Research Paper

JIS

Journal of Information Science
1–16

DOI: 10.1177/0165551520917120
journals.sagepub.com/home/jis
SAGE

SBTM: A joint sentiment and behaviour topic model for online course discussion forums

Xian Peng
College of Education, Zhejiang University, China

Qinmei Xu
College of Education, Zhejiang University, China

Wenbin Gan
National Institute of Informatics, SOKENDAI, Japan

Abstract

Large quantities of textual posts are increasingly generated in course discussion forums, and the accumulation of these data greatly increases the cognitive loads on online participants. It is imperative for them to automatically identify the potential semantic information derived from these textual discourse interactions. Moreover, existing topic models can discover the latent topics or sentimental polarities from textual data, but these models typically ignore the interactive ways of discussing topics, thus making it difficult to further construct topics' semantic space from the perspective of document generation. To solve this issue, we proposed a joint sentiment and behaviour topic model called SBTM, which was an unsupervised approach for automatic analysis of learners' discussed posts. The results demonstrated that SBTM was quantitatively effective on both model generalisation and topic exploration, and rich topic content was qualitatively characterised. Furthermore, the model can be potentially employed in some practical applications, such as information summarisation and behaviour-oriented personalised recommendation.

Keywords

Discussion forums; sentiment and behaviour topic extraction; topic model

1. Introduction

Small Private Online Courses (SPOCs) [1] have provided a feasible platform for blended learning that integrates classroom teaching and online teaching. This new practical mode is generally favoured and recognised by colleges and universities [2]. Discussion forums are the primary scenarios of communication for participants (learners, TAs and instructors) to share views and express opinions in SPOCs [3]. With the ever-growing number of posts in discussion forums, it is difficult for participants to quickly locate and track their focused textual content. For learners, they want to be aware of topic interests in the current phase of course discussions and find some posts similar to their personal thoughts and behavioural patterns. For instructors, they want to identify learners' reported topics about the technical issues, course logistics and course content discussions (e.g. audio-visual glitches, exam deadline, professional knowledge), as well as these topics' sentimental polarities, in order to conduct teaching interventions for learners. Thus, an appropriate approach should be exploited to automatically detect the potential semantic information hidden from the textual content in online learning communities.

One of the primary tasks for semantic understanding is to discover topics learners are concerned about in the process of discussion activities. When evaluating these topics, learners tend to express their opinions from both positive and negative

Corresponding author:
Xian Peng, College of Education, Zhejiang University, Hangzhou 310058, China.
Email: px87374006@126.com

P1: 第一次接触数据结构课程的相关概念，时间复杂度存在较多的疑点，怎么计算时间复杂度的问题？比较难懂，烦啊！ #发帖
It is the first time to contact the related concepts of data structure curriculum, there are some doubts on time complexity. How to compute the problem of time complexity? It's hard to understand, boring! #Post

P2: 我觉得时间复杂度的计算挺简单的啊，也就是最里层的语句的总运行次数，则算法的时间度量则由大O表示，多做几个试题，你就懂了！#回帖
I see, the calculation of time complexity is very simple, that is the executed number of basic statements. The time complexity of algorithm is represented as O. Do more tests, you will understand! # Reply

Figure 1. Two examples of posts by two learners from the course discussions. They express the same topic through different behaviours and opinions.

aspects, for example, 80% positive and 20% negative. In this case, the distribution of sentiments for learners' focused topics should also be discerned. Previous studies [4–7] have shown that compared with the single topic model latent Dirichlet allocation (LDA), the unified language models of topic and sentiment present more advantages in guiding the generative process of documents such as blogs and reviews. Furthermore, in course forums, textual content is closely associated with their interactive discussion behaviours (e.g. posting, replying and forwarding). These behavioural data generally exhibit different behavioural interaction patterns that might reveal learners' intrinsic behaviour motivation. Goffman [8] pointed out that people usually presented themselves to different images through different types of behaviours. As shown in Figure 1, although P1 and P2 are both about the time complexity of data structure, the obvious differences are observed in their sentimental and behavioural tendencies towards the topic. The former is expressed with positive emotion in a way of 'posting'. The latter is involved with negative emotion in a way of 'replying', which indicates the passive attitude of the learner for the specific knowledge. Thus, for the latter, instructors should take appropriate interventions to help the learner adjust his or her mentality and overcome difficulties. From the perspective of teaching practice, instructors and managers not only want to discover the implied topics derived from learners' discussed posts, but also want to reveal the associated opinions and behaviours of topics.

Existing topic models are mainly applied in the field of business intelligence [5,9–12] and social media [4,13–15], which may be inappropriate with online learning platforms. First, since the character length of each post or review in social platforms (e.g. Twitter, Weibo) is limited and short, some topic models (e.g. sentence-latent Dirichlet allocation (SLDA) [16] and Twitter-Latent Dirichlet Allocation (TLDA) [17]) have weakened the assumptions of LDA that takes sentences or documents as the minimum input unit in topic sampling. Nevertheless, in SPOCs, the character length of each post is relatively long, which might involve multiple topics. Second, for the purpose of providing learners with personalised learning services, our model assumes that the learner layer is considered as the input object, instead of the document layer. Last, although most supervised topic models have shown to be effective in semantic extraction [18,19], constructing labelled corpora and manually annotating data will greatly increase teaching cost and workload. Moreover, the labelling samples of course-independent will lead to a serious effect reduction across courses.

Based on the aforementioned analysis, this article proposes an unsupervised topic model by combining sentimental and behavioural interactions with posts for learner-level topic mining in online learning platforms. In particular, taking into account of interdependencies among topics, sentiments, behaviours and words at learner level, we develop a unified probabilistic topic model called sentiment and behaviour topic model (SBTM). By inferring SBTM, the resulting language model is characterised from not only the probability distribution over topics of each learner's posts and the probability distribution over words for topics, but also the sentimental and behavioural tendencies towards topics. We evaluate SBTM model on a real-life dataset from course discussion posts. The experimental results demonstrate that the prediction ability of SBTM and the effectiveness of topic exploration are better than the other unsupervised topic models.

The contributions of this study can be described as follows:

- We put forward a joint SBTM to discover potential topics, and the sentiment and behaviour distribution over topics derived from course posts in the online learning platform.
- We use Gibbs sampling to estimate the hidden model parameters.
- Experiments on the real dataset show that SBTM can achieve better results in the quality of topic generation, and it also has the advantage of explaining learners' profiles in the online learning platform.

2. Related work

In this section, we introduce the related studies on single topic models without considering the extra variables in guiding the process of topic generation, and we also discuss the complex topic models incorporating the corresponding semantic and interactive features (e.g. sentiment and behaviour) derived from learners-generated textual content.

Undoubtedly, Latent Dirichlet Allocation (LDA) has been utilised as a mainstream approach for topic modelling [20]. Some studies [16,17,21–25] tended to weaken the original hypothesis of LDA and combined the meta-data of text set such as author, title, location, link in guiding the generative process of language model. For example, Jo and Oh [16] introduced a probabilistic topic model sentence-latent Dirichlet allocation (SLDA) that posited the words in each sentence 'shared' the same topic. Zhao et al. [17] considered that the length of each post on Twitter was usually short, and each post would not cover multiple interactive topics. In this case, they proposed Twitter-Latent Dirichlet Allocation (TLDA) that assumed words in each post were assigned to a local topic. Rosen-Zvi et al. [25] integrated the author information of document to put forward a novel author-topic model (ATM). This model constrained that each document was depicted by several authors and each author followed a multinomial distribution over topics. Unlike ATM, SBTM proposed by our study focus on the topic distribution with the learner level as the model input, rather than the document level. Moreover, SBTM adjusts the constraint that each document or post is distributed by only one author, which is more consistent with the discussion forums in e-learning environments.

However, without considering the distinct characteristics of sentiment or behaviour associated with textual content, it is insufficient to explain the semantic structure and connotation of topics. Since 2007, researchers have gradually integrated sentimental information into topic model in the field of business intelligence. They demonstrated that the informative topics could be detected, and the topics were closely coupled with sentiments [4–7,16,26–30]. For example, Mei et al. [26] first taken sentiment feature as an independent unknown variable to construct a novel model topical-sentiment mixture (TSM). The model assumed that each word came from either an emotion model or a topic model. Lin and He [27] considered the interdependence of topic and sentiment and proposed a four-layer joint sentiment-topic (JST) model. When generating a word in a document, the sentiment label of the word was sampled first and then the word was sampled from the distribution of sentiment topic. Jo and Oh [16] based on the JST Model and devised an aspect and sentiment unification model (ASUM). This model weakened the original hypothesis of LDA and pointed out that sentence was the smallest unit in the document. In the field of education, Liu et al. [31] hypothesised that each sentence associated with the corresponding topic and emotion labels was regarded as the basic model input. They proposed an emotion-oriented topic model (EOTM) that captured different pairs of emotion-topic over words to identify students' emotions towards the focused aspects in course forums. Along this line, Liu et al. [32] incorporated the time and emotion characteristics of course posts into the probability topic model and proposed an unsupervised temporal emotion-aspect model (TEAM). The model could detect what students are concerned about over time in online course forums. Different from these sentiment-topic models, our model targeted at the exploitation of various pairs of topic-sentiment over the distribution of words from textual content. That is, our model can discern the sentimental intensities of different topics, not just the sentimental polarities of topics. Ramesh et al. [33] devised a weakly supervised joint model of topic and sentiment, which modelled the dependencies between topic interests and sentiment in online course discussion forums. Although this method was utilised in online learning scenarios, they did not adopt an unsupervised method based on topic modelling.

In addition, to better interpret the users' interactive content, several studies have integrated the behavioural characteristic into the generative process of topic model [4,34,35]. Xu et al. [34] modelled user posting behaviour on the social media platform, Twitter. They integrated a mixture topic model by combing three aspects: breaking news, posts from social friends and users' inherent interests. Qiu et al. [35] put forward an LDA-based behaviour topic model (B-LDA), which could jointly reveal users' topic interests and behavioural patterns in Twitter. However, sentiment considered as a distinct feature is typically ignored in constructing users' topic profiles. Based on the emoticons and the personal characteristics of micro-bloggers, Huang et al. [4] constructed a multi-modal joint sentiment topic (MJST) model for topic recognition and emotion analysis in microblogging. Like our model, MJST integrates multiple features derived from textual information for the model of topic and sentiment, whereas it ignores the importance of users' social behaviours in explaining how they interact with various pairs of topic and sentiment.

The topic models mentioned above all deal with social texts such as product/restaurant/movie reviews, twitter, blogs, other than forum posts. Notably, to the best of our knowledge, embedding sentimental and behavioural features simultaneously into the generative process of topic modelling has not yet been exploited in online learning platforms.

3. Sentiment and behaviour topic model

In this section, we first propose a unified SBTM that simultaneously incorporates sentimental and behavioural features over posts in guiding the generation of the whole corpus. Second, we display the process of model inference by Gibbs

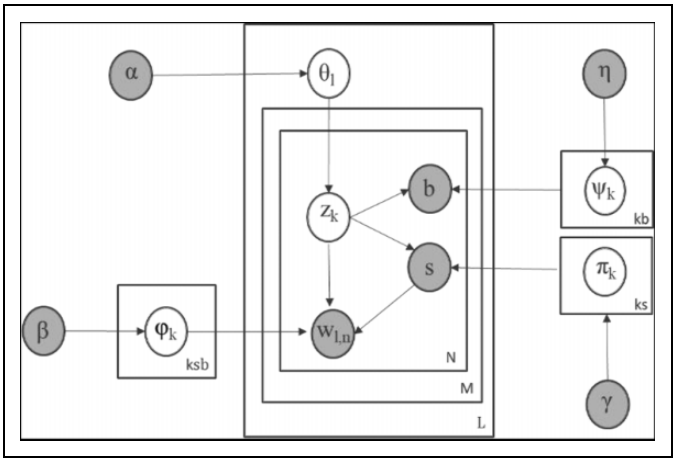

Figure 2. Graphical representation of SBTM model.

sampling and present the calculation formula of the unknown parameters. Finally, we depict the process of model algorithm in detail.

3.1. Model description

The probability topic model can be understood as a formal mathematical representation that simulates the physical process of document generation. Starting from the standard topic model LDA, it is a three-layer Bayesian probability graph model composed of document, topic and word. The basic notion is that each document draws the multinomial distribution of a mixture of several topics, and each topic is attached to the multinomial distribution over words.

By extending LDA, we propose the SBTM, as shown in Figure 2. It is an unsupervised machine learning algorithm that automates the extraction of underlying semantic information, such as topics, from large amounts of textual content. SBTM merges both sentimental and behavioural features into LDA and enhances the generative process of posts when assigning words to the corresponding topic tags. The framework of SBTM consists of four hierarchical layers (i.e. learner layer, topic layer, sentiment-behaviour layer and word layer), which constructs a mapping correlation between topic and sentiment-behaviour layer. The words in a post are closely attached to a pair of topic-sentiment and a pair of topic-behaviour. As with LDA, the common hypotheses are that there are K fixed topics, and each topic has a probability distribution of words. However, SBTM constrains that the posts from each learner serve as the minimum input unit of the model, instead of document level. Moreover, SBTM imposes a constraint that each topic is sampled from learner-topic distribution, rather than document-topic distribution. Therefore, our model can capture the implicit topics discussed by learners and uncover how learners express different attitudes and behaviours towards the topics. In addition, as a novel topic model, SBTM can be generalised to other practical contexts for business intelligence and social interaction, such as product reviews, blogs, twitter.

Specifically, SBTM assumes that when sampling one word from a post published by the learner, a multinomial distribution over topics is first determined. Then the topic is randomly selected from the distribution and draws a Bernoulli distribution over sentiments (positive sentiment and negative sentiment) and a multinomial distribution over behaviours. Last, the word is chosen from the topic combining the sentimental and behavioural tags.

As shown in Figure 2, the probability graph model SBTM is essentially a directed acyclic graph based on Bayesian network, where the nodes represent random variables. The solid circles represent known observed variables, such as words in the document; hollow circles represent unknown hidden variables, such as topics hidden in the document. The directed arrows indicate the conditional dependence between variables, such as the unknown topic variable z_k that is related to the behaviour variable b and sentiment variable s. Table 1 presents the symbolic variables of SBTM.

The specific generation process of SBTM is as follows. First of all, it is assumed that a total number of L learners participate in the discussion forum. Each learner publishes M_l course posts, and each post is denoted as $R = \{r_1, r_2, \ldots, r_m\}(1 \leq m \leq M)$. In SBTM, each post is made up of a series of potential topics $Ln = \{z_1, z_2, \ldots, z_K\}$. Second, topic is closely related to sentiment s and behaviour b. Each topic draws the binomial sentiment distribution π_{ks}, and each sentiment topic z_k follows the multinomial behaviour distribution ψ_{kb}. When randomly sampling a word in a post, the word may come from a topic word or a sentiment word. Then assigning the topic label of the word, and the topic has the probability distribution over word φ_{kw}. Since this topic is associated with

Table 1. Description of notations in SBTM.

Notions	Description	Notions	Description
L	Number of learners	α	Dirichlet prior of θ
M	Total number of posts	β	Dirichlet prior of φ
N	Number of words	η	Dirichlet prior of ψ
K	Number of topics	γ	Dirichlet prior of π
$\mathbf{z}$	Topic index	θ	Distribution of learner-topic
s	Sentiment index	φ	Distribution of topic-word
b	Behaviour index {TP, RE, QU, CP}	ψ	Distribution of topic-behaviour
$\mathbf{w}$	Word index	π	Bernoulli distribution of topic-sentiment

TP: thread posting; RE: replying; QU: quoting; CP: common posting; SBTM: sentiment and behaviour topic model.

sentiment, behaviour and word, this topic composed of a set of words can be represented as $z_{ksb} = \{w_1, w_2, \ldots, w_n\}(1 \leq n \leq N)$. When k is equal to a fixed value, sampling each word for a specific topic iteratively in corpus until SBTM reaches a stable state. Finally, SBTM assumes that the prior distributions θ_{lk}, φ_{kw}, π_{ks} and ψ_{kb} are α, β, γ and η, respectively.

3.2. Model inference

In SBTM, Gibbs sampling considered a random simulation approach is adopted to conduct iterative sampling of the whole corpus. According to the corresponding dependencies among the observed and hidden variables in Figure 2 and the probability graph theory, a joint probability distribution formula integrating the multiple variables is represented as

$$p(w, b, s, z|\alpha, \beta, \gamma, \eta) = p(w|z, s, \beta) \cdot p(s|z, \gamma) \cdot p(b|z, \eta) \cdot p(z|\alpha) \tag{1}$$

The factors on the right side of the above formula are respectively derived into

$$\begin{aligned} p(z|\alpha) &= \int p(z|\theta)p(\theta|\alpha)d\theta \\ &= \int \prod_{l}^{L} \frac{1}{\Delta\alpha} \prod_{k=1}^{K} \theta_{l,k}^{n_l^{(k)+\alpha_{k-1}}} d\theta_l \\ &= \prod_{l=1}^{L} \frac{\Delta\left(n_m^{(k)} + \alpha\right)}{\Delta\alpha} \end{aligned} \tag{2}$$

$$\begin{aligned} p(b|z, \eta) &= \int p(b|z, \psi)p(\psi|\eta)d\psi \\ &= \int \prod_{z=1}^{K} \frac{1}{\Delta\eta} \prod_{b=1}^{B} \psi_{z,b}^{n_z^{(b)+\gamma_{b-1}}} d\psi_z \\ &= \prod_{z=1}^{K} \frac{\Delta\left(n_z^{(b)} + \eta\right)}{\Delta\eta} \end{aligned} \tag{3}$$

$$\begin{aligned} p(s|z, \gamma) &= \int p(s|z, \pi)p(\pi|\gamma)d\pi \\ &= \int \prod_{z=1}^{K} \frac{1}{\Delta\gamma} \prod_{s=1}^{S} \pi_{z,s}^{n_z^{(s)+\gamma_{s-1}}} d\pi_z \\ &= \prod_{z=1}^{K} \frac{\Delta\left(n_z^{(s)} + \gamma\right)}{\Delta\gamma} \end{aligned} \tag{4}$$

$$
\begin{aligned}
p(w|z,s,\beta) &= \int p(w|z,s,\varphi)p(\varphi|\beta)d\varphi \\
&= \int \prod_{z=1}^{K} \frac{1}{\Delta\beta} \prod_{w=1}^{V} \varphi_{z,w}^{n_z^{(w)}+z_{w-1}} d\varphi_z \\
&= \prod_{z=1}^{K} \frac{\Delta\left(n_z^{(w)}+\beta\right)}{\Delta\beta}
\end{aligned}
\tag{5}
$$

Then, each word in the textual sets based on the updating rule of Gibbs sampling is randomly sampled, which assigns the corresponding topic label for the current word z_{-i} through the co-occurrence frequency distribution of other words. The calculation formula is given by

$$
\begin{aligned}
&p(z_i, s_i, b_i | z_{-i}, s_{-i}, b_{-i}, w_{-i}, \alpha, \beta, \gamma, \eta) \\
&= \frac{p(z, s, b, w, \alpha, \beta, \gamma, \eta)}{p(z_{-i}, s_{-i}, b_{-i}, w_{-i}, \alpha, \beta, \gamma, \eta)} \\
&\propto \frac{B(n_{l,k}+\alpha)}{B(n_{l,k,-i}+\alpha)} \cdot \frac{B(n_{z,w}+\beta)}{B(n_{z,w,-i}+\beta)} \cdot \frac{B(n_{z,s}+\gamma)}{B(n_{z,s,-i}+\gamma)} \cdot \frac{B(n_{z,b}+\eta)}{B(n_{z,b,-i}+\eta)}
\end{aligned}
\tag{6}
$$

$$
B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}, \ \ \Gamma(x+1) = x\Gamma(x), x > 0 \tag{7}
$$

Thus, the posterior distribution formula is jointly demonstrated as

$$
\begin{aligned}
&p(z_i = k \mid z_{-i},\ w_{-i},\ s_{-i},\ e_{-i}) \\
&\propto \frac{N_{lk,-i}^{LK} + \alpha}{\sum_{k=1}^{K} N_{lk,-i}^{Lk} + K \cdot \alpha} \ \frac{N_{kw,-i}^{KW} + \beta}{\sum_{w=1}^{V} N_{kw,-i}^{KW} + V \cdot \beta} \ \frac{N_{kjw,-i}^{KSW} + \gamma}{\sum_{s=1}^{S} N_{kjw,-i}^{KEW} + E \cdot \gamma} \ \frac{N_{kbw,-i}^{KBW} + \eta}{\sum_{b=1}^{B} N_{kbw,-i}^{KBW} + B \cdot \eta}
\end{aligned}
\tag{8}
$$

where z_i denotes the topic of a word w_i in the post and z_{-i} denotes the topics of the other words except for the current word w_i. E represents the sentiment distribution of a pair of topic and word, and B represents the behaviour distribution of a pair of topic and word. N_{lk}^{LK} indicates the total number of words that are assigned with topic k by the learner l. N_{kw}^{KW} indicates the total number of words that are assigned with topic k. N_{kjw}^{KSW} indicates the total number of words that are assigned with topic k and sentiment j. N_{kbw}^{KBW} indicates the total number of words that are assigned with topic k and behaviour b.

3.3. Parameter estimation

Through iterating the process of Gibbs sampling, SBTM can reach a convergence state. As a result, four unknown parameters leaner-topic θ_{lk}, topic-word φ_{kw}, topic-sentiment ψ_{kb} and topic-behaviour π_{ke} are approximately computed as follows

$$
\theta_{lk} = \frac{N_{lk}^{LK} + \alpha}{\sum_{k=1}^{K} N_{lk}^{Lk} + K \cdot \alpha} \tag{9}
$$

$$
\varphi_{kw} = \frac{N_{kw}^{KW} + \beta}{\sum_{w=1}^{V} N_{kw}^{KW} + V \cdot \beta} \tag{10}
$$

$$
\pi_{kjw} = \frac{N_{kjw}^{KSW} + \gamma}{\sum_{s=1}^{S} N_{kjw}^{KEW} + E \cdot \gamma} \tag{11}
$$

$$
\psi_{kbw} = \frac{N_{kbw}^{KBW} + \eta}{\sum_{b=1}^{B} N_{kbw}^{KBW} + B \cdot \eta} \tag{12}
$$

Algorithm 1. SBTM algorithm.

Input: the posts corpus $|P_M| = \{p_1, p_2, \ldots, p_m\}(1 \le m \le M)$, initial parameters α, β, γ and η, number of iterations 500, number of topics K.
Output: learner-topic θ_{lk}, topic-word φ_{kw}, topic-sentiment π_{ke}, topic-behavior ψ_{kb}.
1: Randomly assign the corresponding topic k associated with sentiment s and behavior b for each word w, initial matrix θ_{lk}, φ_{kw}, π_{ke}, ψ_{kb}.
2: **for** each topic z_k discussed by learners, k={1, … , K}
3: draw a multinomial distribution of topic-word: $\varphi_k \sim Dir(\beta)$
4: draw a multinomial distribution of topic-behavior: $\psi_{kb} \sim Dir(\eta)$
5: draw a binomial distribution of topic-sentiment: $\pi_{ks} \sim Dir(\gamma)$
6: **end for**
7: **for** each learner $l = \{1, 2, \ldots, L\}$
8: Draw a multinomial distribution of learner-topic: $\theta_{lk} \sim Dir(\alpha)$
9: **for** each word n the learner post, $n = \{w_1, w_2, \ldots w_{lm}\}$
10: sample a topic z_{lm}, $z_{lm} \sim Multi(\theta_{l_k})$
11: sample a behavior label b_{lm} for z_{lm}, $b_{lm} \sim \psi_{z_{lm}}$
12: sample a sentiment label s_{lm} for z_{lm}, $s_{lm} \sim \psi_{z_{lm}}$
13: sample a word w_{lmn} coupled with a pair of s_{lm} and b_{lm}, $w_{lmn} \sim Multi(\varphi_{z_{l_m}})$
14: **end for**
15: **end for**
16: Using Gibbs Sampling to update the unknown parameters of SBTM θ_{lk}, φ_{kw}, π_{ke}, ψ_{kb}.
17: Repeating until the of optimization of SBTM.

SBTM: sentiment and behaviour topic model.

In this section, with the help of equations (9)–(12), we can not only uncover the probability distribution of learners' focused topics as well as the fine-grained semantic vectors, but also detect their opinions and interactions towards the topics.

3.4. Model algorithm process

The modelling process of SBTM algorithm can be generally divided into the following three steps: the first step is to input the posts set and the known parameters of model; the second step is to assign each word with a topic label involving the sentimental and behavioural properties by using Gibbs sampling; the last step is to compute the unknown parameters when the sampling results are stable. The detailed introduction of SBTM algorithm is shown in Algorithm 1.

4. Experiment settings

4.1. Dataset and pre-processing

The data of this study was collected from a Chinese university's cloud classroom called starC (http://spoc.ccnu.edu.cn/starmoocHomepage) that was a hybrid SPOC platform. Under the guidance of network information security and personal information ethics, learners' online click behaviours and abundant textual data in discussion forum were recorded and stored in the databases. In the experiment, 30 online courses (15 courses from arts and sciences, respectively) discussed highly in the next semester of 2016 were selected as the analysis dataset. The statistics of the dataset are shown in Table 2.

The main steps of data pre-processing were as follows. First, the Chinese word segmentation tool called ICTCLAS [36] was adopted to split each post into a set of non-redundant words, and the four types of vocabulary categories including adjectives, adverbs, verbs and nouns were only retained. Second, the stop words (such as '和/and', '他/he', '这样/so'), the low-frequency words (if the number of occurrences of a word is less than 5), the noise characters (such as starting with http, www) and some characters that did not conform to the provisions of the format character label (e.g. [font] [b] [size = 12pt] [img]) were all removed. Finally, we took Gibbs sampling to match the corresponding sentimental and behavioural labels for the filtered words. Notably, we adopted java programming to operate the course post datasets automatically and parsed them into the generation process of topic modelling. Moreover, we employed java programming to perform topic extraction, sentiment and behaviour analysis.

Table 2. Statistics of the experimental dataset.

Number of courses	Number of learners	Number of posts	Number of characters	Avg. number of characters per post	Avg. number of posts per learner	Avg. number of posts per course
30	1527	15,357	4,167,245	324.36	10.06	511.90

Table 3. Category coding of learners' discourse behaviours.

Label	Behavioural categories	Basic classification of behavioural categories
TP	Thread posting	The post is distributed by a learner to launch a new thread (including popular topics, unresolved issues, etc.)
RE	Replying	The post that contains '/reply', '@' or 'mention' is published to answer the others' messages
QU	Quoting	The post that involves 'quote', 'blockquote', 'http' or other similar symbols is used to verify someone's views
CP	Common posting	Common posting is a special behaviour of replying and is not a direct response towards the others' messages

4.2. Sentiment lexicon and semantic rules

A Chinese sentiment lexicon was constructed from three types of authoritative collection of emotional vocabularies, including a Chinese sentimental dictionary [37], HowNet [38], and a Chinese commendatory and derogatory dictionary v1.0 [39]. In total, there were 13,267 negative and 9725 positive terms. When sampling a word label, if it was included in the sentiment lexicon, the word would be assigned with the corresponding sentiment property. Otherwise, it was randomly given with a sentiment token.

To ensure more accurate internal information of topics, this study designed some semantic rules considering the context polarity and dependency of words in SPOCs. For the former, if the preceding sequence of the positive sentiment word $p1$ was a negative word $n1$, $p1$ and $n1$ would be jointly represented as a negative sentiment word. For the latter, if the preceding sequence of the adjective was an adverb, they would be jointly combined with a new word expressed as 'adv + adj'.

4.3. Behaviour categories

According to the characteristics of learners' discourse interaction behaviours in the SPOC discussion forum, this study defined four types of discourse behaviours, including thread posting (TP), replying, quoting and common posting. As shown in Table 3, the coding of learners' discussion behaviour categories and the classification criteria of specific behaviour categories are described. This article pointed out that a post distributed by learners only corresponded to one behaviour category. When sampling the words of each post, these words were assigned with the same behaviour label.

4.4. Parameters setting and comparison of models

In our experiment, the known hyper-parameters α, β, γ and η were empirically set as 0.1, 0.01, 0.1 and 0.01, respectively. The number of words N was computed as 969,127. In addition, the number of iterations of the model was uniformly set as 500. In the last 400 iterations, we selected 200 samples with a gap of 5 to assign values to all the latent parameters θ_{lk}, φ_{kw}, ψ_{kb} and π_{ke}.

A total of three tasks were determined to evaluate the performance of SBTM model quantitatively and qualitatively. The first task used the perplexity index [12] to obtain the optimal number of topics and to ensure the best generalisation ability of model. The second task found out the topics learners are more interested in, and how they expressed the sentiments and behaviours towards the topics. The third task compared SBTM with the four unsupervised topic models (i.e. SLDA [16], TLDA [17], learner-oriented L-SLDA [25] and document-oriented D-TLDA) in terms of perplexity of the model, similarity between topics [26] and information entropy [35].

4.5. Evaluation metrics

Based on the method of internal evaluation of model, three indexes including perplexity, similarity and entropy are adopted. The detailed formulas are presented in the following.

Perplexity is used to evaluate the generalisation ability of the model for unknown data processing by calculating the test set of corpus data. The smaller perplexity value is preferred and the specific formula is given by

$$Perplexity(D_{test}|\text{Model}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{13}$$

where M is the total number of learners, w_d is the joint probability value of the words in the generated document d and N_d is the number of words in the post d by the learner m.

Similarity is adopted to distinguish the degree of similarity between topics grouped by the model. Like LDA, if the similarity value between the topic-word is lower, the discrimination ability between topics will be higher. We also computed the similarity of various topics related to the sentiments and behaviours. The following three formulas indicate the similarity between topics and words, the similarity between topics and sentiments and the similarity between topics and behaviours, respectively

$$Sim(z_i, z_j) = \frac{1}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\sum_{w=1}^{V}\varphi_{iw}\varphi_{jw}}{\sqrt[2]{\left(\sum_{w=1}^{V}\varphi_{iw}^2\right)\left(\sum_{w=1}^{V}\varphi_{jw}^2\right)}} \tag{14}$$

$$Sim(z_i, z_j) = \frac{1}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\sum_{s=1}^{S}\pi_{is}\pi_{js}}{\sqrt[2]{\left(\sum_{s=1}^{S}\pi_{is}^2\right)\left(\sum_{s=1}^{S}\pi_{js}^2\right)}} \tag{15}$$

$$Sim(z_i, z_j) = \frac{1}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\sum_{b=1}^{B}\psi_{ib}\psi_{jb}}{\sqrt[2]{\left(\sum_{b=1}^{B}\psi_{ib}^2\right)\left(\sum_{b=1}^{B}\psi_{jb}^2\right)}} \tag{16}$$

Entropy is used to measure the degree of consistence and coherence of words within the topic generated by the evaluation model. Like LDA, if the entropy value is lower, the consistency of words within the topic is higher. Moreover, it is expected to uncover a set of topics that are highly identified with the dominant sentiments and behaviours. The following three formulas indicate the entropy between topics and words, the entropy between topics and sentiments and the entropy between topics and behaviours, respectively

$$Entropy = \frac{1}{K}\sum_{k=1}^{K}\sum_{w=1}^{V}(-\varphi_{kw}\log\varphi_{kw}) \tag{17}$$

$$Entropy = \frac{1}{K}\sum_{k=1}^{K}\sum_{s=1}^{S}(-\pi_{ks}\log\pi_{ks}) \tag{18}$$

$$Entropy = \frac{1}{K}\sum_{k=1}^{K}\sum_{b=1}^{B}(-\psi_{kb}\log\psi_{kb}) \tag{19}$$

where S denotes the number of sentimental categories and B indicates the number of behavioural categories. Other symbols are defined in Section 3.1.

5. Results and discussion

5.1. Comparison of model generalisation

The number of latent topics is a crucial indicator in affecting the effectiveness and extensibility of model generation [12]. Considering the large amount of input data for the textual posts, the initial number of hidden topics of SBTM was empirically set as 20, and the number of iterations was set as 500. In each test, 20 topics were gradually increased and the prior parameters were uniformly kept fixed. When the perplexity value is smaller in a stable state, the number of topics will be selected. Figure 4 shows the change trend of perplexity value of the model with the number of $K \in \{20, 40, 60, 80, 100, 120, 140, 160\}$ in 500 iterations.

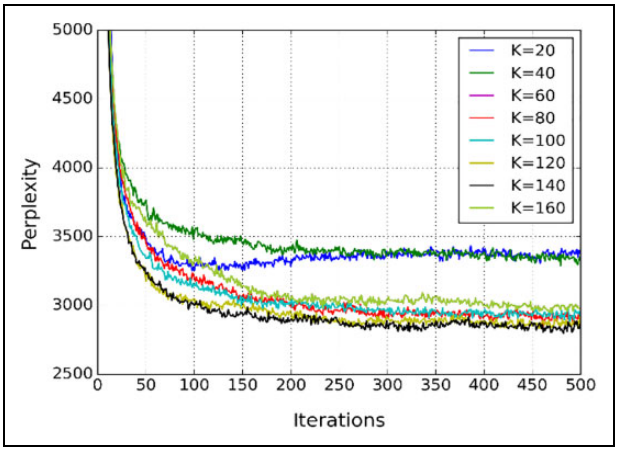

Figure 3. Perplexity values for different topics.

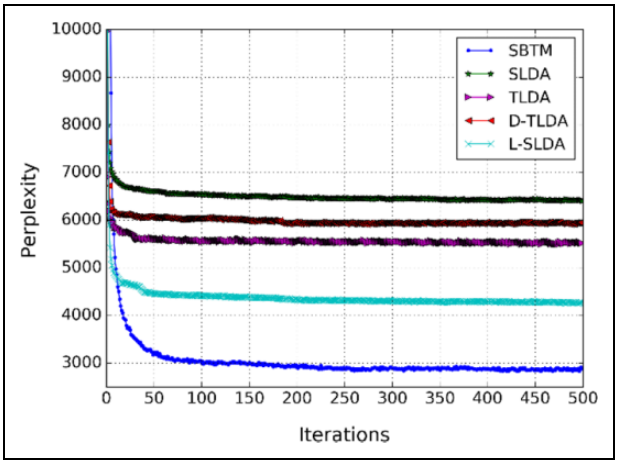

Figure 4. Comparison of the perplexity values of different topic models.

As can be seen from Figure 3, when the number of iterations of the model was within the range of {1–50}, the model perplexity values of all curves tended to change significantly and showed a direct sharp decline. However, when the number of iterations was more than 50, all the topic curves gradually converged until the entire iteration process was completed. Specifically, when the number of topics was equal to 120, the overall average perplexity value dropped to the lowest 3813.03. Furthermore, when K were set as 140 and 160, respectively, the overall average perplexity value no longer decreased and showed a slight upwards trend. This indicated that if the number of topics exceeded 120, the generalisation ability of the model would no longer increase but decrease. Therefore, the optimal preset value for determining the number of topics in this section was set as 120.

In Figure 4, with the increasing of iterations, the average perplexity value of SBTM was lower than SLDA, TLDA, L-SDA and D-TLDA, indicating that the generalisation ability of topic extraction by SBTM was better than the other unsupervised models in the online course posts dataset. The analytic result provides supplementary evidence that the multivariate topic model has a greater advantage of guiding the generation of topic content than the single topic model [6].

5.2. Topic extraction

Topic extraction is one of the primary tasks in topic modelling. Three example topics separately extracted from SBTM, SLDA and TLDA are illustrated in Table 4. Through logic reasoning, the listed topic labels were assigned with the corresponding semantic content so as to be understood easily.

From Table 4, we presented a global topic in the middle column (i.e. Topic 80, Topic 58, Topic 93, Topic 90 and Topic 70) and two local topics (i.e. Topic 44 and Topic 84, Topic 11 and Topic 15, Topic 6 and Topic 74, Topic 73 and Topic 39, Topic 39 and Topic 40) on both sides of the middle column for the five topic models. The former was mainly related to course logistic, while the latter evolved more professional knowledge and terminologies about course-specific content. The

Table 4. Examples of topics extracted from the experimental dataset.

SBTM			SLDA			TLDA		
Topic 44 Chemistry	Topic 80 Course logistic	Topic84 Linear algebra	Topic 11 Chemistry	Topic 58 Course logistic	Topic15 Linear algebra	Topic 6 Chemistry	Topic 93 Course logistic	Topic 74 Linear algebra
反应	学习	线性	反应	学生	题	反应	学生	矩阵
平衡	大学	矩阵	浓度	学习	矩阵	条件	教师	题
浓度	生活	题	平衡	生活	线性	形成	教育	证明
产物	老师	代数	速率	大学	向量	凝结	学习	线性
速率	时间	证明	产物	老师	证明	平衡	问题	空间
近似	同学	向量	温度	问题	考试	浓度	教学	因子
自由	专业	因子	近似	时间	空间	水汽	职业	向量
级数	问题	组	条件	孩子	组	近似	作业	行列式
温度	知识	空间	自由	工作	代数	空气	工作	变换
系统	社会	行列式	分子	学会	计算	产物	知识	解答
reaction	study	linear	reaction	student	item	reaction	student	matrix
balance	university	matrix	concentration	study	matrix	condition	instructor	item
concentration	life	item	balance	life	linear	formation	education	proof
product	teacher	algebra	velocity	university	vector	congeal	study	linear
velocity	time	proof	product	teacher	proof	balance	issue	space
approximate	classmate	vector	temperature	issue	exam	concentration	instruction	divisor
freedom	major	divisor	approximate	time	space	moisture	occupation	vector
series	issue	group	condition	children	group	approximate	homework	det
temperature	knowledge	space	freedom	job	algebra	air	job	transform
system	society	det	molecule	learn	count	product	knowledge	solve

L-SLDA			D-TLDA		
Topic 73 Chemistry	Topic 90 Course logistic	Topic39 Linear algebra	Topic 39 Chemistry	Topic 70 Course logistic	Topic 40 Linear algebra
反应	学生	矩阵	反应	教师	矩阵
浓度	教师	线性	表达式	学生	线性
平衡	教育	题	查询	教育	向量
速率	学习	向量	属性	知识	空间
产物	教学	证明	基准	教学	题
近似	问题	组	子句	职业	证明
自由	知识	空间	浓度	学习	组
级数	职业	因子	平衡	能力	行列式
温度	孩子	行列式	态势	专业	因子
条件	能力	相关	产物	工作	运算
reaction	student	matrix	reaction	instructor	matrix
concentration	teacher	linear	expression	student	linear
balance	education	item	inquire	education	vector
velocity	study	vector	property	knowledge	space
product	instruction	proof	criterion	instruction	item
approximate	issue	group	clause	occupation	proof
freedom	knowledge	space	concentration	study	group
series	occupation	divisor	balance	ability	det
temperature	children	det	situation	major	divisor
condition	ability	relation	product	job	compute

SBTM: sentiment and behaviour topic model; SLDA: sentence-latent Dirichlet allocation; TLDA: twitter- latent Dirichlet allocation.

10 words were listed for each example topic with different probability values in descending order. Through further observation, these words attained a high degree of internal compactness and coherence to each example topic. For example, in the column 2 of Table 4, the words like ‘study’, ‘life’, ‘time’, ‘major’ and ‘issue’ are closely tied to course logistic.

Different from SLDA, TLDA, L-SLDA and D-TLDA, SBTM can not only detect the core topics of learners’ overall discussions, but also locate the topics that the learner is interested in. Table 5 lists three examples of learner-topic probability distribution by SBTM.

Table 5. Examples of learner-topic probabilistic matrix of SBTM.

Learner	Topic	Probability	Learner	Topic	Probability	Learner	Topic	Probability
Learner 1	15	0.1259	Learner 2	15	0.1867	Learner 3	44	0.3098
	48	0.1384		38	0.0592		90	0.1665
	80	0.4683		80	0.5755		101	0.0785

SBTM: sentiment and behaviour topic model.

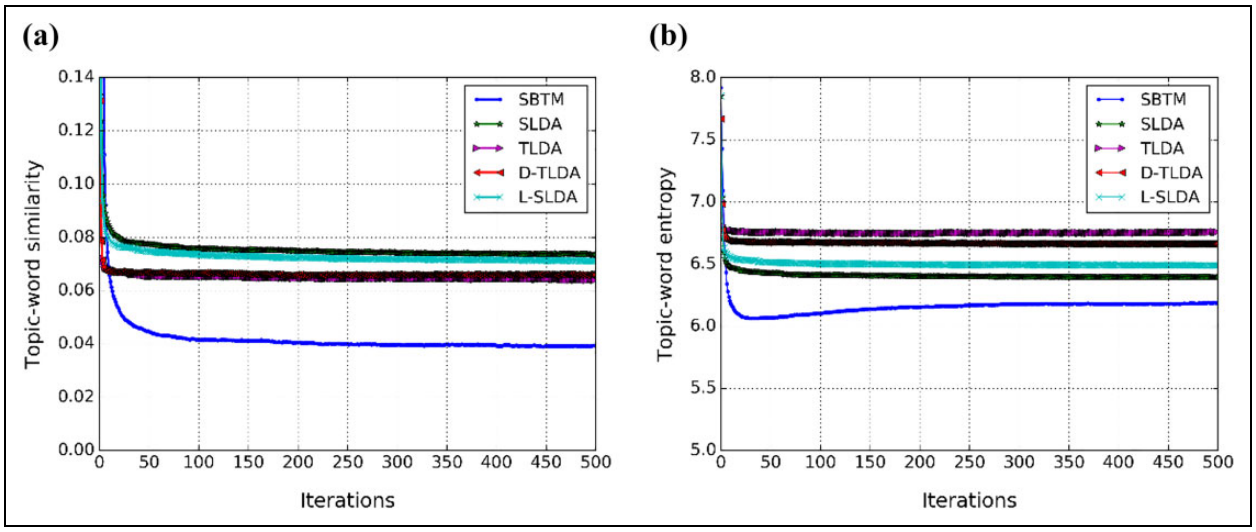

Figure 5. Comparison of the effectiveness of topic-word generation: (a) comparison on topic-word similarity and (b) comparison on topic-word entropy.

By observing the focused topics' probability distribution, learner 1 and learner 2 seemed to share similar course content like topic 15 and topic 80. However, unlike learner 1 and learner 2, learner 3 was more interested in discussing other topics. That is, different learners were interested in different topics, which might be attributed to the individual preference of learners and the arrangement of course discussions. Thus, the analytical results can provide some informative insights that are helpful for instructors to conduct personalised interventions according to the learner-topic probability distribution.

Furthermore, we conducted quantitative analysis for further model validation. The two common indexes similarity and entropy were utilised to evaluate the quality of topic generation among the topic models in Figure 5(a) and (b).

As shown in Figure 5(a) and (b), when the number of iterations was in the top 50, the similarity of topic-word and the aggregation of topic-word of each model declined rapidly and reached a relatively stable trend soon. Furthermore, SBTM model had a lower topic-word similarity value and a lower topic-word entropy value than the other four models SLDA, TLDA, L-SLDA and D-TLDA. That is, SBTM outperformed the other topic models and had better scalability (i.e. the discrimination ability between topics as well as the degree of consistence and coherence of words within the topic) in generating topic content.

5.3. Topics-specific sentimental and behavioural identification

Except for exploring the topics discussed by learners, we also put emphasis on the learners' sentimental attitudes and behavioural patterns in expressing these topics. In this section, we discovered the topics-specific sentimental and behavioural information extracted by SBTM in Tables 6 and 7. Moreover, we made a comparison of topics from SBTM, SLDA, TLDA, L-SLDA and D-TLDA in terms of similarity and entropy on sentimental distribution and behavioural distribution in Figures 6 and 7. Notably, the sentimental and behavioural variables were also embedded into the generative progress of SLDA, TLDA, L-SLDA and D-TLDA for the fair comparison with SBTM in the same dimension.

From Table 6, we illustrated the three topics with the corresponding sentimental distribution and sentiment words. If the positive probability value of the topic was higher than the negative probability value of the topic, the topic was defined with a positive label (+), versa. Intuitively, learners tended to discuss the topic from both positive and negative aspects. Taking topic 84 as an example, we found that learners held a relatively negative attitude towards linear algebra which was

Table 6. Examples of topics-specific sentimental distribution extracted by SBTM.

Topics	Sentiment distribution	Positive words		Negative words	
Topic 44 (+)	(0.5369, 0.4631)	稳定/stabilization	亲和/agreeableness	压力/pressure	复杂/complex
Chemistry		完成/finish	活泼/active	消耗/consume	过量/excess
		适用/applicable	优先/precedence	阻力/obstruction	看不到/fail to see
Topic 80 (+)	(0.6441, 0.3559)	学会/learn better	好/good	迷茫/at sea	不喜欢/not like
Course logistic		认真/earnest	喜欢/like	不适应/inadaptation	挂科/fail an exam
		勇敢/brave	热情/warm	不好/not good	不清楚/not clear
Topic 84 (–)	(0.4863, 0.5136)	行/can	帅/handsome	错/wrong	难/difficult
Linear algebra		简单/easy	帮忙/helpful	未知数/unknown	心慌/nervous
		详细/in detail	很好/very good	出错/make mistakes	不明白/do not see

SBTM: sentiment and behaviour topic model.

Table 7. Examples of topics-specific behavioural distribution extracted by SBTM.

Topics	Topic Label	Sentiment Label	TP	RE	QU	CP
Topic 44 (+)	Chemistry	Positive	0.1571	0.1003	0.0128	0.7298
Topic 80 (+)	Course logistic	Positive	0.9997	0.0001	0.0001	0.0001
Topic 84 (–)	Linear algebra	Negative	0.6322	0.0128	0.0001	0.3549

TP: thread posting; RE: replying; QU: quoting; CP: common posting; SBTM: sentiment and behaviour topic model.

'difficult', 'unknown' and 'nervous'. While this topic was classified into a negative one, learners also expressed the positive evaluation using sentiment words 'easy', 'helpful' and 'very good'. Therefore, we had reasons to believe that learners generally showed a bad experience when discussing topic 84. To sum up, our model can not only identify the emotional polarities of learners' focused topics, but also uncover the specific emotional distributions and the corresponding emotional items of these topics. This will help instructors visually and quickly locate learners' attitudes when they have different interests and face different difficulties in online course discussions.

From Table 7, it could be realised how learners interacted with the above topics. Some differences were observed in the behavioural tendencies of three topics. Topic 44 was mainly dominated by thread posting (TP), replying (RE) and common posting (CP); topic 80 was just dominated by TP; the topic was mainly dominated by TP and CP. Taking topic 84 as an example, learners tended to publish topic posts and common posts to convey their personal views and attitudes, whereas they rarely used replies or quotes to communicate with others. This indicated that when discussing topic 84, learners generally showed more initiative in participating in the course discussions. Therefore, our model can identify learners' emotional tendencies towards the focused topics as well as these topics' behavioural patterns, which will further help to understand the intrinsic motivation of learners' interactive behaviours and is also important for constructing more accurate learners' profiles.

Besides, we expected that our model could generate the high quality of topics that were dominated by sentimental and behavioural features. In this case, we conducted quantitative analysis to measure the effectiveness of topic-sentiment and topic-behaviour in terms of similarity and entropy.

From Figure 6, each curve was similar in iteration rules, and the evaluation indexes could reach a stable trend quickly. That is, when the number of iterations was in the top 50, the similarity of topic-sentiment and the aggregation of topic-sentiment of each model declined rapidly. In the rest of iterations, each model reached a relatively stable trend soon. Although SBTM was slightly higher than TLDA and D-TLDA in terms of generating topic-sentiment similarity and entropy, it still generally shows better advantages than SLDA and L-SLDA. Perhaps, the reason is that compared with the multiple word characteristics within topics, there are only two categories of emotional attribute for topics. Thus, the consistency of topic-sentiment information might be constrained and limited.

From Figure 7, when the number of iterations of each model was within the range of {1–50}, all curves on the similarity of topic-behaviour and the convergence of topic-behaviour had obvious changes, presenting a direct sharp decline. However, when the number of iterations was within the range of {51–500}, all curves gradually reached a stable trend until the entire iteration process was completed. Evidently, topics detected from SBTM model had lower similarity and entropy values than the other topic models SLDA, TLDA, L-SLDA and D-TLDA. This indicated that topics captured from SBTM tended to be characterised by some dominant types of behaviours and were distinctly discriminated. In a

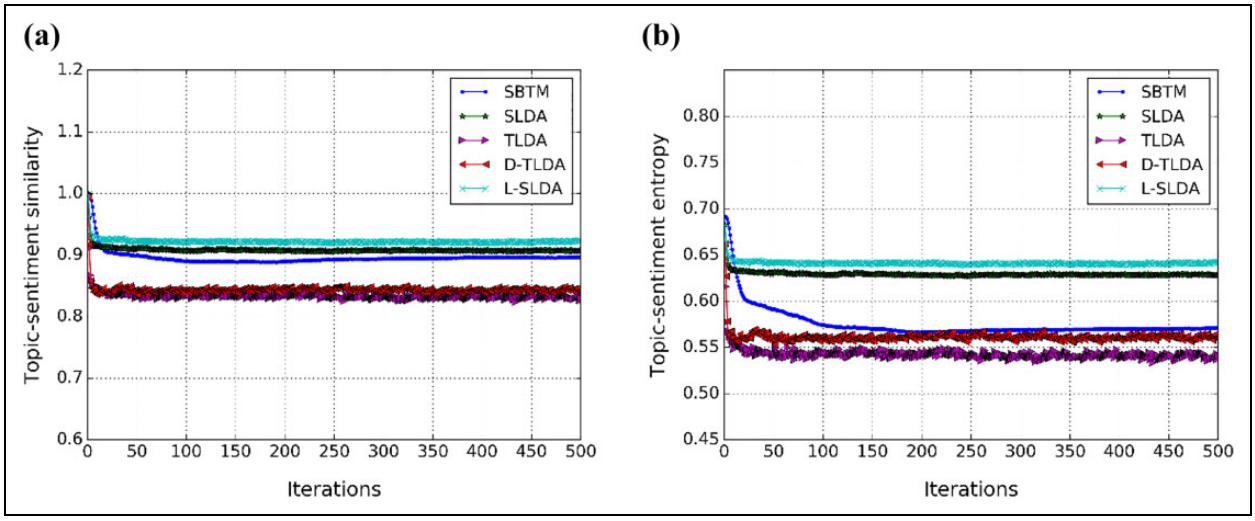

Figure 6. Comparison of topics in terms of similarity and entropy on sentiment distribution: (a) comparison on topic-sentiment similarity and (b) comparison on topic-sentiment entropy.

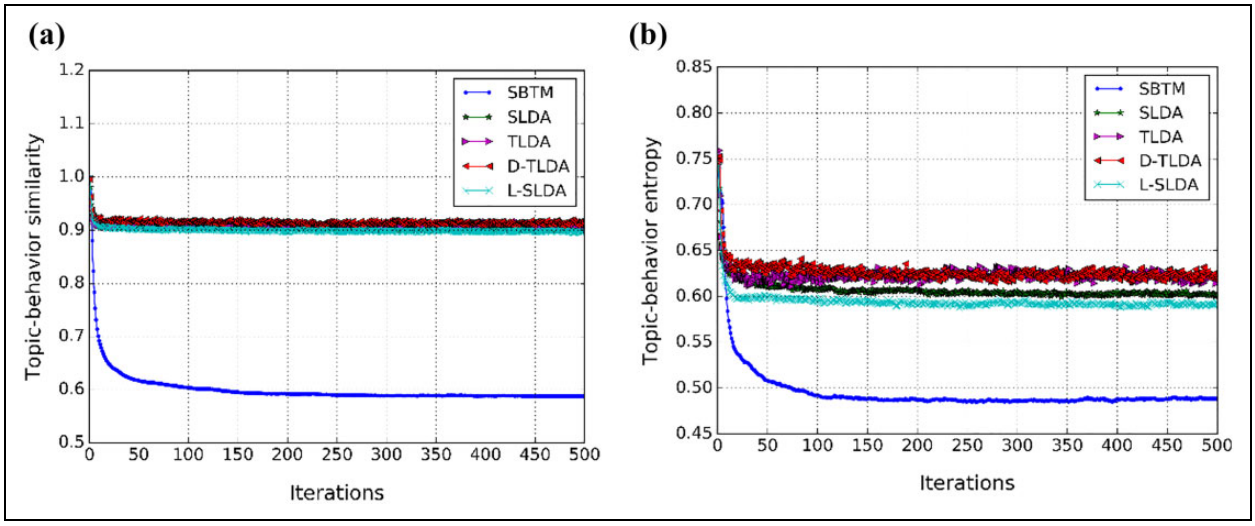

Figure 7. Comparison of topics in terms of similarity and entropy on behaviour distribution: (a) comparison on topic-behaviour similarity and (b) comparison on topic-behaviour entropy.

nutshell, SBTM significantly outperformed the other topic models in terms of topic-behaviour similarity and topic-behaviour aggregation.

6. Conclusion, limitations and future work

In this article, an unsupervised topic model named SBTM by combining the sentimental and behavioural features was proposed. We employed SBTM to automatically extract hidden semantic information derived from learner-generated posts in the discussion forum of a SPOC. In SBTM, the learner-oriented level course posts were utilised to model the process of topic generation. Each topic generated by SBTM drew the binomial sentiment distribution, and the model could compute the positive and negative sentiment intensity of each topic. The experimental results on the real course posts dataset demonstrated that SBTM could not only improve the interpretability of topic representation, but also presented better performance in generalisation ability and stronger differentiation in topic quality. Furthermore, SBTM has the potential of providing some insights into the learner' personal behaviours and discourse information, which are helpful for instructors to conduct adequate personalised interventions.

The study also has several limitations. The input data source of the model is derived from the course forum posts expressed in Chinese language; thus, the findings of the study might be constrained in the use of other languages, such as English. Moreover, as the model is designed for online learning platforms, different types of textual content (e.g. business reviews, social posts) might affect the performance of the model. In future work, we will further validate the effectiveness of SBTM and test it on rich datasets in Chinese or other languages in the context of business intelligence and social media, such as product/restaurant/movie reviews, twitter, weibo.

Acknowledgements

The authors sincerely thank Zhi Liu for his help in data collection and syntax checking.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was supported by the Research funds from the China Mobile Research Foundation of the Ministry of Education (grant no. MCM20160401) and the Research Funds from the National Natural Science Foundation of China (grant no. 61702207, L1724007).

ORCID iD


Xian Peng https://orcid.org/0000-0003-0390-3929

References

[1] Reich J. Rebooting MOOC research. *Science* 2015; 347(6217): 34–35.
[2] Fox A. From MOOCs to SPOCs. *Commun ACM* 2013; 56(12): 38–40.
[3] Ezen-Can A, Boyer KE, Kellogg S, et al. Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In: *Proceedings of the 5th international conference on learning analytics and knowledge*, Poughkeepsie, NY, 16–20 March 2015, pp. 146–150. New York: ACM.
[4] Huang F, Zhang S, Zhang J, et al. Multimodal learning for topic sentiment analysis in microblogging. *Neurocomputing* 2017; 253: 144–153.
[5] Xiong S, Wang K, Ji D, et al. A short text sentiment-topic model for product reviews. *Neurocomputing* 2018; 297: 94–102.
[6] Bagheri A. Integrating word status for joint detection of sentiment and aspect in reviews. *J Inf Sci* 2019; 45: 736–755.
[7] Kim EHJ, Jeong YK, Kim Y, et al. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *J Inf Sci* 2016; 42(6): 763–781.
[8] Goffman E. *The presentation of self in everyday life*. London: Harmondsworth, 1978, p. 56.
[9] Xu H, Zhang F and Wang W. Implicit feature identification in Chinese reviews using explicit topic mining model. *Knowl-Based Syst* 2015; 76: 166–175.
[10] Liu Y, Xiong Q, Sun J, et al. Topic-based hierarchical Bayesian linear regression models for niche items recommendation. *J Inf Sci* 2019; 45(1): 92–104.
[11] Liu Y, Du F, Sun J, et al. iLDA: an interactive latent Dirichlet allocation model to improve topic quality. *J Inf Sci* 2020; 46: 23–40.
[12] Wu L, Wang D, Zhang X, et al. MLLDA: multi-level LDA for modelling users on content curation social networks. *Neurocomputing* 2017; 236: 73–81.
[13] Lim KW and Buntine W. Twitter opinion topic model: extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In: *Proceedings of the 23rd ACM international conference on information and knowledge management*, Shanghai, China, 3–7 November 2014, pp. 1319–1328. New York: ACM.
[14] Gupta M and Gupta P. Research and implementation of event extraction from twitter using LDA and scoring function. *Int J Inform Techn* 2019; 11(2): 365–371.
[15] Liu Y, Wang J, Jiang Y, et al. Identifying impact of intrinsic factors on topic preferences in online social media: a nonparametric hierarchical Bayesian approach. *Inform Sciences* 2018; 423: 219–234.
[16] Jo Y and Oh AH. Aspect and sentiment unification model for online review analysis. In: *Proceedings of the 4th ACM international conference on web search and data mining*, Hong Kong, China, 9–12 February 2011, pp. 815–824. New York: ACM.
[17] Zhao WX, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models. In: *Proceedings of the European conference on information retrieval*, Dublin, 18–21 April 2011, pp. 338–349. Berlin; Heidelberg: Springer.

[18] Zhang Y, Wang Z, Yu Y, et al. LF-LDA: a supervised topic model for multi-label documents classification. *Int J Data Warehous* 2018; 14(2): 18–36.
[19] Brody S and Diakopoulos N. Cooooooooooooooollllllllllll!!!!!!!!!!!!!! Using word lengthening to detect sentiment in microblogs. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, Edinburgh, 27–31 July 2011, pp. 562–570. Stroudsburg, PA: Association for Computational Linguistics.
[20] Blei DM, Ng AY and Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
[21] Wallach HM. Topic modeling: beyond bag-of-words. In: *Proceedings of the 23rd international conference on machine learning*, Pittsburgh, PA, 25–29 June 2006, pp. 977–984. New York: ACM.
[22] Blei DM and Lafferty JD. A correlated topic model of science. *Ann Appl Stat* 2007; 1(1): 17–35.
[23] Jin J, Geng Q, Mou H, et al. Author–Subject–Topic model for reviewer recommendation. *J Inf Sci* 2019; 45: 554–570.
[24] Bagheri A, Saraee M and De Jong F. ADM-LDA: an aspect detection model based on topic modelling using the structure of review sentences. *J Inf Sci* 2014; 40(5): 621–636.
[25] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents. In: *Proceedings of the 20th conference on uncertainty in artificial intelligence*, Banff, AB, Canada, 7–11 July 2004, pp. 487–494. Arlington, VA: AUAI Press.
[26] Mei Q, Ling X, Wondra M, et al. Topic sentiment mixture: modeling facets and opinions in weblogs. In: *Proceedings of the 16th international conference on World Wide Web*, Banff, AB, Canada, 8–12 May 2007, pp. 171–180. New York: ACM.
[27] Lin C and He Y. Joint sentiment/topic model for sentiment analysis. In: *Proceedings of the 18th ACM conference on information and knowledge management*, Hong Kong, China, 2–6 November 2009, pp. 375–384. New York: ACM.
[28] Tang F, Fu L, Yao B, et al. Aspect based fine-grained sentiment analysis for online reviews. *Inform Sciences* 2019; 488: 190–204.
[29] Ali F, Kwak D, Khan P, et al. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl-Based Syst* 2019; 174: 27–42.
[30] Nguyen TH, Shirai K and Velcin J. Sentiment analysis on social media for stock movement prediction. *Expert Syst Appl* 2015; 42 (24): 9603–9611.
[31] Liu Z, Wang T, Pinkwart N, et al. An emotion oriented topic modeling approach to discover what students are concerned about in course forums. In: *Proceedings of the 2018 IEEE 18th international conference on advanced learning technologies (ICALT)*, Mumbai, India, 9–13 July 2018, pp. 170–172. New York: IEEE.
[32] Liu Z, Yang C, Rüdian S, et al. Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums. *Interact Learn Envir* 2019; 27: 598–627.
[33] Ramesh A, Kumar SH, Foulds J, et al. Weakly supervised models of aspect-sentiment for online course discussion forums. In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing*, Beijing, China, 26–31 July 2015, pp. 74–83. Stroudsburg, PA: Association for Computational Linguistics.
[34] Xu Z, Zhang Y, Wu Y, et al. Modeling user posting behavior on social media. In: *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, Portland, OR, 12–16 August 2012, pp. 545–554. New York: sACM.
[35] Qiu M, Zhu F and Jiang J. It is not just what we say, but how we say them: LDA-based behavior-topic model. In: *Proceedings of the 2013 SIAM international conference on data mining*, Austin, TX, 2–4 May 2013, pp. 794–802. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
[36] Zhang HP, Liu Q, Cheng XQ, et al. Chinese lexical analysis using hierarchical hidden Markov model. In: *Proceedings of the 2nd SIGHAN workshop on Chinese language processing*, Sapporo, Japan, 11–12 July 2003, vol. 17, pp. 63–70. Stroudsburg, PA: Association for Computational Linguistics.
[37] Ku LW, Liang YT and Chen HH. Opinion extraction, summarization and tracking in news and blog corpora. In: *Proceedings of the 21st national conference on artificial intelligence*, Boston, MA, 16–20 July 2006, pp. 100–107. Palo Alto, CA: AAAI Press.
[38] Dong ZD. *HowNet's home page*, 2013, http://www.keenage.com
[39] Li J. Chinese derogatory dictionary v1.0, 2011, http://nlp.csai.tsinghua.edu.cn/site2/index.php/zh/resources/13-v10