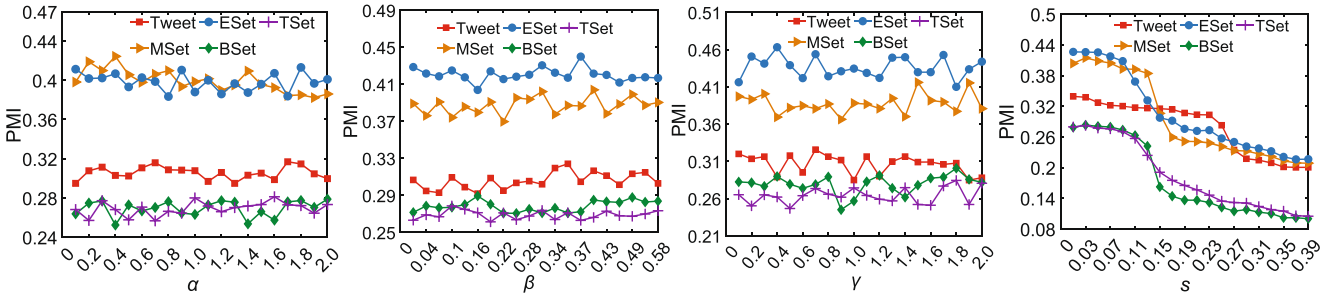
In this subsection, we present the experimental results of our TATM model and baselines, which are associated with the above six tasks.

4.3.1 Influence of parameters (for Q1)

In this part, we answer **Q1** with the experimental results of Task 1. The hyper-parameters (α, β, γ, and the similarity threshold s) could affect the performance of our TATM model. Therefore, we analyze their impact on TATM by conducting experiments under different parameter settings.

The experiments are divided into four groups: (1) changing the value of α in the range of [0.1, 2.0] with a step of 0.1 under $\beta = 0.01$, $\gamma = 0.1$, and $s = 0.025$ (see Fig. 2a); (2) changing the value of β in the range of [0.01, 0.58] with a step of 0.03 under $\alpha = 0.1$, $\gamma = 0.1$, and $s = 0.025$ (see Fig. 2b); (3) changing the value of γ in the range of [0.1, 2.0] with a step of 0.1 under $\alpha = 0.1$, $\beta = 0.01$, and $s = 0.025$ (see Fig. 2c); and (4) changing the value of s in the range of [0.025, 0.5] with a step of 0.025

[3] https://github.com/ptnplanet/Java-Naive-Bayes-Classifier



(a) different values of α (b) different values of β (c) different values of γ (d) different values of s

Fig. 2 Performance of TATM under different values of **a** α, **b** β, **c** γ, and **d** s, respectively

under $\alpha = 0.1$, $\beta = 0.01$, and $\gamma = 0.1$ (see Fig. 2d). And the number of topics is fixed as $K = 20$.

Figure 2 a and b plot the performance of TATM under the different values of α and β, which control the target-topic distribution and topic-word distribution, respectively. When increasing the value of α or β, we can see that, the performance of TATM is stable on the five datasets, especially on the TSet dataset. Therefore, it is reasonable for our parameter settings as $\alpha = 0.1$ and $\beta = 0.01$.

Figure 2c plots the performance of TATM under the different value of γ, which controls the document-target distribution, i.e., the degree that each document focuses on a specific aspect. We can see that, the performance of TATM is more sensitive to the change of γ than that of α and β, i.e., the curve fluctuates more than that of Fig. 2 a and b, especially when γ is in the range of [0.2, 0.8]. This is because that the targeted aspects covered by each short text do not overlap, which makes the document-target distribution is sensitive to γ.

Figure 2d plots the performance of TATM under the different value of s, which controls the degree of similarity between the words in the same targeted aspect. When increasing the value of s, we can see that, the performance of TATM is more stable on the Tweet dataset than the rest four news datasets, especially when s is in the range of [0.15, 0.27]. This result demonstrate that TATM can effectively identify the targeted aspects by gathering similar words together, which helps achieve topic homogeneity. In addition, the similarity threshold s should not be too large, because the proposed model performs target-level modeling, which helps achieve topic completeness.

Summary In general, our TATM model can capture highly coherent topics by employing the E-DMM process and performing target-level topic modeling. The result demonstrates that the two hyper-parameters in the E-DMM process, i.e., γ and s, are both critical for TATM, each of both can help effectively identify the targeted aspects with highly similar words. Moreover, in order to achieve both topic homogeneity and topic completeness, the similarity threshold s should not be too large.

4.3.2 Target-Iteration (for Q2)

In this part, we answer **Q2** with the experimental results of Task 2, which demonstrate the effectiveness of our TATM model for identifying targeted aspects and its high convergence speed. We compare TATM with two clustering baselines GSDMM [33] and K-means [16] on two datasets Tweet and MSet. As analyzed in the parameter settings part, for TATM and K-means, we set the initial number of targeted aspects T_{init} to be close to the true numbers of datasets, namely, $T_{init} = 200$ on Tweet and $T_{init} = 2000$ on MSet, respectively. For GSDMM, we set T_{init} to be larger than the true numbers, namely, $T_{init} = 300$ on Tweet and $T_{init} = 3000$ on MSet, respectively. Figure 3 shows the experimental results.

First, we can see that, our TATM model outperforms the two baselines on both datasets. On the one hand, the number of targeted aspects found by TATM converges quickly and almost gets stable within 10 iterations. On the other hand, the stable number of targeted aspects found by TATM is closer to the true numbers than that of baselines. Second, the performance of GSDMM is comparable on the Tweet dataset, but its performance becomes poor on the larger dataset MSet (i.e., Tweet has only 89 target labels while MSet has more than 1300 target labels). Comparing the two datasets Tweet and MSet, the result indicates that TATM has better robustness than that of GSDMM when the size of data increases.

Summary 1 In general, our TATM model is effective for identifying targeted aspects with the closest result to the true number. In addition, TATM is efficient for this task with the fastest convergence speed, even compared to the state-of-the-art method GSDMM. In particular, TATM has more robust performance on large-scale data than baselines.

4.3.3 Topic quality (for Q3)

In this part, we answer **Q3** with the experimental results of Task 3, which evaluates both topic homogeneity with a

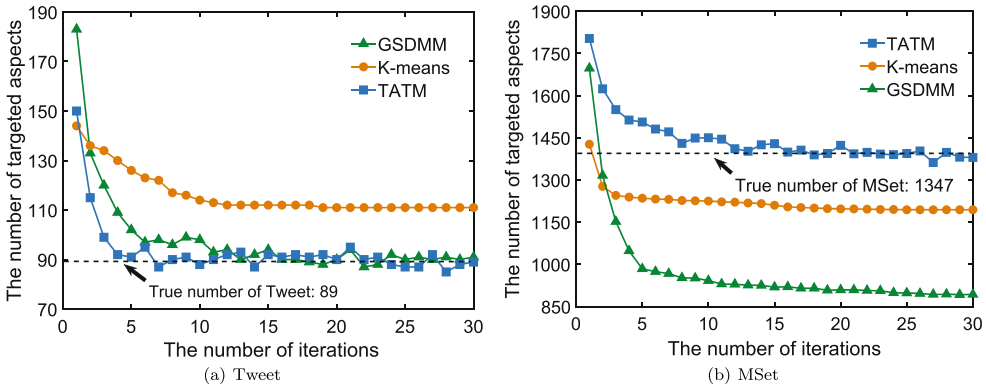


Fig. 3 Performance comparison with two clustering methods of number of targeted aspects found under different number of iterations on two datasets **a** Tweet and **b** MSet

quantitative comparison and topic completeness with a qualitative comparison. For all experiments in this part, the higher PMI score indicates the better topic quality.

Quantitative evaluation To conduct the quantitative evaluation, we first present the PMI scores of the topics captured by all the compared methods with a fixed number of topics $K = 70$. This is to show the improvement of our TATM model on all the baselines. Table 3 presents this result with the best performance is in bold font. Second, we investigate the performance of all compared methods under the different number of topics K changing from 10 to 70 with a step of 10. Figure 4 plots the experimental results on five datasets.

From Table 3, we can see that, TATM outperforms the other four topic models, including two state-of-the-art methods ATM and ASTM. Specifically, on average, TATM improves LDA, ATM, BTM, TTM, and ASTM by 63.91%, 43.99%, 12.88%, 44.22%, 13.44%, respectively. Moreover, the improvement is more significant on the four larger datasets, i.e., BSet, TSet, MSet, and ESet, than that of the smaller dataset Tweet. These results demonstrate that TATM effectively improves topic quality by capturing highly coherent topics and it performs better on large-scale short texts than all the baselines.

Figure 4 plots the PMI scores of all the compared methods under different number of topics K. First, we can see that TATM outperforms the four baselines with the highest PMI scores on all datasets. Second, when K increases, all the compared methods show an upward trend except LDA and ATM. This result demonstrates that, for short texts, the topic models that with targeted aspects (i.e., TTM), global word co-occurrence (i.e., BTM and TATM), or self-aggregation (i.e., ASTM) are more effective, which is consistent with our discussion in Section 2.2. Third, comparing Subfigure (a) to Subfigures (c) to (e), we can see that TTM, which adopts a set of pre-given keywords to identify targeted aspects, cannot always perform better than the other baselines. It indicates that the simple word constraint is not suitable for identifying targeted aspects, especially on large-scale short texts.

Qualitative evaluation To conduct the qualitative evaluation, we investigate the completeness of captured topics on several specific targeted aspects, namely, we evaluate the different angles covered by these topics. We fix $K = 60$ and $N = 10$, and check the correctness of these words according to there relevance to the corresponding targeted aspect. The incorrect words are italicized and marked in red. Tables 4 and 5 show the results on the Tweet and BSet datasets, respectively.

Table 3 Performance comparison with four topic models of topic coherence on five datasets with a fixed number of topics $K = 70$

Method	LDA	ATM	BTM	TTM	ASTM	TATM
Tweet	0.2614	0.2625	0.3198	0.3053	0.2964	**0.3239**
Improvement	+ 0.06(23.91%)	+ 0.06(23.39%)	+ 0.01(1.28%)	+ 0.02(6.09%)	+ 0.03(9.28%)	N/A
BSet	0.2230	0.2326	0.2809	0.2459	0.2877	**0.2965**
Improvement	+ 0.07(32.96%)	0.06(27.47%)	+ 0.02(5.55%)	+ 0.05(20.58%)	+ 0.01(3.06%)	N/A
TSet	0.2137	0.2212	0.2644	0.2047	0.2581	**0.3049**
Improvement	+ 0.10(42.68%)	+ 0.08(37.84%)	+ 0.04(15.32%)	+ 0.10(48.95%)	+ 0.05(18.13%)	N/A
ESet	0.2091	0.2555	0.4007	0.2768	0.3663	**0.4372**
Improvement	+ 0.23(109.09%)	+ 0.18(71.12%)	+ 0.04(9.11%)	+ 0.16(57.95%)	+ 0.07(19.36%)	N/A
MSet	0.2066	0.2800	0.3272	0.2308	0.3712	**0.4357**
Improvement	+ 0.23(110.89%)	+ 0.16(55.16%)	+ 0.11(33.16%)	+ 0.20(88.78%)	+ 0.06(17.38%)	N/A

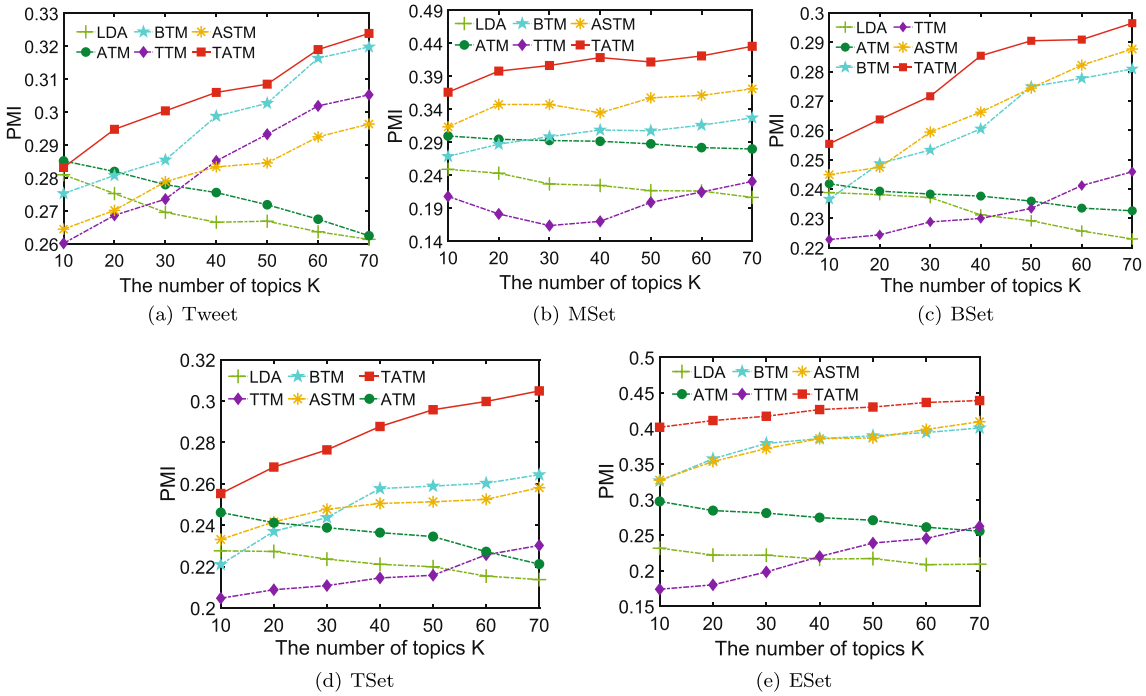


Fig. 4 Performance comparison with four topic models of topic coherence on five datasets **a** Tweet, **b** MSet, **c** BSet, **d** TSet, and **e** ESet under different number of topics K changing from 10 to 70 with a step of 10

Table 4 reports the topical words of a popular aspect *Superbowl* in the Tweet dataset. In addition to the results of this game, two other topics are captured. We refer them to as *commerce*, which discusses the advertisements of this game, and *show*, which is about the singer who sings the national anthem.

First, we can see that TATM captures both topics with the the fewest incorrect words, which means it can discover topics for a targeted aspect from as many angles as possible, namely, it achieves topic completeness. Second, all the compared methods capture the topic *commerce* except ATM. This is mainly because that ATM aggregates all the captured topics, resulting in the topics are general and confusing. Moreover, TTM confuses topic *commerce* with more incorrect words, such as "*Gabrielle*" and "*Giffords*", which indicate a senator of the Democratic Party. In contrast, TATM captures the most complete advertisers that are highly relevant to this topic (e.g., "*Pepsi*" (a well-known beverage) and "*Volkswagen*" (a popular car brand)). Third, the state-of-the-art method ASTM overlooks the topic *show*. Furthermore, the topic *commerce* captured by ASTM is also confused with some unrelated words, such as "*Oscars*" and "*brain*". The most comparable method is BTM. However, comparing topic *show* captured by BTM with that of TATM, we can see that TATM has a better result with the singer's information (e.g., "*Christina*"), the audiences' comments (e.g., "*mess*"), and it even explains the reason for the comments (e.g., "*mistake*").

Table 4 Topics of targeted aspect *Superbowl* in Tweet

Targeted aspect: Superbowl									
LDA		ATM	BTM		TTM		ASTM	TATM	
commerce	show	-	commerce	show	commerce	show	commerce	commerce	show
commercial	sing	news	superbowl	Christina	superbowl	Christina	superbowl	superbowl	Aguilera
superbowl	America	commercial	commercial	national	commercial	Aguilera	commercial	commercial	Christina
advertiser	Christina	*chicken*	advertiser	anthem	Doritos	mess	*brain*	Pepsi	anthem
KIA	video	superbowl	Doritos	Aguilera	advertiser	delay	*Oscars*	Volkswagen	national
car	trendsetter	anthem	Max	fail	*brain*	video	*buildup*	Doritos	mess
optimum	Aguilera	*fishing*	Volkswagen	*news*	*buildup*	star	advertiser	advertiser	fail
supe	anthem	Aguilera	YouTube	lyric	*Gabrielle*	trendsetter	KIA	Audi	lyric
opinion	*ad*	*recipe*	*best*	*ap*	*Giffords*	*traitor*	America	Mercedes	*repeat*
epic	*sounded*	national	watch	banner	car	*camp*	national	YouTube	superbowl
making	fail	*anthem*	*force*	*share*	YouTube	*start*	epic	*big*	mistake

Incorrect words are italicized and marked in red

Table 5 Topics of two targeted aspects *International trading* and *Enterprise* on BSet

Targeted aspect: International trading						Targeted aspect: Enterprise				
BTM		ASTM	TATM			BTM		ASTM	TATM	
banking	oil	banking	banking	oil	gold	computer	e-commerce	-	computer	e-commerce
ECB	crude	ECB	BNP	BP	price	Microsoft	Alibaba	Ebay	earnings	Alibaba
Paribas	BP	bad	ECB	spill	ounce	cloud	Amazon	paypal	Microsoft	camera
BNP	price	triggers	mortgage	crude	silver	sale	GoPro	*split*	CEO	Amazon
plan	trading	*key*	rates	gas	SPDR	drop	mobile	shreholders	*beat*	IPO
JPMorgan	supply	*test*	Paribas	coast	ETF	earnings	IPO	*Asian*	cloud	wall
Fannie	gas	market	home	refinery	forex	street	TV	*rejected*	surface	street
Deutsche	*report*	*downtown*	FHA	WTI	metal	wall	advertise	*asks*	sale	digital
post	spill	*sources*	Fannie	*report*	*review*	CEO	camera	*vote*	wall	*files*
fed	WTI	loan	sanctions	price	*news*	*help*	*users*	triggers	street	raise
mortgage	*Twitter*	DJIA	*post*	*ahead*	fund	*big*	*Twitter*	*making*	revenue	phone

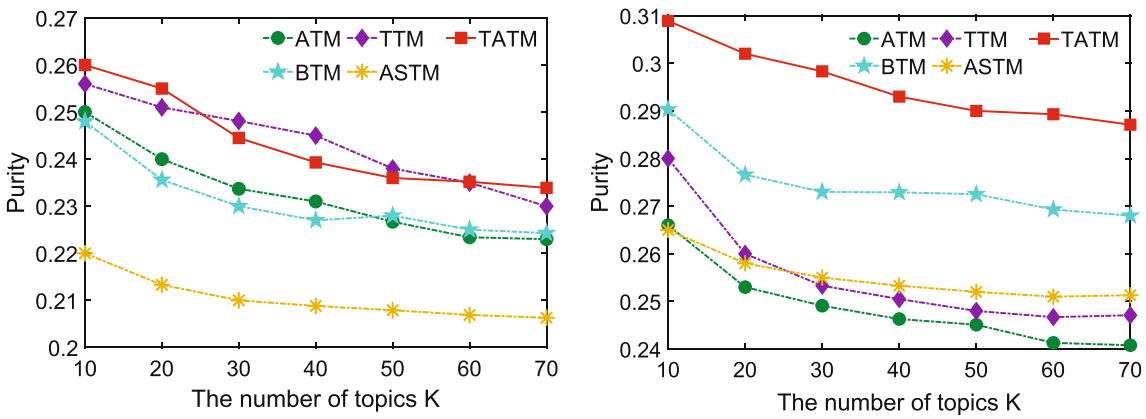
Incorrect words are italicized and marked in red

Table 5 reports the topical words of two business aspects *International trading* and *Enterprise* in BSet. Since the comprehensive comparison of one targeted aspect has been presented in Table 4, now we only compare TATM with ASTM and BTM, which have better performance than LDA, ATM, and TTM.

When the targeted aspect is *International trading*, there are three captured topics, which are referred to as *banking*, *oil*, and *gold*, respectively. First, we can see that TATM captures all topics with the fewest incorrect words. Second, all the three methods capture the topic *banking*. Specifically, BTM only groups many names of banks (e.g.,"*ECB*" (European Central Bank),"*BNP Paribas*" (Bank of Paris France)), which are synonyms but make this topic look confused and vague. In addition, ASTM introduces too many meaningless words (e.g., "*test*" and "*downtown*"). In contrast, our TATM model captures more specific words for this topic, such as "*FHA*" (Federal Housing Administration) and "*sanctions*". These specific words help target the news event that Fannie Mae sued the world's nine largest banks. Third, although the topic *oil* captured by BTM is comparable, BTM cannot capture the topic *gold*, which is about the international gold transactions.

When the targeted aspect is *Enterprise*, the results of TATM and BTM are similar, because they both capture the two popular topics *computer* and *e-commerce*, which are about computer enterprises and electronic commerce enterprises, respectively. Specifically, TATM covers less incorrect words than BTM. Furthermore, the topic captured by ASTM is not interpretability. This is mainly because its pre-trained knowledge highly lies in statistics of words, which makes the topic is not clear for a specific aspect.

Summary 2 TATM can effectively capture high-quality topics for targeted aspects compared to four baseline topic models. On the one hand, the quantitative evaluation demonstrates that TATM significantly improves topic homogeneity with highly coherent topics. On the other hand, the qualitative evaluation demonstrates that TATM can capture topics for a targeted aspect from as many angles as possible, which means it achieves topic completeness. In addition, the qualitative evaluation also demonstrates that for short texts, the mechanism of biterm (i.e., word pairs) adopted by BTM and TATM is more effective than the unigram mechanism adopted by the other baselines.



(a) Tweet (b) TSet

Fig. 5 Performance comparison with three baseline topic models of topic purity on two datasets **a** Tweet and **b** TSet

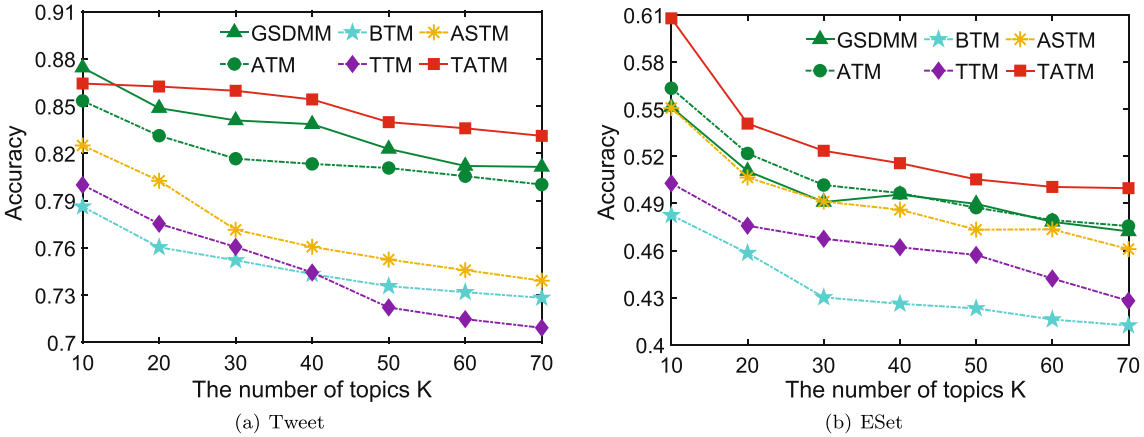


Fig. 6 Performance comparison with four baselines of the short text classification on two datasets **a** Tweet and **b** ESet

4.3.4 Topic purity (for Q4)

In this part, we answer **Q4** with the experimental result of Task 4, which evaluates the consistency between the captured topics and their corresponding targeted aspects. We use the purity metric and compare TATM with three baselines on two datasets. We fix the number of most probable words in each topic as $N = 10$ when K changes from 10 to 70 with a step of 10. Figure 5 plots the purity scores on (a) the Tweet dataset and (b) the TSet dataset, respectively. For all experiments in this part, the higher purity score indicates the better topic purity.

First, we can see that TATM outperforms all the compared baselines, especially on the TSet dataset, which means that the topics captured by TATM are as consistent with the corresponding targeted aspect as possible. Specifically, on average, TATM improves ATM, BTM, TTM, and ASTM by 7.99%, 6.45%, 15.73%, and 4.67%, respectively. Second, on the Tweet dataset, the performance of TTM is the most comparable to our TATM model. However, its performance quickly becomes poor on the TSet dataset as plotted in Fig. 5b. This is because that the pre-given keywords cannot completely summarize every angle of the corresponding targeted aspect, which aggravates the sparsity problem of short texts when the size of data increases. Third, when K increases, all the compared methods show downward trends. This can be explained with common sense that the more topics we discuss, the more likely it is to introduce unrelated words.

Summary 3 In general, our TATM model can effectively capture the topics that are highly consistent with each targeted aspect, which demonstrates the E-DMM process of TATM can gather similar words together as many as possible and introduce fewer noisy words than baselines. Moreover, compared to baselines, TATM maintains good topic purity when the size of data increases.

4.3.5 Short text classification (for Q5)

In this part, we answer **Q5** with the experimental results of Task 5, which evaluates the performance of short text classification of all the compared methods. To conduct this experiment, the datasets are randomly divided into two parts, i.e., the testing set and the training set, with a ratio of 3:7. This ratio is optimized with cross-validations in {1:9, 2:8, 3:7, 4:6, 5:5}, respectively. Figure 6 plots the results with the accuracy metric on (a) the Tweet dataset and (b) the ESet dataset, respectively.

First, we can see that TATM outperforms the four baselines on both datasets, including two state-of-the-art methods ATM and ASTM. Specifically, on average, TATM improves GSDMM, ATM, BTM, TTM, and ASTM by 3.75%, 4.25%, 13.94%, 17.33%, and 8.77%, respectively. Second, compared to GSDMM and BTM, which adopt the standard one-topic assumption, TATM performs better. This indicates that, for short text topic discovery, our one-aspect assumption is more

Table 6 Time complexity of TATM and four baseline topic models in an original form and a unified form

Method	Time complexity (original form)	Time complexity (unified form)
LDA	$O(N_{iter}DKl)$	$O(N_{iter}DKl)$
ATM	$O(N_{iter}DKl)$	$O(N_{iter}DKl)$
BTM	$O(N_{iter}DKB)$	$O(N_{iter}DKl(l-1)/2)$
TTM	$O(N_{iter}(Dl + (2+K)V))$	$O(N_{iter}(Dl + (2+K)Dl/x))$
ASTM	$O(N_{iter}(s + (DKl)/5))$	$O(N_{iter}DK(l(l-1)/2 + l/5))$
TATM	$O(N_{iter}D_tKB)$	$O(N_{iter}D_tKl(l-1)/2)$

Table 7 Performance comparison of running time (in seconds) after 1000 iterations for Gibbs Sampling

Method	Tweet	BSet	TSet	ESet	MSet
LDA	29.50	106.11	72.96	163.19	64.38
ATM	32.77	108.21	75.45	170.53	70.91
BTM	53.63	121.78	113.57	181.14	106.67
TTM	87.48	276.37	130.99	184.01	285.95
ASTM	95.54	223.29	181.47	258.22	186.68
TATM	10.29	55.11	48.31	72.07	40.43

reasonable than the one-topic assumption. Third, compared to the targeted aspects oriented method TTM, TATM has a better result when the size of data increases as plotted in Fig. 6b. It demonstrates that, for short texts, the mechanism of biterms (i.e., word pairs) is more effective than the mechanism of unigram that TTM adopts. In addition, the performance of ATM is the most comparable to our method, especially on the ESet dataset. This is mainly because that ATM aggregates all the captured topics by estimating their Jensen-Shannon divergence, which improves the topic coherence. However, this aggregation mechanism cannot distinguish the topics that are associated with different targeted aspects.

Summary 4 Overall, our TATM model can learn high-quality document-topic features for short text classification, especially when dealing with large-scale data. The significant improvement of TATM is benefits from its one-aspect assumption and the E-DMM process, which can gather similar words together as many as possible for each targeted aspect.

4.3.6 Time complexity (for Q6)

In this part, we answer **Q6** with the experimental results of Task 6. Table 6 presents the time complexity of different methods, and Table 7 presents the average running time.

Suppose the number of iterations for Gibbs Sampling is N_{iter}, the number of topics is K, the number of documents is D, the number of documents in a targeted aspect is D_t, and the average number of words in a document is l, the time complexity of all compared methods are summarized in Table 6. We analyze the time complexity in an original form and a unified form, respectively.

For TATM and BTM, which adopt the mechanism of biterm (i.e., word pair), we use an average value $l(l-1)/2$ to replace the number of extracted word pairs B. For TTM, which assigns a relevance status to each word in each iteration, we replace the size of vocabulary V with a ratio $1/x$ of Dl. This ratio is around 1/4 on Tweet dataset and 1/25 on four NewsTitle datasets. For ASTM, which samples some fragments for each document in each iteration, we replace the number of fragments s with an average value $l(l-1)/2$. Consequently, the time complexity reduced by TATM can be approximated and the improvement depends on l and D_t. We set $l = 9$ as an average value as shown in Table 2. In summary, the time complexity of total Gibbs Sampling iterations of LDA, ATM, BTM, TTM, and ASTM reduced by TATM is approximately $O(N_{iter}K(9D - 36D_t))$, $O(N_{iter}K(9D - 36D_t))$, $O(36N_{iter}K(D - D_t))$, $O(N_{iter}(15D + 3KD - 36K))$, and $O(N_{iter}(36K(D - D_t) + 9/5))$, respectively.

To validate the above theoretical analysis, we further compare the running time of TATM and four baseline topic models on all datasets. Table 7 presents the average values of each method after 500 Gibbs Sampling iterations.

From Table 7, first, we can see that TTM and ASTM spend the most time since TTM needs to update the keywords in each iteration and ASTM needs to sample attentional fragments in each iteration. TTM and ASTM enhance their performance of topic discovering by adopting the additional operations but sacrifice a lot of time performance. Second, the results of LDA and ATM are comparable, since LDA and ATM sacrifice topic homogeneity by performing a full analysis. Third, comparing TATM with BTM, which extracts the co-occurrence word pairs, TATM has a better result as it reduces the running time by filtering some documents that are not included in a targeted aspect as noise.

Summary Overall, our TATM model is more efficient for discovering topics on targeted aspects, even compared to state-of-the-art methods. Specifically, the mechanism of word pairs can help reduce the running time on a target-level modeling process, which also helps achieve topic completeness.

5 Conclusion

Motivated by the observation that people always tend to find focused topics on some special aspects (or events), in this paper, we have proposed a novel targeted aspects oriented topic model (TATM) for short texts. On the one hand, for achieving topic homogeneity, TATM identifies targeted aspects via an enhanced Dirichlet Multinomial Mixture (E-DMM) process, which groups similar words as many as possible. On the other hand, for achieving topic completeness, TATM performs a target-level topic modeling by adopting the mechanism of biterm (i.e., word pair). The extensive experiments conducted on five real-world short text datasets demonstrate that our TATM model is superior to the state-of-the-art baselines in terms of convergence speed, topic quality, topic purity, short text classification, and time efficiency.

In future work, we will consider improving our model with self-adaptive sampling process to make it fits more types of data. We are also interested in employing optimization techniques for parameter settings and applying the model to novel

scenarios, such as the detection of aggregation for fake research papers..

Acknowledgements This work has been supported by the National Key Research and Development Program of China under grant 2016YFB1000901, the National Natural Science Foundation of China under grant 91746209 and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education of China under grant IRT17R32.

References

1. Ahuja A, Wei W, Carley KM (2016) Microblog sentiment topic model. In: Proceedings of the 2016 IEEE 16th international conference on data mining workshops (ICDMW), pp 1031–1038
2. Beykikhoshk A, Arandjelović O, Phung D, Venkatesh S (2018) Discovering topic structures of a temporally evolving document corpus. Knowl Inf Syst 55(3):599–632
3. Blair S J, Bi Y, Mulvenna M D (2020) Aggregated topic models for increasing social media topic coherence. Appl Intell 50(1):138–156
4. Blei D M, Ng A Y, Jordan M I (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
5. Chang J, Gerrish S, Wang C, Boyd-Graber J L, Blei D M (2009) Reading tea leaves: How humans interpret topic models. In: Proceedings of the 20th annual conference on neural information processing systems, NIPS 2009, pp 288–296
6. Chen W, Wang J, Zhang Y, Yan H, Li X (2015) User based aggregation for biterm topic model. In: Proceedings of the 53rd annual meeting of the association for computational linguistics, ACL 2015, pp 489–494
7. Cheng X, Yan X, Lan Y, Guo J (2014) Btm: Topic modeling over short texts. IEEE Trans Knowl Data Eng 26(12):2928–2941
8. Esposito M, Damiano E, Minutolo A, De Pietro G, Fujita H (2020) Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. Inform Sci 514: 88–105
9. Finegan-Dollak C, Coke R, Zhang R, Ye X, Radev D (2016) Effects of creativity and cluster tightness on short text clustering performance. In: Proceedings of the 54th annual meeting of the association for computational linguistics, ACL 2016, pp 654–665
10. Griffiths T L, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235
11. Hayashi T, Fujita H (2019) Word embeddings-based sentence-level sentiment analysis considering word importance. Acta Polytechnica Hungarica 16(7):152–52
12. He J, Li L, Wu X (2017) A self-adaptive sliding window based topic model for non-uniform texts. In: Proceedings of the 2017 IEEE international conference on data mining, ICDM 2017, pp 147–156
13. Hisano R (2019) Learning topic models by neighborhood aggregation. In: Proceedings of the 28th international joint conference on artificial intelligence, IJCAI 2019, pp 2498–2505
14. Huang R, Yu G, Wang Z, Zhang J, Shi L (2012) Dirichlet process mixture model for document clustering with feature partition. IEEE Trans Knowl Data Eng 25(8):1748–1759
15. Ibrahim R, Elbagoury A, Kamel M S, Karray F (2018) Tools and approaches for topic detection from twitter streams: Survey. Knowl Inf Syst 54(3):511–539
16. Jain AK (2008) Data clustering: 50 years beyond k-means. In: Proceedings of joint European conference on machine learning and knowledge discovery in databases, pp 3–4
17. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media?. In: Proceedings of the 19th international conference on World Wide Web, WWW 2010, pp 591–600
18. Li X, Li C, Chi J, Ouyang J (2018) Short text topic modeling by exploring original documents. Knowl Inf Syst 56(2):443–462
19. Mikolov T, Sutskever I, Chen K, Corrado G S, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th annual conference on neural information processing systems, NIPS 2013, pp 3111–3119
20. Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. In: Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics, pp 100–108
21. Nigam K, McCallum A K, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using em. Mach Learn 39(2-3):103–134
22. Pedrosa G, Pita M, Bicalho P, Lacerda A, Pappa G L (2016) Topic modeling for short texts with co-occurrence frequency-based expansion. In: Proceddings of the 5th Brazilian conference on intelligent systems, BRACIS 2016, pp 277–282
23. Qiang J, Chen P, Wang T, Wu X (2017) Topic modeling over short texts by incorporating word embeddings. In: Proceedings in the 21st Pacific-Asia conference on knowledge discovery and data mining, PAKDD 2017, pp 363–374
24. Quan X, Kit C, Ge Y, Pan S J (2015) Short and sparse text topic modeling via self-aggregation. In: Proceedings of the 24th international joint conference on artificial intelligence, IJCAI 2015, pp 2270–2276
25. Rahman M M, Wang H (2016) Hidden topic sentiment model. In: Proceedings of the 25th international conference on World Wide Web, WWW 2016, pp 155–165
26. Shi B, Lam W, Jameel S, Schockaert S, Lai K P (2017) Jointly learning word embeddings and latent topics. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2017, pp 375–384
27. Teh Y W, Newman D, Welling M (2007) A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In: Proceedings of the 20th annual conference on neural information processing systems, NIPS 2006, pp 1353–1360
28. Wang H, Lu Y, Zhai C (2011) Latent aspect rating analysis without aspect keyword supervision. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, SIGKDD 2011, pp 618–626
29. Wang J, Chen L, Qin L, Wu X (2018) Astm: An attentional segmentation based topic model for short texts. In: Proceedings of the 2018 IEEE international conference on data mining, ICDM 2018, pp 577–586
30. Wang S, Chen Z, Fei G, Liu B, Emery S (2016) Targeted topic modeling for focused analysis. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, SIGKDD 2016, pp 1235–1244
31. Wang Y, Wang M, Fujita H (2019) Word sense disambiguation: A comprehensive knowledge exploitation framework. Knowl-Based Syst P 190:105030
32. Yan X, Guo J, Lan Y, Xu J, Cheng X (2015) A probabilistic model for bursty topic discovery in microblogs. In: Proceedings ot the 29th AAAI conference on artificial intelligence, AAAI 2015, pp 353–359
33. Yin J, Wang J (2014) A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, SIGKDD 2014, pp 233–242
34. Yin J, Wang J (2016) A text clustering algorithm using an online clustering scheme for initialization. In: Proceedings of the 22nd

ACM SIGKDD international conference on knowledge discovery and data mining, SIGKDD 2016, pp 1995–2004
35. Yu G, Huang R, Wang Z (2010) Document clustering via dirichlet process mixture model with feature selection. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, SIGKDD 2010, pp 763–772
36. Zhang Y, Song D, Zhang P, Li X, Wang P (2019) A quantum-inspired sentiment representation model for twitter sentiment analysis. Appl Intell 49(8):3093–3108
37. Zhao W X, Jiang J, Weng J, He J, Lim E P, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European conference on information retrieval, ECIR 2011, pp 338–349
38. Zhou X, Ouyang J, Li X (2018) Two time-efficient gibbs sampling inference algorithms for biterm topic model. Appl Intell 48(3):730–754
39. Zuo Y, Zhao J, Xu K (2016) Word network topic model: A simple but general solution for short and imbalanced texts. Knowl Inf Syst 48(2):379–398

Jin He is currently pursuing the Ph.D. degree with the School of Computer Science and Information Engineering at Hefei University of Technology, Hefei, China. She received the B.E. degree from Anhui University of Finance and Economics, Bengbu, China, in 2015. Her current research interests include data mining and topic modeling.

Lei Li received the B.S. degree from Jilin University, China, in 2004, the M.S. degree from the Memorial University of NewFoundland, St. John's, NL, USA, in 2006, and the Ph.D. degree from Macquarie University, Sydney, NSW, Australia, in 2012. His is an Associate Professor with the School of Computer Science and Information Engineering at Hefei University of Technology, Hefei, China. His current research interests including graph computing, social computing, and data mining.

Yan Wang received the B.E., M.E., and Ph.D. degrees in Computer Science and Technology from the Harbin Institute of Technology (HIT), China, in 1988, 1991, and 1996, respectively. He is a professor in the Department of Computing, Macquarie University, Sydney, Australia. His current research interests include trust computing, social computing, service computing and recommender systems. He is a senior member of the IEEE.

Xindong Wu received the Ph.D. degrees in artificial intelligence from the University of Edinburgh, Edinburgh, U.K.. He is a professor of computer science at the University of Louisiana at Lafayette, USA. His current research interests include data mining, knowledge based systems, and Web information exploration. He is the Steering Committee chair of IEEE International Conference on Data Mining (ICDM). He is the editor-in-chief of Knowledge and Information Systems (KAIS) and ACM Transactions on Knowledge Discovery from Data (TKDD). He is a fellow of IEEE and the AAAS.