

Article

BiTTM: A Core Biterms-Based Topic Model for Targeted Analysis

Jiamiao Wang ¹, Ling Chen ², Lei Li ³  and Xindong Wu ^{4,*}

¹ School of Information and Engineering, Sichuan Tourism University, Chengdu 610100, China; wjmzjx@163.com

² Centre for Artificial Intelligence, University of Technology Sydney, P.O. Box 123, Broadway, NSW 2007, Australia; Ling.Chen@uts.edu.au

³ School of Computer Science and Information Engineering, Hefei University of Technology, Tunxi Road, Hefei 230009, China; lilei@hfut.edu.cn

⁴ Mininglamp Academy of Sciences, Mininglamp Technology, Beijing 100084, China

* Correspondence: xwu@hfut.edu.cn

Abstract: While most of the existing topic models perform a *full analysis* on a set of documents to discover all topics, it is noticed recently that in many situations users are interested in fine-grained topics related to some specific aspects only. As a result, *targeted analysis* (or *focused analysis*) has been proposed to address this problem. Given a corpus of documents from a broad area, targeted analysis discovers only topics related with user-interested aspects that are expressed by a set of user-provided query keywords. Existing approaches for targeted analysis suffer from problems such as topic loss and topic suppression because of their inherent assumptions and strategies. Moreover, existing approaches are not designed to address computation efficiency, while targeted analysis is supposed to provide responses to user queries as soon as possible. In this paper, we propose a *core BiTerms-based Topic Model* (BiTTM). By modelling topics from *core biterms* that are potentially relevant to the target query, on one hand, BiTTM captures the context information across documents to alleviate the problem of topic loss or suppression; on the other hand, our proposed model enables the efficient modelling of topics related to specific aspects. Our experiments on nine real-world datasets demonstrate BiTTM outperforms existing approaches in terms of both effectiveness and efficiency.

Keywords: AI; text analysis; topic model; biterm; content analysis; targeted modeling



Citation: Wang J.; Chen L.; Li L.; Wu X. BiTTM: A Core Biterms-Based Topic Model for Targeted Analysis. *Appl. Sci.* **2021**, *11*, 10162. <https://doi.org/10.3390/app112110162>

Academic Editor: Chang Young Song

Received: 9 October 2021

Accepted: 26 October 2021

Published: 29 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Topic modelling as unsupervised learning has become a prevalent text mining tool for discovery of hidden semantic structures in a text body. Given a collection of documents, most of the existing topic models perform a full analysis to discover *all* topics occurring in the corpus. However, it was recently noticed [1] that in many situations users are interested in focused topics related to some specific aspects only. For example, given a set of Amazon product reviews, a user might be interested only in bedding products. A conventional topic model performing full analysis will identify all topics from the entire corpus such as “furniture”, “food” and “clothing”. Although the topic of “furniture” is related to the user interested aspect of “bedding products”, it is too coarse as the user might be more interested in fine-grained topics like “bed frames” and “mattress”. As a result, targeted (or focused) analysis is proposed by Wang et al. [1] to discover topics relevant to targeted aspects only. Particularly, given a corpus of documents from a broad domain and a set of user-provided keywords representing user-interested aspects, targeted analysis aims to discover topics related with the queried aspects only.

Methods for targeted analysis can be generally categorised into two groups: (1) conventional topic models incorporating filtering strategies and (2) specialised topic models. However, methods of both categories suffer from problems such as *topic loss* and *topic suppression*, because of the limitations of their respective assumptions and strategies.

For algorithms in the first group, both pre-filtering and post-filtering strategies can be adopted to empower full-analysis topic models to find topics related to queried aspects. Basically, the pre-filtering strategy retains only documents containing the query keywords and extracts topics from the retained “partial data”. The quality of the discovered topics thus heavily depends on user-supplied query keywords. If the keywords are not appropriate or comprehensive enough, many relevant documents will be filtered, which incurs a significant *topic loss*. For example, if a user provides “bath” as a query keyword, documents without the keyword but containing the synonyms like “shower” and similar words like “bathtub” will be filtered although such documents are actually relevant. Consequently, there is a great possibility to lose topics if modelling from the retained partial data. A post-filtering strategy applies conventional topic models to identify first all topics in the corpus and then filter the topics that do not contain the query keywords in the results. However, as analysed in [1], such a strategy may result in *topic suppression* when the query keywords are infrequent in the database. *Topic suppression* means that topics related to the user interested aspect are suppressed by general topics.

For algorithms in the second group, TTM [1] is the first and the state-of-the-art. TTM is a sparse topic model designed to directly mine focused topics based on user-provided query keywords. TTM simulates two topic-word distributions: ϕ^r for relevant topics and ϕ^{ir} for irrelevant topics. It considers documents at the sentence level and introduces a variable r to indicate the status of a sentence (e.g., relevant or irrelevant). Words are then sampled from ϕ^r or ϕ^{ir} according to the sentence status. Although TTM can accomplish the targeted analysis to a certain extent, the effectiveness of TTM is handicapped by its scheme of processing at the sentence level and its assumption that each sentence focuses on only one topic. By considering sentences individually and separately, topic information between consecutive sentences may be lost, which results in inferior topic qualities and possible topic loss. By assuming that each sentence is related with only one aspect, it is very likely for TTM to mistakenly assign relevance status for sentences related with multiple topics, which is often the case for long sentences. The wrong assignment of sentence statuses will in turn lead to possible missing of meaningful topics.

A common challenge faced by algorithms of both categories is the computation efficiency, while full analysis of topics is largely performed offline, targeted analysis is more likely an online module that is supposed to respond to user queries as soon as possible. However, existing algorithms for targeted analysis, especially the post-filtering strategy and the specialised topic models, are not devised to address this issue. The pre-filtering strategy may gain efficiency by modelling topics from a reduced set of “partial data”, but it achieves this at the cost of losing important topics.

To address the aforementioned issues, we propose a novel *Core BiTerm-based Topic Model* (BiTTM) for targeted analysis, which directly models fine-grained topics related to the queried aspect from a set of *core biterms*. Biterm, proposed in BTM [2], is a word-pair consisting of two different words that appear together in a fixed-size window and represent co-occurrence information. Improving biterms, we introduce *core biterms* as a set of selected biterms that have strong connections with query keywords. By modelling topics from the set of core biterms, BiTTM is expected to achieve better performance than existing specialised topic models in terms of the following aspects:

1. The existing specialised topic models for targeted analysis (i.e., TTM and AP-SUM [3]) process at either the sentence level or the word level so that the semantic information between consecutive sentences will be lost. In contrast, since a biterm may consist of two words coming from two successive sentences, information across the whole document can be captured by BiTTM to alleviate the issue of losing topics.
2. The TTM model samples relevance status at the sentence level which may be too coarse. When a sentence is related to multiple topics, it would be difficult to infer the relevance status of the sentence as a binary value. In contrast, the APSUM model [3] samples relevance status for individual words which may be too specific, because it cannot handle phrases that make sense when multiple words are considered together. Biterms, as

a scheme in-between sentences and words, are expected to achieve more accurate inference of relevance status.

3. Existing specialised topic models do not have any finesse to accelerate the calculation without significant semantic information losing. Instead, BiTTM introduces a heuristic preprocessing based on core biterms for speeding topic modelling while alleviating information loss, which makes it a more pragmatic solution for targeted analysis according to user queries.

To comprehensively evaluate the performance of BiTTM, extensive experiments have been conducted on real-world datasets including short texts, medium texts and long texts. Moreover, we select a large number of targets with different word and document frequencies to explore the adaptability of BiTTM to various types of queries. The experimental results show that (1) BiTTM improves the quality of topics, alleviates topic losing, and outperforms baselines especially for query keywords of low frequencies; (2) the time cost of BiTTM is most outstanding and stable compared to those of the baselines, which demonstrates the high applicability of BiTTM on datasets with different characteristics.

The remainder of this paper is organised as follows. Prior research and related works are reviewed in Section 2. We provide technical details of BiTTM in Section 3, and discuss the experimental results in Section 4. Finally, Section 5 closes this paper with some conclusive remarks.

2. Related Work

In this section, we introduce works related to our research in three parts. Firstly, we review existing specialised topic models for targeted topic analysis. Secondly, we describe the model of BTM that introduces the concept of biterms for topic modelling. Thirdly, we discuss other topic models relevant to our proposed BiTTM.

2.1. Targeted Topic Models

Specialised topic models for targeted analysis are still rarely seen, which are mainly used for information retrieval [4], abstract extraction [3,5] and opinion mining [1,6]. TTM [1] and APSUM [3] are the two most representative models.

Wang et al. first study the problem to detect relevant and user-concerned topics from a given dataset [1] and propose the model TTM as illustrated in Figure 1. The main idea of TTM is to introduce a relevance variable r to indicate whether a sentence is related with a specified aspect. The variable r determines whether each word in a sentence is generated by a related topic or an irrelevant topic. Moreover, the relevant topic-word distribution φ^r is sparse because the number of words related to the target is usually less than that of the irrelevant words.

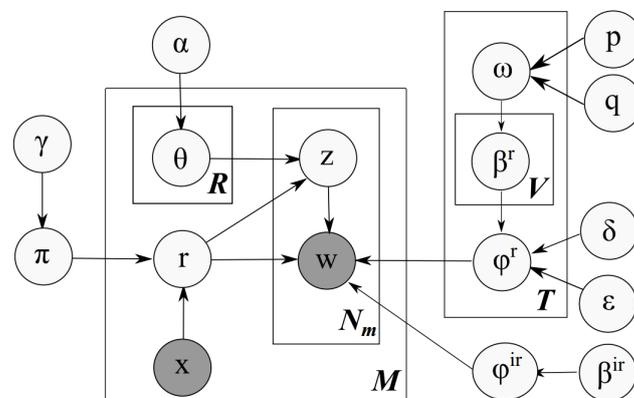


Figure 1. The graphical model of TTM.

The steps of generative process are illustrated as follows:

1. Draw $\varphi^{ir} \sim \text{Dirichlet}(\beta^{ir})$

2. For each relevant topic $k \in \{1, 2, \dots, T\}$
 - (a) Draw $k \sim \text{Beta}(p, q)$.
 - (b) For each word $v \in \{1, 2, \dots, V\}$
 - i. Draw $\beta_{t,v}^r \sim \text{Bernoulli}(\omega_t)$.
 - (c) Draw $\varphi_t^r \sim \text{Dirichlet}(\beta_t^r \delta + \epsilon)$.
3. For each document $m \in \{1, 2, \dots, M\}$
 - (a) Draw $\pi_m \sim \text{Beta}(\gamma)$.
 - (b) raw relevance status r based on keyword indicator x and $\text{Bernoulli}(\pi_m)$.
 - (c) If the document is relevant
 - i. Draw $z \sim \text{Multinomial}(\theta^r)$.
 - ii. Draw $w_i \sim \text{Multinomial}(\varphi_z^r)$
 - (d) If the document is irrelevant
 - i. Draw $w_i \sim \text{Multinomial}(\varphi^{ir})$.

Therefore, TTM considers the status r at the sentence-level. It is difficult to determine whether a sentence is related to the target when a sentence contains multiple topics. The wrong assignment of sentence status will negatively affect the quality of topics.

APSUM [3] is a generative aspect summarisation model designed for fine-grained summaries of online reviews. Compared with TTM, APSUM is different in terms of the following two aspects. Firstly, while TTM models the relevance at the sentence level, APSUM considers at the word level. As discussed in Section 1, the former might be too coarse to determine the relevance status for sentences appropriately; the latter is not able to handle phrases where it makes sense only when multiple words are considered together. Secondly, APSUM introduces an additional component called *document aggregator* to mitigate the issue of aspect sparsity, which refers to the circumstances where there are not enough text data related with specific aspects. The basic idea is to cluster similar documents through document aggregator and sample topics for documents at the document aggregator level.

Essentially, both TTM and APSUM try to identify potentially related words that can serve as bridges to link relevant documents, especially those without containing query keywords. TTM searches such related words from sentences containing query keywords; APSUM attempts from semantically similar documents through document aggregator. However, both models fail to capture the semantic information between neighbouring sentences, which does exist in natural language [7].

2.2. BTM

BTM [2] is a topic model for short texts. To alleviate the problem of insufficient information with short texts, this model extracts topics by modelling from *biterms* representing word co-occurrences. A biterm is an unordered word pair consisting of two different words in a fixed-length text window. BTM replaces documents with a biterm set that can reveal the correlation between words in depth.

The graphical model of BTM is depicted in Figure 2 with a generative process as follows:

1. Draw $\theta \sim \text{Dirichlet}(\alpha)$
2. For each topic $k \in \{1, 2, \dots, K\}$
 - (a) Draw $\phi_k \sim \text{Dirichlet}(\beta)$.
3. For each biterm $b_i \in B$
 - (a) Draw $z_i \sim \text{Multinomial}(\theta)$
 - (b) Draw $w_{i,1}, w_{i,2} \sim \text{Multinomial}(\phi_{z_i})$

BTM is designed for full analysis on short texts, while the concept of biterms is used by BTM to seize word occurrences to alleviate data sparsity, we borrow the idea in our targeted analysis model to capture words closely related with the query keywords provided by users.

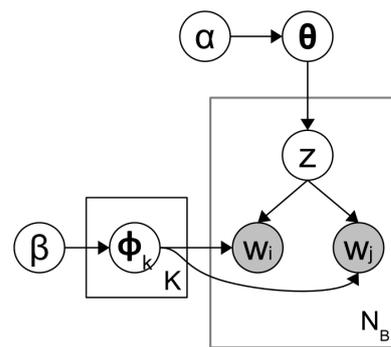


Figure 2. The graphical model of BTM.

2.3. Other Topic Models

Topic models have been widely studied and used in different applications, such as analysing public opinions and trends [8–12], providing personalised user services [13–15] and news tracking [16–18]. Among a wide range of existing topic models, in this subsection, we discuss two types of topic models that are related with our BiTTM.

Firstly, since we consider user queried aspects, our model is linked with topic models considering user information (e.g., user profiles and behaviours). In the literature, there are many topic models that take into account user information to obtain in-depth analysis [19–24]. For example, Viet et al. exploit users' browsing histories to propose a keyword-topic model [19] for contextual advertising. Kalyanam et al. [20] simultaneously consider textual data and user behaviours, such as forwarding and commenting, to explore the evolution of topics. Sordo et al. [21] consider the topological changes of users' co-authorship network to identify groups of researchers. Although these models can incorporate user information into topic analysis, they cannot extract fine-grained topics related with user-interested specific aspects.

Secondly, BiTTM is also connected with sparse topic models [25–32]. The notable feature of these models is the consideration of *distribution skewness* that can be divided into two categories. Firstly, a document is related with only a few topics among all topics available in the data set. Secondly, a topic involves only a small part of the dictionary. A lot of sparse topic models have been devised based on the two types of distribution skewness. For example, Williamson et al. [30] and Chen et al. [28] address the document skewness, while Wang et al. [33] take into account the topic skewness. Moreover, the method called “dual-sparse topic model” [25] implements both types of skewness simultaneously. Generally speaking, the sparsity is addressed by incorporating the “Spike and Slab” priors: the “Spike” is used to control the selections of words; and the “Slab” is used to smooth distributions to avoid ill-defined distributions where some words never appear. Our BiTTM also considers both document skewness and topic skewness through the spike and slab. Differently, we address the sparsity by taking into account user-interested aspects at the same time.

3. BiTTM

In this section, we describe BiTTM for efficient topic analysis of targeted aspects. In Section 3.1, we introduce the concept of core biterms and the process to generate core biterms. Sections 3.2 and 3.3 discuss the generative process and inference of BiTTM, respectively.

3.1. Core Biterms

Considering the user-specified aspect usually involves only part of the data, we believe data preprocessing is an indispensable step for efficient targeted analysis. However, existing specialised topic models perform directly on the entire dataset ignoring the efficiency issue. Existing methods incorporating pre-filtering strategies, as discussed before, achieve certain efficiency by modelling topics from a reduced data set; nevertheless, the reduced data set may lose relevant documents if the targets are not expressed appropriately or

comprehensively. For example, Table 1 enumerates three situations where query keywords may easily be incomplete, resulting in possible loss of relevant documents and topics.

Table 1. Three examples where targeted aspect can be expressed by multiple different keywords.

Example 1	Synonyms	Two candidate query keywords: “bath” and “shower”.
Example 2	Domain restriction	Two candidate query keywords, “crib” and “bed”, in the Amazon review dataset <i>Baby</i> .
Example 3	Event Description	Two candidate query keywords, “mistake” and “osars-fail”, in the Twitter dataset <i>Oscars</i> .

1. Synonyms. For example, if the supplied query keyword is “bath”, relevant documents containing words representing similar semantics, such as “shower”, may be missed.
2. Words referring to the same targeted aspect in a particular domain. For example, when the domain is confined to Amazon reviews of baby products, the keywords “crib” and “bed” represent the same aspect, although they are not exactly synonyms.
3. Words describing the same event. Users often use diverse words to refer to the same event, especially in social networks. For example, considering the Twitter dataset of Oscars, both “mistake” and “osarsfail” are used to describe the event of a wrong envelope for the Best Picture Award.

To address the aforementioned issues, we propose an efficient data preprocessing method based on *core biterms*.

As introduced in BTM [2], a biterm consists of any two distinct words in a fixed-length window so that it captures the co-occurrence information in the document. As the window may span two or more sentences, the semantic information between consecutive sentences can be captured. Compared with TTM and APSUM, processing at the level of biterms addresses potential loss of information between successive sentences. Therefore, we consider biterms as the base unit of our preprocessing.

To handle the situations exemplified in Table 1, we consider to use “core words” to complement query keywords so that relevant documents that do not explicitly contain query keywords can be considered. Intuitively, if core words represent the same aspect indicated by query keywords, they should appear together with query keywords very often. Hence, we first extract “core words” that frequently co-occur with query keywords from biterms, and then extract frequent biterms containing core words as “core biterms”. The algorithm is illustrated in Algorithm 1, which can be summarised in three steps as follows:

Step 1: Calculate the desired size of the set of core words, scw , and rank all biterms $\in B_{all}$ in descending order according of frequency (Lines 1–2).

Step 2: Acquire core words from top frequent biterms containing target, and then calculate the average frequency of biterms containing core words as $threshold$ (Lines 3–15).

Step 3: Select core biterms according to two conditions. Firstly, the biterm has at least one core word. Secondly, the frequency of the biterm has to be greater than $threshold$ (Lines 16–20).

We will then model targeted topics from the generated core biterms, which yields a threefold benefit as follows: (1) the context information between neighbouring sentences is preserved; (2) sampling relevance status based on biterms is more accurate; and (3) modelling topics from core biterms is more efficient.

Algorithm 1: Preprocessing based on biterns

Input: size of dictionary W , biterns B_{all}
Output: biterns B_{core} with semantic links to the target

```

1  $scw = W * 2.5\%$ ;
2  $B_{all} = sort(B_{all})$ ;
3 for bitern  $b_i(w_{i1}, w_{i2}) \in B_{all}$  do
4   if  $coreWords.size \geq scw$  then
5     break;
6   if  $w_{i1} == target || w_{i2} == target$  then
7      $coreWords.add(w_{i1})$ ;
8      $coreWords.add(w_{i2})$ ;
9  $threshold = 0$ ;
10  $sum = 0$ ;
11 for bitern  $b_i(w_{i1}, w_{i2}) \in B_{all}$  do
12   if  $coreWords.contains(w_{i1}) || coreWords.contains(w_{i2})$  then
13      $threshold+ = b_i.freq$ ;
14      $sum ++$ ;
15  $threshold = Math.ceil(threshold/sum)$ ;
16 for bitern  $b_i(w_{i1}, w_{i2}) \in B_{all}$  do
17   if  $coreWords.contains(w_{i1}) || coreWords.contains(w_{i2})$  then
18     if  $b_i.freq > threshold$  then
19        $B_{core}.add(b_i)$ ;
20 return  $B_{core}$ 

```

3.2. Model Description & Generative Process

In this subsection, we describe the model and the generative process of BiTTM. Table 2 lists the notations used in this paper.

The generative process is as follows:

1. Draw $\theta \sim Dirichlet(\alpha)$.
2. Draw $\phi^{ir} \sim Dirichlet(\beta^{ir})$.
3. For each target-relevant topic $k \in \{1, 2, \dots, K\}$
 - (a) Draw $k \sim Beta(p, q)$.
 - (b) For each word $w \in \{1, 2, \dots, W\}$
 - i. Draw $\beta_{k,w}^r \sim Bernoulli(\omega_k)$.
 - (c) Draw $\phi_k^r \sim Dirichlet(\beta_k^r \delta + \epsilon)$.
4. For each bitern $b_i(w_{i1}, w_{i2}) \in B$
 - (a) Draw $\pi_b \sim Beta(\gamma)$.
 - (b) Compute r based on x and $Bernoulli(\pi_b)$.
 - (c) If b_i is relevant to the target
 - i. Draw $z_1, z_2 \sim Multinomial(\theta)$.
 - ii. Draw $w_{i,1} \sim Multinomial(\phi_{z_1}^r)$, and
 - iii. Draw $w_{i,2} \sim Multinomial(\phi_{z_2}^r)$.
 - (d) If b_i is irrelevant
 - i. Draw $w_{i,1}, w_{i,2} \sim Multinomial(\phi^{ir})$.

Graphical representation of BiTTM is shown in Figure 3. Following the above procedure, the generative process can be summarised into three parts. Firstly, we draw two global parameters θ and ϕ^{ir} . The former is a topic distribution which models on the entire corpus instead of one document, and the latter is a topic-word distribution of irrelevant topic. In other words, two words in an irrelevant biterns are drawn from only one irrele-

vant topic. Secondly, ϕ_k^r is drawn for each target-relevant topic $k \in \{1, 2, \dots, K\}$. Please note that two smoothing parameters, smoothing prior δ and weak smoothing prior ϵ , are used for dual-sparsity [25]. Thirdly, status r of b_i is determined by both target indicator x and $Bernoulli(\pi_b)$. According to two different types of status, relevant or irrelevant, we draw a word from ϕ^r or ϕ^{ir} .

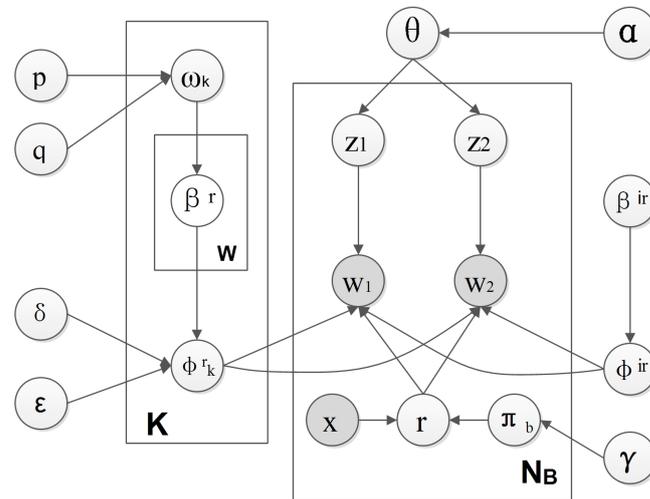


Figure 3. The graphical model of BiTTM.

Table 2. Symbol description.

Notation	Meaning
B	the set of (core) biterms
W	the set of words
D	the set of documents
π_b	the bernoulli distribution over biterm b
$\gamma, \beta^{ir}, \alpha$	beta prior of π_b , Dirichlet prior of ϕ^{ir}, θ
ϕ^{ir}	topic-word distribution over the irrelevant topic
ϕ_k^r	topic-word distribution over the k th relevant topic
p, q	beta prior of ω
δ, ϵ	word smoothing prior, weak word smoothing prior
ω_k	bernoulli distribution of word selector β_k^r
x, w, z, r	target indicator, word, topic, status
$\beta_{w k}^r, \beta_{* k}^r$	word selector of word w under topic k , the sum of word selector $\beta_{w k}^r$.
β_w^r, β_*^r	word selector of word w , the sum of word selector β_w^r
$\beta_{-w,*}^r$	the sum of word selector except the word w
$\beta_{-w,* k}^r$	the sum of word selector under topic k except the word w
$n_{-i,w}^r, n_{-i,w}^{ir}$	the number of times word w is relevant (irrelevant) excluding biterm b_i
n_{-i}^r, n_{-i}^{ir}	the number of relevant (irrelevant) biterms excluding biterm b_i .
$w_{i,m}$	the m th word in biterm b_i , where $m = 1, 2$
$n_{w_{i,m} k}^{r,-im}$	the number of times that word $w_{i,m}$ assigned to topic k under relevant status excluding $w_{i,m}$
$n_{* k}^{r,-im}$	the total number of words assigned to topic k under relevant status excluding $w_{i,m}$
$n_{w_{i,m} k}^{ir,-i}$	the number of times that word $w_{i,m}$ assigned to topic k under irrelevant status excluding biterm b_i
$n_{* k}^{ir,-i}$	the total number of words assigned to topic k under irrelevant status excluding biterm b_i
n_w^r	the number of times that word w is relevant
$n_{-w,*}^r$	the total number of words that are relevant excluding word w

Different from the generative process of BTM, BiTTM draws two topics for a relevant biterm and each word in the biterm may be assigned a different topic. The reason why we choose this strategy is that it is inappropriate to assume that the two words in a biterm share the same topic for targeted analysis, while for BTM, it is probably sufficient to draw one topic for a biterm as it is a full-analysis model which aims to mine coarse-grained topics.

Here is an example to elaborate the difference between full-analysis and targeted analysis. When dealing with the biterms $b_1(battery, larger)$ and $b_2(lens, larger)$, BTM is prone to assign the same topic to the two biterms because of the shared word “larger”. This allocation might be fine for full analysis since it does not pursue fine-grained topics so that it is not necessary to distinguish between “battery” and “lens”. However, for targeted analysis, “battery” and “lens” represent two different aspects and should be recognised distinctively. Note that, although the sampling process of the two words in a biterm is independent to each other, their combined effects determine the status (i.e., relevant or irrelevant) of the biterm.

3.3. Inference

Following BTM [2] and TTM [1], we choose Gibbs Sampling [34] to infer the model parameters. All notations used in this section are shown in Table 2.

We first sample the status of every biterm. Intuitively, if a biterm contains a query keyword, then it is relevant with the target aspect. Let d be a binary variable and $d = 1$ indicates a biterm contains the keyword provided by users. Then, we define the probability that a biterm is relevant as $P(r|x = d, \beta^r, \beta^{ir}, \gamma) = 1$ if $d = 1$. Otherwise, we define the probability as shown below:

$$P(r|x = d, \beta^r, \beta^{ir}, \gamma) = \begin{cases} 1 & d = 1 \\ \frac{\prod_{w \in b_i} \Gamma(\beta_w^r n_{-i,w}^r + \beta_w^r \delta + \epsilon + 1)}{\Gamma(\sum_w (\beta_w^r n_{-i,w}^r + 1) + |\beta_w^r| \delta + |W| \epsilon)} & d = 0, r = relevant \\ \frac{\prod_{w \in b_i} \Gamma(n_{-i,w}^{ir} + \beta_w^{ir} + 1)}{\Gamma(\sum_w n_{-i,w}^{ir} + 1 + |\beta_w^{ir}|)} & d = 0, r = irrelevant \end{cases} \tag{1}$$

Next, we sample word selector β_w^r for all words $w \in W$. Applying Gibbs Sampling similar to TTM [1], we can obtain the equation $P(\beta_w^r | \beta_{-w}^r, w, \delta, \epsilon, p, q) \propto P(\beta^r, w | \delta, \epsilon, p, q)$. Then,

$$P(\beta_w^r | \beta_{-w}^r, \delta, \epsilon, p, q) \propto \iint P(\beta^r, w, \omega, \phi | \delta, \epsilon, p, q) d\omega d\phi \tag{2}$$

$$\propto \begin{cases} \Gamma(n_w^r + \delta + \epsilon) * \Gamma(|\beta_{-w,*}^r| \delta + |W| \epsilon + n_{-w,*}^r) * \\ \Gamma(|\beta_{-w,*}^r| \delta + \delta + |W| \epsilon) * (p + |\beta_{-w,*}^r|) & \beta_w^r = 1 \\ \Gamma(\delta + \epsilon) * \Gamma(|\beta_{-w,*}^r| \delta + \delta + |W| \epsilon + n_{-w,*}^r) * \\ \Gamma(|\beta_{-w,*}^r| \delta + |W| \epsilon) * (q + |W| - |\beta_{-w,*}^r| - 1) & \beta_w^r = 0 \end{cases}$$

For a biterm $b_i(w_{i,1}, w_{i,2})$, the probability of sampling k as the topic for $w_{i,m}$ can be computed as Equation (3).

$$P(z_{i,m} = k | \alpha, \beta^r, \beta^{ir}, \delta, \epsilon) \propto \begin{cases} P(z_{i,m} = k | \alpha, \beta^r, \delta, \epsilon) & r_{b_i} = relevant \\ P(z_{i,m} = k | \alpha, \beta^{ir}) & r_{b_i} = irrelevant \end{cases} \tag{3}$$

As mentioned before, two words in a biterm may be assigned different topics if the biterm is relevant. Therefore, we can have

$$\begin{aligned}
 P(z_{i,m} = k | \alpha, \beta^r, \delta, \epsilon) &\propto \tag{4} \\
 &\int P(z_{i,m} = k | \theta) P(\theta | \alpha) d\theta \int P(w_{i,m} | \phi_k) P(\phi_k | \beta_k^r \delta + \epsilon) d\phi \\
 &\propto \frac{n_{*|k}^{r,-im} + \alpha}{\sum_K (n_{*|k}^{r,-im} + \alpha)} \frac{\beta_{w_{i,m}|k}^r n_{w_{i,m}|k}^{r,-im} + \beta_{w_{i,m}|k}^r \delta + \epsilon}{\sum_W (\beta_{w_{i,m}|k}^r n_{w_{i,m}|k}^{r,-im} + \beta_{w_{i,m}|k}^r \delta + \epsilon)} \\
 &\propto \frac{n_{*|k}^{r,-im} + \alpha}{\sum_K n_{*|k}^{r,-im} + K\alpha} \frac{\beta_{w_{i,m}|k}^r n_{w_{i,m}|k}^{r,-im} + \beta_{w_{i,m}|k}^r \delta + \epsilon}{n_{*|k}^{r,-im} + \beta_{*|k}^r \delta + |W|\epsilon}
 \end{aligned}$$

If a biterm b_i is irrelevant (i.e., $r_{b_i} = 0$), then we directly sample a topic from ϕ^{ir} for two words $w_{i,m}$. Therefore, we can obtain the conditional probability:

$$\begin{aligned}
 P(z_{i,m} = k | \alpha, \beta^r, \beta^{ir}, \delta, \epsilon) &\propto \tag{5} \\
 &\begin{cases} \frac{n_{*|k}^{r,-im} + \alpha}{\sum_K n_{*|k}^{r,-im} + K\alpha} \frac{\beta_{w_{i,m}|k}^r n_{w_{i,m}|k}^{r,-im} + \beta_{w_{i,m}|k}^r \delta + \epsilon}{n_{*|k}^{r,-im} + \beta_{*|k}^r \delta + |W|\epsilon} & r_{b_i} = 1 \\ \frac{n_{*|k}^{ir,-i} + \alpha}{\sum_K n_{*|k}^{ir,-i} + K\alpha} \frac{(n_{w_{i,1}|k}^{ir,-i} + \beta^{ir})(n_{w_{i,2}|k}^{ir,-i} + \beta^{ir})}{(n_{*|k}^{ir,-i} + |W|\beta^{ir})(n_{*|k}^{ir,-i} + |W|\beta^{ir} + 1)} & r_{b_i} = 0 \end{cases}
 \end{aligned}$$

At last, we sample a word selector $\beta_{w|k}^r$ for a topic k , where $w \in W$ and $k \in K$.

$$\begin{aligned}
 P(\beta_{k,w}^r = s | \beta_k^r, k, \delta, \epsilon, p, q) &\propto \tag{6} \\
 &\begin{cases} \Gamma(|\beta_{-w,*|k}^r| \delta + |W|\epsilon + n_{-w,*|k}^r) * (p + |\beta_{-w,*|k}^r|) \\ * \Gamma(|\beta_{-w,*|k}^r| \delta + \delta + |W|\epsilon) * \Gamma(n_{w|k}^r + \delta + \epsilon) & s = 1 \\ \Gamma(|\beta_{-w,*|k}^r| \delta + |W|\epsilon) * (q + |W| - |\beta_{-w,*|k}^r| - 1) \\ * \Gamma(|\beta_{-w,*|k}^r| \delta + \delta + |W|\epsilon + n_{-w,*|k}^r) * \Gamma(\delta + \epsilon) & s = 0 \end{cases}
 \end{aligned}$$

4. Experimental Results

4.1. Baselines and Metrics

Baselines. Three methods are chosen to be compared with BiTTM, including Targeted Topic Model (TTM), Biterm Topic Model-Partial Data (BTM-PD), and Biterm Topic Model with a post-filtering strategy (BTM*).

- **TTM**. Targeted Topic Model is the first method for focused analysis that extracts related topics according to a target keyword provided by users. We select TTM rather than APSUM as the baseline of specialised topic models for targeted analysis because TTM outperforms APSUM in terms of *topic coherence* when the number of topics is less than 50 [3]. For targeted analysis of fine-grained topics, we believe the number of topics in a given corpus is usually less than 50. Moreover, TTM serves as the most valuable comparison because APSUM is not exactly designed for targeted analysis.
- **BTM***. As our model is developed based on biterms, we also compare with two variations of BTM that are adapted for targeted analysis. BTM is a state-of-the-art topic model for short texts, which also applies to long texts [2]. As a typical full-analysis model, BTM aims to find all topics (or all aspects) from the entire corpus. We then use a filtering strategy to eliminate topics that do not contain the target keywords. This approach is named as BTM* for simplicity.

- **BTM-PD.** This is another variation of BTM which applies the pre-filtering strategy to perform focused analysis. We use only the subset of documents containing the target keywords to model topics. As discussed before, the pre-filtering strategy is handicapped by the variability of target keyword—relevant documents may be filtered so that topics may be missed out.

Metrics. We adopt two techniques to evaluate the quality of topics: *topic coherence* [35] and *precision@n* [1] ($P@n$ for short). The former is a popular evaluation method to evaluate the quality of discovered topics [36–40]. As an automated evaluation metric, topic coherence mainly measures the interpretability of topics instead of target-relevance. More specifically, topic coherence measures document-level mutual information of keywords in topics, however, it does not reflect the relationship between topics and targets. In order to evaluate whether topics are target-relevant, we employ the metric $P@n$, also used by TTM [1], which is an evaluation based on human judgment to assess the relevance between the target and topics.

Considering the M most probable words in topic k , the topic coherence of k is defined as Equation (7).

$$TC(k) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{|\{doc(w_{k,m}, w_{k,l})\}| + 1}{|\{doc(w_{k,l})\}|} \quad (7)$$

where $|\{doc(w_{k,m}, w_{k,l})\}|$ is the number of documents containing both $w_{k,m}$ and $w_{k,l}$; $|\{doc(w_{k,l})\}|$ is the number of documents containing $w_{k,l}$, and $w_{k,l}$ is the l th most probable word in topic k . Basically, for the m th probable word, the measure considers its co-occurrence with the $m - 1$ more probable words. A smoothing count of 1 is added to avoid leading the logarithm to zero. Basically, the more the measure approximates to zero, the more coherent the discovered topics are.

Given the set of topics discovered by all models, suppose there are Ku topics that have been verified by users to be related with the target aspect. Moreover, from all topics discovered by a particular model m , suppose there are Km topics related with the target. Then, the precision of model m at rank position n is defined as follows:

$$P_m@n = \frac{\sum_{z=1}^{K_m} |\{CorrectWords(z)\}|}{\sum_{z=1}^{K_u} |\{CorrectWords(z)\}|} \quad (8)$$

where $|\{CorrectWords(z)\}|$ is the number of words, among the top n words of topic z , which are relevant to the target (Note that, if a discovered topic is potentially related with multiple semantic topics, the best semantic topic based on the top 20 words will be adopted).

Therefore, the two evaluation methods have different merits and objectives. For example, topic coherence is an automated evaluation metric reflecting the interpretability of topics. $P@n$ demands human judgement and assesses the relevance between the discovered topics and the queried target. For the sake of fairness, we use $P@n$ to evaluate all comparing models (i.e., BiTTM, TTM, BTM-PD and BTM*) to find out the effectiveness of the models in performing focused analysis. However, we only compare BiTTM and TTM in terms of topic coherence to evaluate the topic quality since the other two models are variations of BTM, which is essentially designed for full-analysis of topics.

4.2. Data sets & Experimental Settings

Data sets. In order to comprehensively evaluate the performance of our proposed model, we conduct experiments on different types of text. In particular, three types of documents are considered, including short, medium and long texts. For each type of documents, we select three data sets. The description of the nine datasets used in our experiments is provided in Table 3. The datasets are all publicly available at the URLs listed in the bottom of Table 3.

Experimental Settings. In our experiments, we use various words as target queries to analyse the influence exerted by diverse targets on performance. For parameter settings, we follow the hyper-parameter setting in TTM: $\alpha = \gamma = 1, \beta^{ir} = 0.001, p = q = 1$, and the two smoothing priors are set as $\delta = 0.001, \epsilon = 1 \times 10^{-7}$. Other baselines follow the parameter settings in their respective papers.

Table 3. Datasets.

Type	Datasets			
	Source	Domain	Length	Size (KB)
short	cigar ^a	Twitter	2.947836	641
	ecig ^a	Twitter	3.499578	708
	Oscars ^b	Twitter	3.906165	565
medium	baby ^c	Amazon	28.07813	141
	camera ^a	Amazon	79.08307	1285
	computer ^a	Amazon	80.9001	1295
long	home ^c	Amazon	179.4867	619
	food ^c	Amazon	258.4938	195
	care ^c	Amazon	675.1493	1523

^a <http://jmcauley.ucsd.edu/data/amazon/>, accessed on 28 October, 2021; ^b <https://github.com/shuaiwanghk/TTM>, accessed on 28 October, 2021; ^c <https://www.kaggle.com/madhurinani/oscars-2017-tweets>, accessed on 28 October, 2021.

4.3. Quantitative Evaluation

In this subsection, we analyse the quality of discovered topics from two aspects: *topic coherence* (representing topic interpretability or semantic coherence) and *P@n* (indicating topic relevance).

Analysing the results of topic coherence: The average topic coherence achieved by BiTTM and TTM is shown in Table 4, the more the score approximates to zero, the more coherent the discovered topics are. As we can see from the table, BiTTM is not comparable to TTM for analysing short texts in terms of topic coherence. However, with the increase of document length, BiTTM starts to outperform TTM. The reason why BiTTM generally works better than TTM on medium and long texts is because TTM is a sentence-based model for which the information between consecutive sentences will be lost. In contrast, by considering core biterms that may come from neighbouring sentences, our BiTTM model captures the semantics crossing sentences so that more interpretable topics can be generated. However, since it is quite often for a short text document to contain only one sentence, the limitation of sentence-based TTM cannot be reflected. Generally, by beating TTM on non-short text documents, BiTTM has a broader applications in text data analysis.

To evaluate the model performance with respect to different queries, we randomly sample query keywords from the documents according to word frequency distributions. We plot the comparative results of BiTTM and TTM in Figure 4, where the horizontal axis represents the word frequency of the target keyword, and the vertical axis indicates the percentage of documents containing the target. There are three types of symbol in the figure: red dots, green squares and blue triangles. Each symbol corresponds to a comparison between the topics discovered by BiTTM and TTM with respect to a query. In particular, a green square means BiTTM obtains a better topic coherence than TTM for this query, while a blue triangle implies the opposite. For a red dot, it indicates that TTM fails to discover the specified number of topics or words under some topics for this particular query. For example, we set the number of topics to 5 for the experiments in Figure 4 and consider the top 10 words for each topic. However, TTM discovers less than 5 topics or less than 10 words for a topic when handling queries corresponding to red dots. Note that, this situation does not happen for BiTTM.

Table 4. Topic coherence.

Datasets	BiTTM					
	TopM = 5	TopM = 10	TopM = 15	TopM = 20	TopM = 25	TopM = 30
cigar	-43.85160224	-213.938738	-515.6799558	-942.5318026	-1493.639111	-2170.963155
ecig	-47.18005764	-225.7818028	-537.0456906	-975.5867746	-1540.959263	-2233.885475
Oscars	-41.56838081	-203.696938	-495.7966598	-913.9376539	-1451.063314	-2102.622643
baby	-18.77161638	-97.92745779	-258.704099	-512.2443085	-862.3028153	-1314.514524
camera	-11.60928852	-57.83209167	-151.6040027	-295.3511581	-493.2409214	-748.8210303
computer	-14.82764795	-66.09833632	-163.5166346	-310.8723517	-517.6532656	-789.6438756
home	-9.146932399	-45.84159746	-117.7837843	-239.7455224	-418.9237371	-656.914665
food	-7.47342957	-42.52194503	-109.7998671	-220.2148208	-370.0466113	-563.1454128
care	-11.52868834	-49.17233659	-116.4456707	-224.1156048	-375.6415389	-585.8790988

Datasets	TTM					
	TopM = 5	TopM = 10	TopM = 15	TopM = 20	TopM = 25	TopM = 30
cigar	-34.68247162	-197.0478114	-493.1432304	-921.1591797	-1481.235217	-2165.174583
ecig	-34.67967533	-200.4202621	-502.3021289	-935.8301943	-1495.777983	-2182.933002
Oscars	-31.17239749	-178.8829525	-456.1151923	-858.0192159	-1383.122495	-2032.967068
baby	-16.2223621	-101.2752879	-281.5019366	-562.4756204	-937.8116571	-1440.487996
camera	-11.28057698	-62.01834755	-164.0824617	-318.5467601	-529.0200582	-801.4709905
computer	-13.16215736	-68.04826101	-174.2534566	-332.9993329	-549.2408003	-824.3810116
home	-9.934943893	-51.89310174	-131.4846734	-254.4960003	-428.6019462	-643.0140473
food	-11.33222981	-54.36687511	-128.2609949	-242.0789656	-399.9506445	-600.8773534
care	-9.739145131	-53.9780039	-133.7955495	-254.3544155	-432.013784	-666.1259446

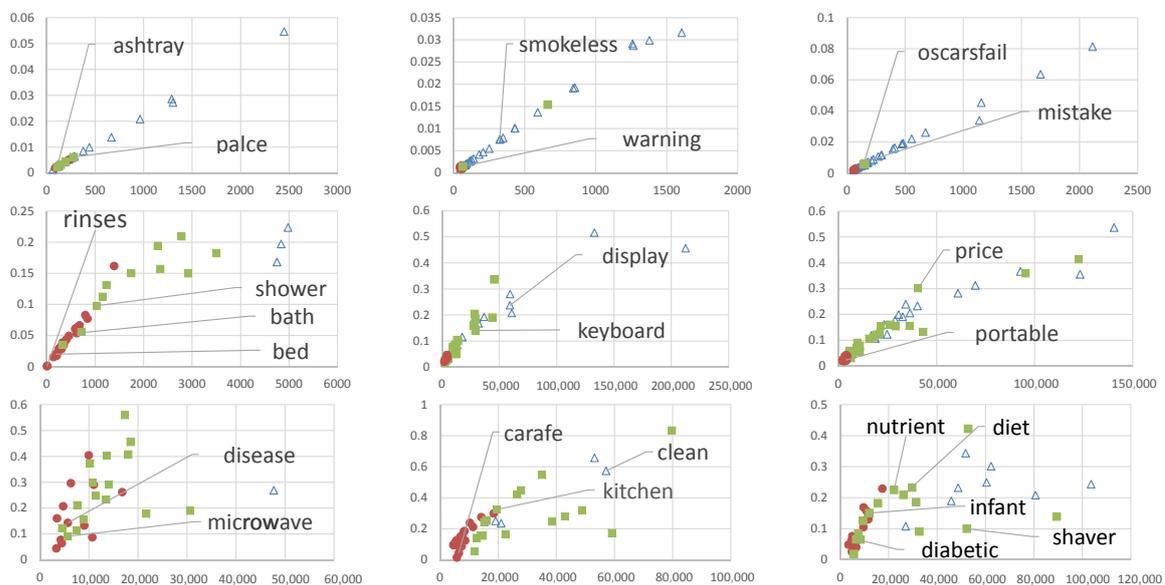


Figure 4. Effect of various targets on topic coherence. The x-axis shows the word frequency of targets, and the y-axis represents the percentage of documents containing a target.

The most obvious trend that can be observed from Figure 4 is that the red dots usually appear in the lower left corner, the blue triangles gather in the upper right corner, and the green squares fall in between. The red dots in the lower left corner imply that TTM is prone to miss out topics when dealing with infrequent targets. The blue triangles in the right corner suggest that TTM performs better when the targets appear very frequently in many documents. However, the number of such target keywords may be limited. On the

contrary, BiTTM achieves satisfactory performance for a diverse range of targets even if they are infrequent in the corpus. This also verifies the effectiveness of using core words to enrich the semantic information in the context of the target keyword (i.e., BiTTM strategy) than taking words in same sentences as bridges to connect potentially target-related words (i.e., TTM strategy).

Analysing the results of $P@n$: To calculate the measure of $P@n$, similar to TTM, three human labelers familiar with the data sets are engaged to label the results. The $P@n$ values at the rank positions of 5, 10 and 20 are reported in Table 5, from which several interesting outcomes can be observed. Firstly, the performance of the two variations of BTM (i.e., BTM-PD and BTM*) is generally worse than that of the two specialised topic models (i.e., BiTTM and TTM), which demonstrates that full-analysis topic models with filtering strategies are not suitable for targeted analysis because they are prone to detect general topics instead of fine-grained target-related topics. In addition, comparing the two BTM variations, BTM-PD is better than BTM* in most cases, which proves that the pre-filtering strategy is more effective in removing irrelevant words than the post-filtering strategy. Secondly, the average $P@n$ of BiTTM achieves a gain of more than 10% compared with TTM, and more than 26% compared with BTM-PD, over all queries in the table and the settings of n . Moreover, the performance difference among the three types of document is not significant, whereas the different target queries have influence on the $P@n$ results, which will be explained later using concrete examples. Thirdly, TTM is the second best model for $P@5$. However, for $P@10$, TTM achieves the best performance than all other models for some queries. It suggests the tendency of TTM to put target-related words in lower-ranked positions.

Table 5. $P@n$ scores of all models over a set of 18 targets on 9 data sets. n is set to 5, 10 and 20.

Type	Datasets	Targets	BiTTM			TTM			BTM-PD			BTM*		
			P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20
short	cigar	ashtray	0.92	0.84	0.6	0.92	0.66	0.46	0.6	0.6	0.39	0.2	0.2	0.135
		place	0.64	0.56	0.43	0.52	0.44	0.29	0.48	0.4	0.26	0.22	0.17	0.135
	ecig	smokeless	0.72	0.64	0.45	0.56	0.5	0.45	0.44	0.4	0.31	0.3	0.24	0.19
		warning	0.6	0.54	0.54	0.48	0.46	0.37	0.36	0.3	0.25	0.4	0.24	0.215
	Oscars	mistake	0.6	0.56	0.49	0.4	0.48	0.45	0.2	0.32	0.28	0.28	0.18	0.14
		oscarsfail	0.64	0.52	0.44	0.36	0.4	0.31	0.24	0.18	0.15	0.26	0.18	0.125
medium	baby	rinses	0.52	0.38	0.3	0.2	0.26	0.24	0.08	0.1	0.11	0.12	0.12	0.085
		shower	0.52	0.48	0.38	0.44	0.4	0.32	0.2	0.36	0.31	0.24	0.19	0.145
	camera	portable	0.52	0.48	0.4	0.4	0.44	0.29	0.32	0.28	0.23	0.1	0.09	0.08
		price	0.52	0.48	0.51	0.48	0.44	0.42	0.44	0.34	0.26	0.08	0.11	0.09
	computer	display	0.48	0.74	0.72	0.4	0.62	0.62	0.28	0.32	0.32	0.1	0.15	0.165
		keyboard	0.52	0.66	0.59	0.44	0.6	0.52	0.32	0.34	0.33	0.18	0.15	0.14
long	home	clean	0.76	0.8	0.69	0.68	0.72	0.64	0.32	0.34	0.32	0.25	0.19	0.18
		kitchen	0.72	0.72	0.66	0.68	0.8	0.66	0.64	0.52	0.44	0.22	0.18	0.16
	food	disease	0.6	0.6	0.6	0.48	0.46	0.42	0.44	0.4	0.38	0.22	0.23	0.19
		microwave	0.84	0.58	0.47	0.68	0.64	0.44	0.32	0.32	0.26	0.18	0.17	0.125
	care	diabetic	0.56	0.62	0.53	0.48	0.44	0.25	0.28	0.28	0.23	0.12	0.13	0.12
		infant	0.48	0.54	0.51	0.36	0.3	0.29	0.2	0.14	0.11	0.18	0.15	0.105
average score			0.62	0.60	0.52	0.50	0.50	0.41	0.34	0.33	0.27	0.20	0.17	0.14
improvement by BiTTM						+0.12	+0.09	+0.10	+0.28	+0.27	+0.24	+0.42	+0.43	+0.38

To explore the influence of different queries, let us take a closer look at two specific targets: “ashtray” (in the short-text data set “cigar”) and “rinses” (in the medium-text data set “baby”). As shown in Figure 4, both targets are infrequent words (appearing in the lower left corner) in respective datasets. However, the $P@n$ scores of BiTTM and TTM for the two queries, as shown in Table 5, are remarkably different. Basically, both models perform well with respect to “ashtray” but not with respect to “rinses”, especially for TTM. The $P@n$ score of TTM for handling “rinse” is unsatisfactory and several inexplicable words, such as “attention” and “entertain”, appear in the discovered topics, which makes it hard to interpret the topics. By examining the datasets, we find that documents containing “ashtray” consistently describe the appearance of ashtrays such as colours and materials. That is, the documents are pretty clean and relevant, which explains why both BiTTM and TTM process the query well. Nevertheless, the documents containing “rinses” are mostly composed of short sentences, such as “It rinses out well and dries quickly.” and “Rinses/Washes easy.”, where the meaningful descriptions are hidden in the context of sentences containing “rinses”. TTM cannot handle this situation since it is a sentence-based model. The two examples explain why the performance varies with respect to query keywords.

Comparing the performance in terms of topic coherence and $P@n$, we notice that BiTTM is more capable to acquire topics related to the target (i.e., high $P@n$ scores) than to generate semantically coherent topics (i.e., better topic coherence values), especially for short text documents. This is because words related to the target do not necessarily have high co-occurrence, which is used to calculate topic coherence. For instance, “Oktoberfest” is an appropriate word related to the target “place” in the dataset *cigar* because a type of cigar named Quesada Oktoberfest is released in October for celebrating the famous Germany beer festival. However, “Oktoberfest” as a low-frequency word can not provide enough mutual information, which directly causes the poor performance in topic coherence. Conversely, a high-frequency word “rolled” contributes to high topic coherence score but it is not selected by BiTTM since it is too general to describe the target “place”.

4.4. Time Efficiency Analysis

As mentioned before, it is ideal for targeted analysis to provide responses to user queries as soon as possible. Therefore, in this experiment, we analyse the time efficiency of the comparative models.

The average time cost of the four methods on each dataset over 40 random queries is shown in Table 6. It can be observed that, generally, BiTTM has the best time efficiency, followed by BTM-PD. TTM is significantly slower without any preprocessing strategy, and BTM* is the most inefficient model since BTM performs full analysis on the complete dataset.

Table 6. The time cost (seconds) of all datasets. The best performance on each dataset is in bold.

Domain	Size (KB)	BiTTM	BTM-PD	TTM	BTM*
cigar	641	6.012	0.378	60.994	314.940
ecig	708	2.960	2.900	107.008	489.180
Oscars	565	3.620	3.934	58.042	418.560
baby	141	16.011	9.753	24.246	1153.740
camera	1285	81.086	164.834	811.684	15,997.380
computer	1295	96.978	147.023	713.068	17,559.660
home	619	37.083	90.283	129.481	7116.300
food	195	9.166	31.833	18.534	2496.600
care	1523	135.040	218.065	1516.927	18,767.580

To clearly demonstrate the impact of data size on the time efficiency, we plot the results in Figure 5 where the grey bars denote the size of datasets and the polylines in different colours indicate the time cost of different methods. Note that, since the time consumption of BTM* is not comparable to the others, only three models (i.e., BiTTM, BTM-PD and TTM) are displayed in the figure. It can be observed that, generally, the time cost of all methods increases with respect to the increment of data size. However, the size of dataset has a greater impact on TTM than the other two methods, which shows that TTM is not suitable for processing large data sets. In contrast, BiTTM and BTM-PD have a better capability to adapt to large data sets. For these two methods, the difference of data size does not make dramatic changes to time consumption since they both have preprocessing strategies to focus on only the portion of data related to query targets. The difference between BiTTM and BTM-PD is that BiTTM is faster than BTM-PD especially when the length of documents increases. The reason is that the pre-filtering strategy adopted by BTM-PD is a simple and rough processing. It selects documents as long as they contain the query keywords. Consequently, irrelevant information contained by such documents will be included and processed as well, which negatively contributes to the time efficiency of BTM-PD.

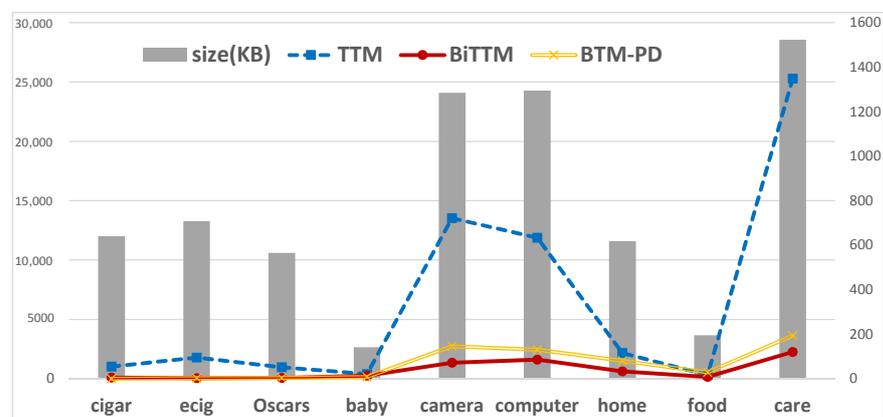


Figure 5. Time cost in datasets of different sizes.

To illustrate the impact of document length on the time efficiency, the percentage histogram of time cost of BiTTM, TTM and BTM-PD is plotted in Figure 6, where the average document length increases from left to right. It can be observed that the time efficiency of TTM is worst on short texts. Recall that the topic coherence of TTM on short texts is better than BiTTM. This experiment shows that TTM achieves this by significantly sacrificing time efficiency, while the topic quality in terms $P@n$ of TTM on short texts is also worse than that of BiTTM. Moreover, we can see that efficiency performance of BTM-PD is worse on long texts, compared to its performance on short texts. This is because BTM-PD is a biterm-based topic model and long texts generally have more biterms than short texts. Although BiTTM is also a biterm-based model, the strategy of selecting “core biterms” removes a lot of irrelevant biterms so that the performance of BiTTM on long texts is also promising.

Therefore, Figures 5 and 6 demonstrate that BiTTM can be widely applied to various types of text data, because both data size and document length have no great impact on its time efficiency, thanks to the core biterm-based preprocessing strategy.

4.5. Qualitative Evaluation

We present qualitative analysis of the result topics generated by comparative models in this subsection. We focus on evaluating from two aspects: performance of discovering as many fine-grained relevant topics as possible and performance of dealing with semantically approximate targets. For exemplified queries discussed in the following, we have shown their word frequency and document frequency in Figure 4.

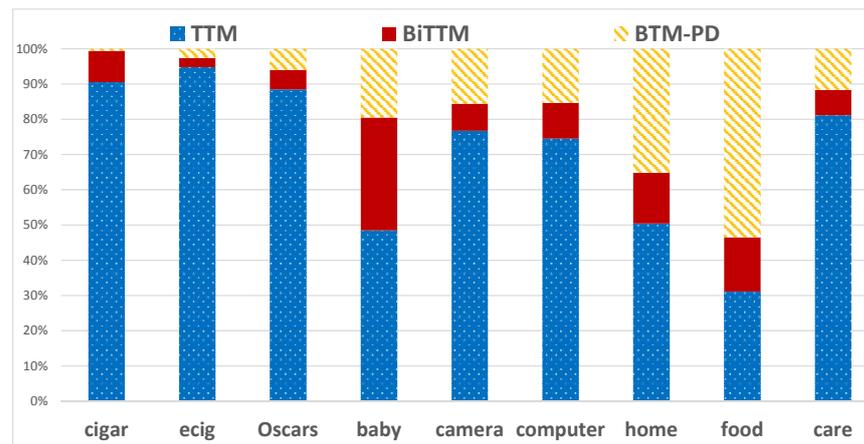


Figure 6. Time cost in datasets with different document lengths.

4.5.1. Discovering Relevant Topics

We take the query “disease” in the dataset *food* as an example. Table 7 shows the topics discovered by the four comparative models, together with the top 10 words of each topic. The third row of Table 7 are the labels we assign manually to summarise the semantics of each topic, where *SFA* is the abbreviation for *Saturated Fatty Acid*. Words that do not semantically align with the topics are displayed in red.

Table 7. The contents of topics on target “disease” in long texts *food*.

Datasets: Food. Target: Disease														
BiTTM					TTM				BTM-PD			BTM*		
Tea	SFA	Research	Risk	Prevention	Tea	SFA	Research	Risk	Tea	SFA	Research	Tea	SFA	Research
tea	fat	study	disease	cherry	disease	fat	heart	risk	tea	protein	cherry	tea	fat	study
effect	cancer	increase	risk	reduce	tea	saturated	study	disease	green	oil	chocolate	antioxidant	saturated	disease
benefit	people	research	heart	tart	flavour	tart	cancer	cherry	weight	fat	tart	disease	disease	risk
work	health	small	high	prevent	bad	chip	reduce	star	fat	palm	study	heart	risk	heart
lower	fruit	antioxidant	green	include	price	sweetener	order	bag	loss	saturated	disease	fruit	oil	reduce
taste	blood	good	cell	body	higher	follow	brand	measure	study	quality	dark	rich	coconut	show
long	sugar	eat	level	find	company	thing	vegetable	calorie	increase	eat	cocoa	provide	find	cherry
kind	animal	amount	show	product	back	production	cook	state	body	diet	sweet	health	study	health
pure	saturated	diet	day	result	nutrient	add	large	price	drink	disease	cancer	substitute	increase	food
rich	drink	water	sweet	add	simply	expensive	drink	shipping	calorie	gram	product	vegetable	health	cancer

Compared with the topics discovered by BiTTM, all the other three methods fail to identify the topic *prevention*, which is clearly a relevant topic of “disease”. Moreover, the two BTM variation models (i.e., BTM-PD and BTM*) miss out the topic *risk*. By taking a closer look, we find that this is because the two BTM models cannot distinguish between the two topics *risk* and *research* that are different delicately. In other words, the two BTM models discover a topic combining *research* and *risk*. This is understandable because BTM as a full-analysis topic model discovers general topics. TTM succeeds in discovering both *research* and *risk*, but the topic quality is poorer than that of BiTTM (e.g., there are more bold words in the two topics discovered by TTM, which means more inconsistent words in results of TTM). Therefore, BiTTM discovers more relevant and fine-grained topics than other models for this example query.

Consider the topic *SFA* that is discovered by all of the four models. Results of BiTTM clearly indicate that saturated fatty acids affect blood sugar and carcinogenesis, but the results of other methods are not satisfactory. For example, TTM tends to find out which foods (e.g., tart, chip and sweetener) have unsaturated fatty acids. BTM-PD and BTM* focus on food ingredients (e.g., palm oil and protein). These results are not related with the target “disease” queried by users. Hence, the topic quality of BiTTM is better than that of other models as well in this example.

4.5.2. Handling Semantically Approximate Targets

When the targets supplied by users are semantically approximate, a set of similar relevant topics are supposed to be discovered. We further examine the performance of the comparative models in handling semantically approximate targets. In particular, we analyse two types of semantically approximate queries mentioned in Section 3.1: synonyms and diverse descriptions of the same event.

An example of the first type is shown in Table 8. We query the dataset “baby” with two targets, “bath” and “shower”, which share similar semantics in the data set of Amazon reviews of baby products. A successful model should return similar topics. As shown in the table, BiTTM is the only model that can obtain the set of four meaningful topics for both queries, while other methods either miss topics or generate vague content for topics.

Table 8. The contents of topics on two approximate semantic targets in dataset “baby”.

BiTTM				TTM			BTM-PD		BTM*		
Target: Bath											
Blanket	Spout	Protection	Sentiment	Blanket	Spout	Sentiment	-	-	Blanket	Spout-1	Spout-2
cover	spout	cute	fit	spout	fit	spout	cover	cover	spout	faucet	
pull	head	play	time	cover	cute	bath	bath	bath	cute	easy	
product	stay	tub	daughter	snap	son	shower	spout	spout	head	spout	
shower	bath	put	faucet	bright	whale	tub	faucet	faucet	fit	cute	
time	whale	easy	give	hate	quickly	cover	head	head	cover	tub	
child	month	buy	nice	read	shower	whale	protect	protect	faucet	cover	
recommend	crib	thing	problem	realize	bend	fit	tub	tub	tub	product	
blanket	easily	kid	bump	mobile	bear	faucet	whale	whale	son	bath	
lot	side	protect	diaper	break	front	knob	fit	fit	whale	head	
start	year	bumper	find	parent	touch	perfect	time	time	bath	whale	
Target: shower											
blanket	spout	cute	son	shower	spout	cover	spout	spout	blanket	spout	spout
buy	shower	thing	pull	gift	tub	fit	shower	shower	buy	cover	fit
cover	head	easy	bath	bath	kid	cute	cover	cover	gift	fit	shower
put	stay	faucet	time	buy	top	whale	cute	whale	swaddle	faucet	cute
big	kid	product	daughter	hole	easy	face	faucet	bath	receive	head	pull
pretty	gift	side	play	blanket	head	easily	head	tub	friend	shower	stay
safe	perfect	problem	nice	time	remove	couple	pull	pull	child	product	cover
stroller	soft	install	protect	thing	worry	snap	tub	mold	day	tub	son
change	worry	child	find	trip	picture	high	whale	kid	hold	time	tub
quality	bag	car	fall	totally	face	expect	bath	faucet	shower	pull	whale

For instance, except BiTTM, the other three methods fail to identify the topic *blanket* with respect to the query “bath”, while TTM and BTM* can retrieve the topic with respect to the target “shower”. According to the results of BiTTM, we find that *blanket* is an important aspect of bath/shower, because most people will cover their babies with a blanket after a bath/shower. Hence, the topic *blanket* is an aspect in which users are interested. Ignoring an important topic hinders downstream analysis and applications, such as high-quality personalised services and commodity recommended systems. As another example, there are two topics discovered by BiTTM only: *sentiment* and *protection*. Checking the content of topic *sentiment*, we learn that users tend to associate emotional expressions (e.g., “have a nice time with daughter/son”) when commenting on shower/bath products. This topic thus implies users’ emotional polarity of products, which is important for applications such as user profiling, recommendation and public opinion monitoring. The topic *protection* describes safety products that can be installed in tubs or on faucets. The safety issue of bath is an important concern especially for baby products, and it is non-ideal for the other three methods to ignore this topic.

Moreover, we find that BTM-PD extracts only two topics for both queries and the content of the topics are too vague to understand (e.g., we are not able to assign semantic labels to the topics). There are six identical words between the two sets of top 10 words, which makes it very hard to distinguish between the semantics of the topics. The same

situation occurs to BTM*—there are two similar topics about “spout”. For example, given the query “bath”, the two topics have eight identical words in the top 10 words. The content of these two topics may be correct, but the information expressed is redundant. It is not useful to generate identical topics but increasing the difficulty of further analysis.

Table 9 shows an example of the second type. Given the dataset *Oscars*, both “mistake” and “oscarfail” refer to the same event that the Best Picture Award, which should belong to *Moonlight*, was wrongly presented to *La La Land* because of a wrong envelope. As we can see from the table, BiTTM can acquire three fine-grained relevant topics, which describes the process of the event development: At the beginning of the event, two guests present the Best Picture to *La La Land*, and no one was aware of the mistake. Many tweets emerge to talk about *La La Land* and express congratulations to the actors and the producer, which can be seen from the content of the topic *beginning*. Next, the error is corrected and the real winner is another movie *Moonlight*. Topic *correction* is a perfect interpretation of this stage. Note that, the top 10 words of this topic with respect to the query “mistake” contains the word “oscarfail”, which demonstrates the usefulness of the core biterns strategy used by BiTTM. The third topic *discussion* covers the discussion of the actors’ reaction after this mistake has happened.

Table 9. The contents of topics on two approximate semantic targets in dataset “Oscars”.

BiTTM		TTM	BTM-PD	BTM*		
Target: Oscarsfail						
Beginning	Correction	Discussion	Discussion	Discussion	Discussion	Discussion
lalaland	moonlight	oscar	oscarfail	win	win	moonlight
hollywood	oscarfail	envelopegate	oscar	oscarfail	vote	bestpicture
winner	award	picture	envelopegate	lalaland	lalaland	lalaland
cast	mistake	actor	majorgroup	vote	moonlight	russians
actress	vote	reaction	pick	short	popular	hack
emma	variety	electoral	scene	white	electoral	election
people	moment	time	hollywood	rating	oscarfail	envelope
white	mahershala	movie	hack	helmets	Irihendry	votetrumpics
barryjenkins	violadavis	word	moana	moonlight	word	neontaster
bestpicture	black	affleck	auliicavalho	documentary	oscar	oscarfail
Target: mistake						
lalaland	moonlight	win	win	lalaland	lalaland	moonlight
realize	winner	mistake	moonlight	mistake	moonlight	winner
announce	real	academy	picture	moonlight	picture	picture
award	oscar	black	moment	crew	win	announce
producer	night	film	realize	realize	realize	lalaland
moment	people	reaction	lalaland	cast	mistake	real
congrat	oscarfail	give	watch	moment	moment	abc
abc	movie	congratulation	mistake	watch	crew	watch
thr	theellenshow	time	cast	win	watch	mistake
happen	russians	support	crew	picture	cast	realize

In contrast, TTM only retrieves the topic *discussion* and the quality is not satisfactory. Some irrelevant words like *Moana*, another movie, appear in the topic. BTM-PD and BTM* also discover only the topic of *discussion* with respect to the target of “oscarfail”, and the quality is low. For example, the word “documentary” which is not related with the two movies appears in results. Although the quality has improved with respect to the target “mistake”, the two topics discovered BTM* are too similar with 6 identical words in top 10.

5. Conclusions

Targeted topic modelling is an increasingly vital task due to the prevalence of texts on the web and the limit of users’ interests. Compared with full-analysis topic models, such as LDA [41] and BTM [2], which are designed to discover all topics in a dataset, targeted analysis models aim to perform an in-depth semantic analysis to extract fine-grained topics about which users are concerned. In this paper, we propose a core bitern-based topic

model for targeted analysis named BiTTM. Motivated by the fact that only part of the entire dataset is related with target aspects and the requirement to efficiently provide responses to user queries, a pre-processing mechanism is indispensable and core biterms related to target queries are proposed to be extracted (from neighbouring sentences) to preserve relevant information and to capture semantics across documents. Fine-grained topics are then modelled from core biterms where different topics are allowed to be sampled for each word in a biterm. Extensive experiments have been conducted to evaluate BiTTM, compared with the state-of-the-arts, in terms of topic coherence, topic relevance and time efficiency on nine real-world data sets including short texts, medium texts as well as long texts with respect to various query keywords randomly sampled from the corpus. The experimental results demonstrate that BiTTM outperforms existing models remarkably in terms of retrieving high quality topics relevant to targets and computation efficiency.

Future research should consider the potential effects of relevance in semantic space more carefully, for example, using multi-source semantic information to enhance the computational accuracy of relevance may significantly improve model performance. Recent studies [42–46] have shown that using word embeddings for topic modelling is potential for text analysis, and this may constitute the object of future studies.

Author Contributions: Conceptualization, J.W. and L.L.; methodology, J.W.; validation, J.W. and L.L.; formal analysis, J.W. and L.C.; writing—original draft preparation, J.W.; writing—review and editing, L.C., L.L. and X.W.; supervision, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was jointly funded by the Sichuan Science and Technology Program, Grant/Award Numbers: 2019ZYZF0169; the A Ba Achievements Transformation Program, Grant/Award Number: 19CGZH0006, R21CGZH0001; the Chengdu Science and technology planning project, Grant/Award Number: 2021-YF05-00933-SN.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: This research was jointly funded by the Sichuan Science and Technology Program, Grant/Award Numbers: 2019ZYZF0169; the A Ba Achievements Transformation Program, Grant/Award Number: 19CGZH0006, R21CGZH0001; the Chengdu Science and technology planning project, Grant/Award Number: 2021-YF05-00933-SN. We would like to thank Wu Deng for his encouragement and guidance throughout this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, S.; Chen, Z.; Fei, G.; Liu, B.; Emery, S. Targeted Topic Modeling for Focused Analysis. In Proceedings of the ACM SIGKDD International Conference, San Francisco, CA, USA, 13–17 August 2016; pp. 1235–1244.
2. Cheng, X.; Yan, X.; Lan, Y.; Guo, J. BTM: Topic Modeling over Short Texts. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2928–2941. [[CrossRef](#)]
3. Rakesh, V.; Ding, W.; Ahuja, A.; Rao, N.; Sun, Y.; Reddy, C.K. A Sparse Topic Model for Extracting Aspect-Specific Summaries from Online Reviews. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, 23–27 April 2018; pp. 1573–1582. [[CrossRef](#)]
4. Kim, H.; Choi, D.; Drake, B.L.; Endert, A.; Park, H. TopicSifter: Interactive Search Space Reduction through Targeted Topic Modeling. In Proceedings of the 14th IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2019, Vancouver, BC, Canada, 20–25 October 2019; pp. 35–45. [[CrossRef](#)]
5. He, J.; Li, L.; Wang, Y.; Wu, X. Hierarchical features-based targeted aspect extraction from online reviews. *Intell. Data Anal.* **2021**, *25*, 205–223. [[CrossRef](#)]
6. Nguyen, T.; Pham, T.; Le, H.; Nguyen, T.; Bui, H.; Ha, Q. A Targeted Topic Model based Multi-Label Deep Learning Classification Framework for Aspect-based Opinion Mining. In Proceedings of the 12th International Conference on Knowledge and Systems Engineering, KSE 2020, Can Tho City, Vietnam, 12–14 November 2020; pp. 165–170.
7. Li, S.; Zhang, Y.; Pan, R.; Mao, M.; Yang, Y. Recurrent Attentional Topic Model. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 3223–3229.
8. Cai, G.; Peng, L.; Wang, Y. *Topic Detection and Evolution Analysis on Microblog*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 67–77.

9. Ye, C.; Liu, D.; Chen, N.; Lin, L. Mapping the topic evolution using citation-topic model and social network analysis. In Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery, Zhangjiajie, China, 15–17 August 2016; pp. 2648–2653.
10. Xia, Y.; Tang, N.; Hussain, A.; Cambria, E. Discriminative Bi-Term Topic Model for Headline-Based Social News Clustering. In Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, FL, USA, 18–20 May 2015; pp. 311–316.
11. Amara, A.; Taieb, M.A.H.; Aouicha, M.B. Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Appl. Intell.* **2021**, *51*, 3052–3073. [[CrossRef](#)]
12. Hu, Y.; Tai, C.; Liu, K.E.; Cai, C. Identification of highly-cited papers using topic-model-based and bibliometric features: The consideration of keyword popularity. *J. Inf.* **2020**, *14*, 101004. [[CrossRef](#)]
13. Zhang, W.; Wang, J. Integrating Topic and Latent Factors for Scalable Personalized Review-based Rating Prediction. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3013–3027. [[CrossRef](#)]
14. Wang, H.; Li, W. Relational Collaborative Topic Regression for Recommender Systems. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1343–1355. [[CrossRef](#)]
15. Zhang, Y.; Chen, W.; Zha, H.; Gu, X. A Time-Topic Coupled LDA Model for IPTV User Behaviors. *IEEE Trans. Broadcast.* **2015**, *61*, 56–65. [[CrossRef](#)]
16. Hu, C.; Hu, Y.; Xu, W.; Shi, P.; Fu, S. Understanding Popularity Evolution Patterns of Hot Topics Based on Time Series Features. In Proceedings of the Web Technologies and Applications—APWeb 2014 Workshops, SNA, NIS, and IoTS, Changsha, China, 5 September 2014; pp. 58–68.
17. Feuerriegel, S.; Ratku, A.; Neumann, D. Analysis of How Underlying Topics in Financial News Affect Stock Prices Using Latent Dirichlet Allocation. In Proceedings of the Hawaii International Conference on System Sciences, HICSS 2016, Koloa, HI, USA, 5–8 January 2016; pp. 1072–1081.
18. Viermetz, M.; Skubacz, M.; Ziegler, C.N.; Seipel, D. Tracking Topic Evolution in News Environments. In Proceedings of the IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, Washington, DC, USA, 21–14 July 2008; pp. 215–220.
19. Phuong, D.V.; Phuong, T.M. A keyword-topic model for contextual advertising. In Proceedings of the Symposium on Information and Communication Technology 2012, SoICT '12, Halong City, Vietnam, 23–24 August 2012; pp. 63–70.
20. Kalyanam, J.; Mantrach, A.; Saez-Trumper, D.; Vahabi, H.; Lanckriet, G. Leveraging Social Context for Modeling Topic Evolution. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 517–526.
21. Sordo, M.; Ogihara, M.; Wuchty, S. Analysis of the Evolution of Research Groups and Topics in the ISMIR Conference. In Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, 26–30 October 2015; pp. 204–210.
22. Zhao, B.; Xu, W.; Ji, G.; Tan, C. Discovering Topic Evolution Topology in a Microblog Corpus. In Proceedings of the Third International Conference on Advanced Cloud and Big Data, Yangzhou, Jiangsu, China, 30 October–1 November 2015; pp. 7–14.
23. Gou, Z.; Li, Y. A method of query expansion based on topic models and user profile for search in folksonomy. *J. Intell. Fuzzy Syst.* **2021**, *41*, 1701–1711. [[CrossRef](#)]
24. Sperrle, F.; Schäfer, H.; Keim, D.A.; El-Assady, M. Learning Contextualized User Preferences for Co-Adaptive Guidance in Mixed-Initiative Topic Model Refinement. *Comput. Graph. Forum* **2021**, *40*, 215–226. [[CrossRef](#)]
25. Lin, T.; Tian, W.; Mei, Q.; Cheng, H. The dual-sparse topic model: Mining focused topics and focused terms in short text. In Proceedings of the 23rd International World Wide Web Conference, WWW '14, Seoul, Korea, 7–11 April 2014; pp. 539–550. [[CrossRef](#)]
26. Chien, J.T.; Chang, Y.L. Bayesian Sparse Topic Model. *J. Signal Process. Syst.* **2014**, *74*, 375–389. [[CrossRef](#)]
27. Slutsky, A.; Hu, X.; An, Y. Learning Focused Hierarchical Topic Models with Semi-Supervision in Microblogs. In Proceedings of the Advances in Knowledge Discovery and Data Mining—19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, 19–22 May 2015; Part II; pp. 598–609.
28. Chen, X.; Zhou, M.; Carin, L. The contextual focused topic model. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 96–104.
29. Pu, X.; Jin, R.; Wu, G.; Han, D.; Xue, G.R. Topic Modeling in Semantic Space with Keywords. In Proceedings of the ACM International Conference on Information and Knowledge Management, Melbourne, VIC, Australia, 19–23 October 2015; pp. 1141–1150.
30. Williamson, S.; Wang, C.; Heller, K.A.; Blei, D.M. The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 1151–1158.
31. Zhu, B.; Cai, Y.; Zhang, H. Sparse Biterm Topic Model for Short Texts. In Proceedings of the Web and Big Data—5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, 23–25 August 2021; Part I; Lecture Notes in Computer Science; Hou, U.L., Spaniol, M., Sakurai, Y., Chen, J., Eds.; Springer: Cham, Switzerland, 2021; Volume 12858, pp. 227–241.
32. Shi, L.; Du, J.; Kou, F. A sparse topic model for bursty topic discovery in social networks. *Int. Arab J. Inf. Technol.* **2020**, *17*, 816–824. [[CrossRef](#)]

33. Wang, C.; Blei, D.M. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In Proceedings of the International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1982–1989.
34. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101* (Suppl. S1), 5228. [[CrossRef](#)] [[PubMed](#)]
35. Mimno, D.; Wallach, H.M.; Talley, E.; Leenders, M.; Mccallum, A. Optimizing semantic coherence in topic models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, UK, 27–31 July 2011; pp. 262–272.
36. Yao, L.; Zhang, Y.; Wei, B.; Qian, H.; Wang, Y. Incorporating Probabilistic Knowledge into Topic Models. In Proceedings of the Advances in Knowledge Discovery and Data Mining—19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, 19–22 May 2015; Part II; pp. 586–597. [[CrossRef](#)]
37. Arora, S.; Ge, R.; Halpern, Y.; Mimno, D.; Moitra, A.; Sontag, D.; Wu, Y.; Zhu, M. A Practical Algorithm for Topic Modeling with Provable Guarantees. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 280–288.
38. Li, C.; Wang, H.; Zhang, Z.; Sun, A.; Ma, Z. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In Proceedings of the International Acm Sigir Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, 17–21 July 2016; pp. 165–174.
39. Allahyari, M.; Kochut, K. Automatic Topic Labeling Using Ontology-Based Topic Models. In Proceedings of the IEEE International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, 9–11 December 2015; pp. 259–264.
40. Huang, J.; Peng, M.; Wang, H.; Cao, J.; Gao, W.; Zhang, X. A probabilistic method for emerging topic tracking in Microblog stream. *World Wide-Web-Internet Web Inf. Syst.* **2017**, *20*, 325–350. [[CrossRef](#)]
41. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
42. Bollegala, D.; Hayashi, K.; Kawarabayashi, K. Think Globally, Embed Locally—Locally Linear Meta-embedding of Words. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, 13–19 July 2018; pp. 3970–3976. [[CrossRef](#)]
43. Zhang, P.; Wang, S.; Li, D.; Li, X.; Xu, Z. Combine Topic Modeling with Semantic Embedding: Embedding Enhanced Topic Model. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 2322–2335. [[CrossRef](#)]
44. Li, S.; Pan, R.; Luo, H.; Liu, X.; Zhao, G. Adaptive cross-contextual word embedding for word polysemy with unsupervised topic modeling. *Knowl. Based Syst.* **2021**, *218*, 106827. [[CrossRef](#)]
45. Inoue, S.; Aida, T.; Komachi, M.; Asai, M. Modeling Text using the Continuous Space Topic Model with Pre-Trained Word Embeddings. In Proceedings of the ACL-IJCNLP 2021 Student Research Workshop, ACL 2021, Online, 5–10 July 2021; Kabbara, J., Lin, H., Paullada, A., Vamvas, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 138–147.
46. Gupta, P.; Chaudhary, Y.; Schütze, H. Multi-source Neural Topic Modeling in Multi-view Embedding Spaces. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, 6–11 June 2021; Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4205–4217.