Dynamic non-parametric joint sentiment topic mixture model

Xianghua Fu, Kun Yang, Joshua Zhexue Huang, Laizhong Cui

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China

A R T I C L E   I N F O

Keywords:
Topic sentiment analysis
Dynamic topic analysis
Non-parametric topic model
Social media
Hierarchical Dirichlet Process
Text mining

A B S T R A C T

The reviews in social media are produced continuously by a large and uncontrolled number of users. To capture the mixture of sentiment and topics simultaneously in reviews is still a challenging task. In this paper, we present a novel probabilistic model framework based on the non-parametric hierarchical Dirichlet process (HDP) topic model, called non-parametric joint sentiment topic mixture model (NJST), which adds a sentiment level to the HDP topic model and detects sentiment and topics simultaneously from reviews. Then considered the dynamic nature of social media data, we propose dynamic NJST (dNJST) which adds time decay dependencies of historical epochs to the current epochs. Compared with the existing sentiment topic mixture models which are based on latent Dirichlet allocation (LDA), the biggest difference of NJST and dNJST is that they can determine topic number automatically. We implement NJST and dNJST with online variational inference algorithms, and incorporate the sentiment priors of words into NJST and dNJST with HowNet lexicon. The experiment results in some Chinese social media dataset show that dNJST can effectively detect and track dynamic sentiment and topics.

C⃝ 2015 Elsevier B.V. All rights reserved.

1. Introduction

The rise of the social media motivates people to express their sentiment and opinions about anything more freely and frequently than ever before. For most commercial organizations and government departments, the user generated reviews represent invaluable source of information. Although many works have been done to extract information from reviews, summarize user’s opinions, and categorize reviews according to opinion polarities [15,18,20,22,26,30,32,45], it is still a challenge for users to easily digest and exploit the large number of reviews due to the inadequate supports for understanding individual reviewer’s opinions at the fine-grained level of topical aspects [9,41,48]. In fact, most existing works detect sentiment in isolation of topic detection. To address this issue, topic models are introduced for simultaneous analysis of topics and sentiment in a document. These studies, which jointly model topic and sentiment, take the advantage of the relationship between topics and sentiment, and are shown to be superior to traditional sentiment analysis tools [17,18,21,37,38,53]. But these methods only consider the static dataset, Bollen’s [5] and Connor’s [24] have shown that sentiment dynamics of online contents have a strong correlation with the fluctuations of macroscopic social and economic indicators in the same time period. Furthermore, social media data are produced continuously by many uncontrolled users, so the dynamic nature of such data requires the sentiment and topic analysis model to be updated dynamically.

To the best of our knowledge, TSM [21] and dJST [11] are the very few studies to detect and track dynamic topic and sentiment based on probability topic model, where TSM is based on probability latent semantic analysis (pLSA) model [12], and dJST is based on latent Dirichlet allocation (LDA) [4]. Since pLSA and LDA are parametric probabilistic model, both of them require to determine the topic number beforehand. It is insufficient for the dynamic and massive social media data. Furthermore, dJST is implemented with the Gibbs sampling algorithm, the drawbacks of which include: they are often hard to access convergence of the Markov chains, and they are not sufficient to deal with massive corpus [29,40].

In this paper, we propose a dynamic non-parametric joint sentiment topic model (dNJST) for detecting and tracking dynamic sentiment and topics of social reviews. We introduce a non-parametric joint sentiment topic model (NJST) through adding a sentiment level to the hierarchical Dirichlet process (HDP) topic model, and then present dynamic NJST (dNJST) which adds time decay dependencies of historical epochs to the current epochs. Compared with the existing sentiment-topic models, the biggest difference of dNJST is that dNJST can determine topic number automatically. Furthermore, we implement dNJST with an online variational inference algorithm, and improve the sentiment identification by HowNet lexicon. The experiment results show that dNJST can effectively detect and track dynamic sentiment and
topic of Chinese social media. The main contributions of this paper are four-folds:

1. We propose the non-parametric joint sentiment and topic model (NJST) and its dynamic version dNJST. Different with the existing sentiment topic mixture models, NJST and dNJST add sentiment levels to the non-parametric HDP topic model, which can determine the topic number of each epoch automatically. To the best of our knowledge, both NJST and dNJST are the first work to attempt dynamic sentiment and topic detection based on the non-parametric HDP topic model.

2. We implement online variational algorithms for NJST and dNJST, which can compute quickly for large corpus. The existing JST and dJST models are implemented with the standard Gibbs sampling algorithms, and they are difficult to deal with massive data because they have to repeatedly sample from the posterior topic assignment for each word token through the entire corpus at each iteration.

3. The main purpose of the sentiment and topic mixture model is to extract the sentiment and topics from social media reviews. We apply our dNJST model to discover the dynamic sentiment and topics of Chinese social media with real social media data from the biggest Chinese Web online forum “Tianya Forum”. We compare the performance of dNJST with NJST, JST and dJST. The experimental results show that dNJST outperform NJST, JST and dJST in extracting topics of specific sentiment orientation, which indicates the effectiveness of our dynamic non-parametric model.

The remainder of the paper is organized as follows. In Section 2, we introduce some related works. The Non-parametric Joint Sentiment-Topic Model is proposed in Section 3. Section 4 introduces the Dynamic NJST model and its online variational algorithm. Experiments and evaluations are reported in Section 5. We conclude the paper in Section 6 with future researches.

2. Related work

In recent years, there have been many research works on sentiment analysis and opinion mining [9,25,33]. Sentiment analysis methods usually can be divided into two categories: (1) the first type is based on part-of-speech (POS) tagging of the words and sentiment lexicons. This type is proposed by Peter [39] at first. Qun Liu [19] established a believable vocabulary on Chinese semantic knowledge named HowNet lexicon, and then get the sentiment polarity of words through comparison with the similarity between the words. Yanlan Zhu [54] succeeded in judging semantic orientation of Chinese online reviews based on the HowNet lexicon. Some other works such as Linhong Xu [47] and Weifu Du [8] adopted this type of method for the sentiment analysis of Chinese online reviews. (2) the second type is based on machine learning algorithms. For example, Whitelaw used support vector machine (SVM) to classify the sentiment orientation of movie reviews [43]. Socher [30] trained a sentiment Treebank with Recursive Neural Tensor Network to identify the sentence level sentiment. Some researchers also used machine learning algorithms to classify the sentiment orientation of Chinese online reviews, e.g., Jun Xu used Naive Bayesian (NB) and maximum entropy [46] to classify sentiment of Chinese news, Yi Hu [13] compared the qualities of sentiment classification between SVM and NB classifier.

Although researchers have got many achievements, most existing works focused on the sentiment classification of the product and service reviews. Only a few researchers pay attention to the sentiment analysis of the social reviews. For example, Mullen and Malouf [23] described preliminary statistical tests on a new dataset of political discussion group postings. Somasundaran [31] explored the utility of sentiment and arguing opinions for classifying stances in ideological debates. Mei [21] proposed Topic-Sentiment Mixture (TSM) model to reveal the latent topical facets in a Weblog collection. Bollen [5] performed a sentiment analysis of tweets, and found that the events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. In Chinese social reviews’ sentiment analysis, Tao [35] proposed an approach for feature extraction of sentiment analysis of the news comments. Yang [50] attempted to construct a new sentiment lexicon with sentiment orientation extent based on existing HowNet and NTUSD, which was applied to a semi-automatic Web public opinion analysis system. Daifeng Li [16] proposed a Topic-level Opinion Influence Model (TOIM) to predict the users’ future opinions on specific topics with a large dataset from Tencent Weibo. In addition, because of the figurative language such as irony, metaphor is generally used in social media, some researchers focus on identifying the figurative language and analyzing its polarity [27,28].

On the other hand, probabilistic topic models provide a principled and elegant way to discover hidden topics from large document collections [3]. pLSA [12] and LDA [4] are widely-used probabilistic topic models. One of the main advantages of LDA is that it can be easily used as a module in more complicated models for more complicated goals. A number of extensions to LDA have been proposed. For example, to detect the topics in social media, Yan [49] proposed a bitterm topic model (BTM) for modeling topics in short texts, and Diao [6] proposed a topic model that captured the similar relation among posts. When the models like LDA are used, the question usually arises that how many topics the estimated model should have, given the document collection [51]. The problem can be addressed by sharing a discrete base distribution among documents. A hierarchical Dirichlet process (HDP) [36] creates such a discrete base distribution for the document Dirichlet processes (DPs) by sampling from another DP. To learn evolutionary topics from a time varying corpus, some works have focused on extending LDA and HDP to dynamic topic models. Wang and McCallum [42] presented a LDA-style topic model called Topic Over Time (TOT) that explicitly modeled time jointly with word co-occurrence patterns. Tang [34] proposed a new generative model to simulate the generation process of both web contents and user’s participation in a unified framework. Kawamae [14] presented the theme chronicle model (TCM) which divided traditional topics into temporal and stable topics to detect the change of each theme over time. Amr Ahmed [1] presented an infinite dynamic topic models (iDTM) based on HDP topic model, which allowed for unbounded number of topics: topics can die or be born at any epoch, and the representation of each topic can evolve according to a Markovian dynamics. Zhang [52] proposed an evolutionary hierarchical Dirichlet process (EvoHDP) model.

Recently, probabilistic topic models are also introduced for simultaneous analysis of topics and sentiment in a document. These studies, which jointly model topic and sentiment, take the advantage of the relationship between topics and sentiment, and are shown to be superior to traditional sentiment analysis tools. The first topic and sentiment model is the Topic-Sentiment Model (TSM) [21], which jointly models the mixture of topics and sentiment predictions for the entire document. Because TSM is essentially based on pLSA model with an extra background component and two additional sentiment subtopies, it suffers from the problems of inference on new document and overfitting the data. Titov and McDonal [37,38] proposed the Multi-grain Latent Dirichlet Allocation model (MG-LDA) to build topics that were
representative of ratable aspects of objects from either a global topic or a local topic, which worked on a supervised setting. One of the most closely related works is that Lin and He [18] proposed a Joint Sentiment/Topic Model (JST). Unlike other machine learning approaches for sentiment classification which often require labeled corpus for classifier training, the JST model is fully unsupervised. There some other related topic-sentiment modeling works described as follows. Li and Huang [17] proposed the Sentiment-LDA model and Dependency-Sentiment-LDA. Unlike the previous models with the assumption that the sentiments of the words in the document are all independent, this model regarded the sentiments of words as a Markov chain. Zhao and Jiang [53] proposed a MaxEnt-LDA hybrid model to jointly discover both aspects and aspect-specific opinion words. They showed that with a relatively small amount of training data, MaxEnt-LDA hybrid model could effectively identify aspect and opinion words simultaneously.

Since social media data are produced continuously by many uncontrolled users, the dynamic nature of such data requires the sentiment and topic analysis model to be updated dynamically. TSM [21] and dJST [11] are the rarely work to detect and track dynamic topic and sentiment based on probability topic model. However, since TSM is essentially based on the pLSA model with a fixed background component and two additional sentiment subtopics, it suffers from the problems of inference on new document and overfitting the data, both of which are known as the deficits of pLSA. dJST is based on the LDA model, which is necessary to determine the topic number beforehand, but it is difficult to social reviews. Furthermore, dJST is implemented with Gibbs sampling algorithms, some drawbacks of which include: they are often hard to access convergence of the Markov chains, and not sufficient to deal with massive corpus [29,40]. To address these issues, we propose a dynamic non-parametric joint sentiment topic (dNJST) model in this paper. Compared with the existing methods, our approach is different in the following aspects: (1) dNJST is the first dynamic sentiment-topic mixture model based on non-parametric HDP topic model; (2) dNJST is the first sentiment-topic model which is implemented with an online variational inference algorithm. To address this issue, Wang [40] presented an alternative stick-breaking construction called Sethuraman’s stick-breaking construction, which was formed by twice applying Sethuraman’s stick-breaking construction of the DPs. For the corpus-level DP draw, this representation is

\[ \beta_0 \sim \text{Beta}(1, \gamma), \quad \beta_k = \beta_0^{k-1}(1 - \beta_1), \quad \phi_k \sim H, \quad G_0 = \sum_{k=1}^{\infty} \beta_k \phi_k \]

(3)

Thus, \( G_0 \) is discrete and has support at the atoms \( \phi = \{\phi_k\}_{k=1}^{\infty} \) with weights \( \beta = \{\beta_k\}_{k=1}^{\infty} \). And then in the document level, Sethuraman’s construction also is applied to each \( G_j \).

\[ \pi_j \sim \text{Beta}(1, \omega), \quad \pi_{jr} = \pi_j \prod_{m=1}^{b_j} (1 - \pi_{jm}), \quad \psi_{jr} \sim G_0, \quad G_j = \sum_{r=1}^{\infty} \pi_{jr} \psi_{jr} \]

(4)

where each document level atom (i.e., the topic of each different word’s in vocabulary) \( \psi_{jr} \) maps to a corpus-level atom \( \phi_j \) in \( G_0 \) according to the distribution defined by \( G_0 \). The indicator variables \( \psi_j = \{\psi_{jr}\}_{r=1}^{\infty} \), where \( \psi_j \sim \text{Mult}(\beta) \), then \( \psi_{jr} = \phi_j \). The weights of document level atoms are \( \pi_j = \{\pi_{jr}\}_{r=1}^{\infty} \).

Given \( G_j \), the generative process for the \( i \)th words in the \( j \)th document, \( w_{ij} \), is described as follows:

\[ z_{ij} \sim \text{Mult}(\pi_j), \quad \theta_{ij} = \psi_{z_{ij}} = \phi_j, \quad w_{ij} \sim \text{Mult}(\theta_{ij}) \]

(5)

The indicator \( z_{ij} \) selects topic parameter \( \psi_{z_{ij}} \), which maps to one topic \( \phi_j \) through the indicator \( c_i \). The corresponding graphical model is shown in Fig. 1(c).

**Fig. 1.** Graphical representation for HDP: circles denote random variables, oval nodes denote parameters, shaded nodes denote observed variables, and plates indicate replication. (a) HDP. (b) The stick-breaking construction of HDP. (c) The Sethuraman’s stick-breaking construction.
### 3.2. Non-parametric joint sentiment-topic model

In order to detect the sentiment and topics from the reviews at the same time, Lin propose the JST model [18]. JST is effectively a four-layer model, where sentiment labels are associated with documents, topics are associated with sentiment labels and words are associated with both sentiment labels and topics. A graphical model of JST is illustrated in Fig. 2(a). Since JST is not insufficient for dynamic social reviews, we add a sentiment level to extend the standard HDP topic model, which is called as the non-parametric joint sentiment-topic model (NJST). The procedure of generating a word \( w_i \) in document \( d_k \) under NJST consists of four stages.

Assumed there is \( L \) sentiment labels in the corpus. First, for each sentiment label \( l, l \in \{1, \ldots, L\} \), the global measure \( G_0 \) is drawn from a DP \((\gamma_1, H_1)\), and a set of measures \( \{G_l\}_{l=1}^{\infty} \) is drawn from a DP with the base measure \( G_0 \). Then, for a word \( w_i \) in the document \( d_k \), a sentiment label \( l_i \) is chosen from the per-document sentiment distribution \( \omega_i \). Then, a topic is chosen according to the multinomial distribution mixture model \( \text{Mult}(\theta_i) \) which corresponds with the sentiment label \( l_i \). Finally, a word is drawn from the per-corpus word distribution conditioned on both topic and sentiment label. The formal definition of the generative process of NJST corresponding to the graphical model shown in Fig. 2(b) is described as follows:

\[
G_0 | H_1 \sim \text{DP} (\gamma_1, H_1), \quad G_l | G_0 \sim \text{DP} (\gamma_l, G_0) \tag{6}
\]

\[
\omega_i \sim \text{Dir}(\nu), \quad l_i | \omega_i \sim \omega_i, \quad \theta_l | G_l \sim G_l, \quad w_i | \theta_l \sim \text{Mult}(\theta_l) \tag{7}
\]

The base distribution \( H_1 \) is a symmetric Dirichlet over the vocabulary simplex, and the atoms of \( G_0 \) are the topics \( \left\{ \phi_k \right\}_{k=1}^{\infty} \), which are independently drawn from \( H_1 \).

It is worth noting that the topic distribution of NJST is different from that of HDP. In HDP, there is only one measure \( C_l \) for each individual document. In contrast, in NJST, each document is associated with \( L \) measures, each of which corresponds to a sentiment label \( l \) with the different number of atoms. This feature essentially provides the ability for the NJST model to predict the sentiment associated with the extracted topics. Furthermore, since the topic number of each sentiment labels is fixed in JST but not fixed in NJST that is also different from JST.

**NJST with representation of Sethuraman’s stick-breaking construction.** To implement an online inference algorithm of NJST, we propose a representation of NJST with Sethuraman’s alternative stick-breaking construction [40], which graphical model is shown in Fig. 2(c).

In NJST, there is a different topic distribution for each sentiment label \( l \). We can think that each sentiment label corresponds with a HDP topic model. So there is a stick-breaking construction for each sentiment label \( l \). Then for the corpus-level DP drawn from the sentiment label \( l \), the stick-breaking representation is

\[
\beta_l \sim \text{Beta}(1, \gamma_1), \quad \beta_l = \beta_l \left( \prod_{m=1}^{k-1} (1 - \beta_m) \right), \quad \phi_k \sim H, \quad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \tag{8}
\]

where \( G_0 \) is discrete and has support at the atoms \( \phi = (\phi_k)_{k=1}^{\infty} \) with weights \( \beta_l = (\beta_l)_{l=1}^{\infty} \).

In the document level, the stick-breaking representation of each \( G_0 \) is

\[
\pi_{l,i} \sim \text{Beta}(1, \gamma_0), \quad \pi_{l,i} = \pi_{l,i} \left( \prod_{j=1}^{i-1} (1 - \pi_{l,j}) \right), \quad \psi_{l,i} \sim G_0, \quad G_{l,i} = \sum_{j=1}^{\infty} \pi_{l,j} \delta_{\psi_{l,j}} \tag{9}
\]

where each document level atom \( \psi_{l,i} \) maps to a corpus-level atom \( \phi_k \). The indicator variables \( c_{l,i} = (c_{l,i})_{i=1}^{\infty} \) is used to represent \( \pi_{l,i} = (\pi_{l,i})_{i=1}^{\infty} \), which maps to one topic \( \phi_k \).

Given \( G_{l,i} \), the generative process for the \( i \)th words of the document \( d_l, w_{l,i} \), is

\[
\omega_{l,i} \sim \text{Dir}(\nu), \quad l_{l,i} \sim \text{Mult}(\omega_{l,i}), \quad z_{l,i} \sim \text{Mult}(\pi_{l,i}), \quad \theta_{l,i} = \psi_{l,i}, \quad w_{l,i} \sim \text{Mult}(\theta_{l,i}) \tag{10}
\]

where the sentiment indicator \( l_k \) of \( w_{l,i} \) is determined from the multinomial sentiment distribution \( w_{l,i} \) of the document \( d_k \). According to the multinomial distribution \( \pi_{l,i} \), a topic \( z_{l,i} \) is chosen from the document level atoms \( \psi_{l,i} \), which maps to one topic \( \phi_k \) through the indicator \( c_{l,i} \). Finally, the generation probability of \( w_{l,i} \) is determined by the topic \( \phi_{c_{l,i}} \) (the topic in \( \phi \) with \( l = l_{l,i}, k = z_{l,i} \)).

### 3.3. Online variational inference algorithm of NJST

We implement our online variational inference algorithm of NJST based on Wang's online algorithm [40]. In NJST, the hidden

---

**Fig. 2.** Graphical representation for HDP: circles denote random variables, oval nodes denote parameters, shaded nodes denote observed variables, and plates indicate replication. (a) JST. (b) NJST. (c) NJST with representation of the Sethuraman’s stick-breaking construction.
variables that we are interested in include the top-level stick proportion \( \{ \beta_i^t \}_{i=1}^K \), bottom level stick proportion \( \{ \eta_j^t \}_{j=1}^R \), indicator vector \( \{ c_{lj}^t \}_{l=1}^L \), and the sentiment-topic distribution \( \{ \phi_{lj}^t \}_{l=1}^L \), the sentiment distribution \( c_l \) of each document, the sentiment label \( l_i \) and the topic index \( j \) for each word \( w_j \).

According to the variational inference of DPs [2], the variational distribution of NJST has the following form,

\[
q(\beta^t, \eta, z, \phi, l, c) = q(\beta^t)q(\eta)q(z)q(\phi)q(l)q(c)
\]

which can be factorized into

\[
q(\beta^t) = \prod_{k=1}^{K-1} q(\beta_k^t | u_k, v_k), \quad q(\eta) = \prod_{j=1}^{R-1} q(\eta_j^t | v_j),
\]

\[
q(\phi) = \prod_{l=1}^{L-1} \prod_{j=1}^{K-1} q(\phi_{lj}^t | a_{lj}, b_{lj}), \quad q(c) = \prod_{l=1}^{L-1} q(c_l)
\]

where the variational parameters are \( \phi_{lj}^t \) (multinomial), \( \beta_k^t \) (multinomial), and \( z_{lj}^t \). They are the parameters of beta distributions. For each sentiment label, the truncations for the corpus and document levels respectively are \( K \) and \( R \). Based on Wang’s algorithm [40], the online variational inference algorithm includes two level updates, the document-level update and the corpus-level update.

**Document-level update:** At the document level, we update the parameters to the per-document stick, the parameters to the per word topic indicators, the parameters to the per word sentiment indicators, and the parameters to the per document and per sentiment topic indices,

\[
a_{lj}^t = 1 + \sum_{r=1}^{R} \tilde{z}_{ljr}^t
\]

\[
b_{lj}^t = \alpha'_0 + \sum_{r=1}^{R} \tilde{z}_{ljm}^t
\]

\[
v_j^t = \beta_j^t + \sum_{i=1}^{L} \tilde{z}_{ij}^t
\]

\[
\phi_{lj}^t \propto \exp \left( \sum_{r=1}^{R} \lambda_{ljr}^t \log p(w_j | \phi_{lj}) + \sum_{r=1}^{R} \log p(l_j | v_j) + \sum_{r=1}^{R} \log p(b_j) \right)
\]

\[
\tilde{z}_{ljr}^t \propto \exp \left( \sum_{r=1}^{R} \phi_{ljr}^t \log p(w_j | \phi_{lj}) + \sum_{r=1}^{R} \log p(l_j | v_j) + \sum_{r=1}^{R} \log \pi_{rj} \right)
\]

The expectations involved above are taken under the variational distribution \( q \), which are calculated by

\[
E_q[\log \beta_k^t] = \sum_{m=1}^{K-1} E_q[\log (1 - \beta_m^t)] + \sum_{m=1}^{K-1} E_q[\log (1 - \beta_m^t)]
\]

**Corpus-level online update:** at the corpus level, we update the parameters to top-level-sticks and the topics online,

\[
\lambda_l^t = \lambda_l^t + \rho \lambda_j^t \partial \lambda_l^t(j)
\]

\[
u_j^t = \nu_j^t + \rho \nu_j \partial \nu_j(j)
\]

\[
v_l^t = v_l^t + \rho v_l \partial v_l(j)
\]

where \( \partial \lambda_l^t(j) \), \( \partial \nu_j(j) \) and \( \partial v_l(j) \) are the components of the natural gradient of the variational objects. The expression of them are

\[
\partial \lambda_l^t(j) = -\lambda_l^t + \eta + \sum_{r=1}^{R} \phi_{ljr}^t \left( \sum_{i=1}^{L} z_{ljir}^t I[w_j = w_i] \right)
\]

\[
\partial \nu_j(j) = -\nu_j + 1 + \sum_{r=1}^{R} \phi_{ljr}^t
\]

\[
\partial v_l(j) = -v_l + \gamma_j + \sum_{r=1}^{R} \phi_{ljr}^t
\]

where \( N \) is the total number of documents in the corpus, and \( \rho_0 \) is the learning rate, and is selected as

\[
\sum_{t=1}^{\infty} \rho_0 = \infty, \sum_{t=1}^{\infty} \rho_0^2 < \infty
\]

The expectations involved above are taken under the variational distribution \( q \), which are calculated by

\[
E_q[\log \beta_k^t] = \sum_{m=1}^{K-1} E_q[\log (1 - \beta_m^t)] + \sum_{m=1}^{K-1} E_q[\log (1 - \beta_m^t)]
\]

\[
E_q[\log \beta_k^t] = \psi(u_k) - \psi(u_k + v_k)
\]

\[
E_q[\log (1 - \beta_m^t)] = \psi(b_j) - \psi(a_j + b_j)
\]

\[
E_q[\log p(w_j = w_i | \phi_{lj})] = \psi(b_j) - \psi \left( \sum_{i=1}^{L} \phi_{lj}^t \right)
\]

\[
E_q[\log p(w_j = w_i | \phi_{lj})] = \psi(b_j) - \psi \left( \sum_{i=1}^{L} \phi_{lj}^t \right)
\]

where \( \psi(\cdot) \) is the digamma function.

For each sentiment label \( t \), the online variational inference algorithm for the NJST is described in Fig. 3.

**4. Dynamic NJST model**

**4.1. Historical dependencies among NJST models**

It is noted that the social media reviews are generated continuously by many uncontrolled users. The dynamic nature of such data requires both the sentiment and topic analysis model to be updated dynamically. Assumed that the reviews are sorted in the ascending order of their timestamps, and the time is divided into \( T \) epochs. The corpus of the epoch \( t \) is denoted by \( D^t \), and the whole corpora over time is denoted by \( D = \{ D^1, D^2, \ldots, D^T \} \). We can model each epoch as a NJST model, and model the multiple correlated time-varying corpora as a series of NJST models with time dependencies, as shown in the graphical representation of Fig. 4. We call it dynamic NJST model (dNJST),

\[
\frac{\partial \lambda_l^t(j)}{\partial l} = -\lambda_l^t + \eta + \sum_{r=1}^{R} \phi_{ljr}^t \left( \sum_{i=1}^{L} z_{ljir}^t I[w_j = w_i] \right)
\]

\[
\frac{\partial \nu_j(j)}{\partial l} = -\nu_j + 1 + \sum_{r=1}^{R} \phi_{ljr}^t
\]

\[
\frac{\partial v_l(j)}{\partial l} = -v_l + \gamma_j + \sum_{r=1}^{R} \phi_{ljr}^t
\]

Fig. 3. The online variational inference algorithm for NJST model.
We assume that the reviews at current epoch could be influenced by previous reviews. Thus, in dNJST, at epoch $t$, for each sentiment label $l$, the current sentiment-word specific word distributions $\phi'_{lk}$ are influenced by the word distributions of previous epochs, and the sentiment-topic weights $\mu_{lk}$ are influenced by the previous epochs. Assumed that in a sliding window $\Delta = \{t - \Delta - 1, \ldots, t - 1\}$, the influences of previous epochs are generated in a time decay way. At epoch $t$, the dependencies sentiment-word distribution is denoted by $\phi'_d$, and the dependencies topic weight denoted by $\mu'_d$, which are defined as follows:

$$\phi'_d = \sum_{\beta=1}^{\beta} \exp\gamma_{\beta}^d \beta^{-\delta}$$

$$\mu'_d = \sum_{\beta=1}^{\beta} \exp\gamma_{\beta,0}^d \beta^{-\delta}$$

where $\Delta$ and $C$ represent the width and decay factor of the time-decay kernel. This defines a $\Delta$-order process where the strength of dependencies between epoch-specific DPs is controlled by $\Delta$ and $C$.

The time dependencies of parameters of NJST models can be illustrated in Fig. 4. Obviously, for each sentiment label $l$, the atoms of base measure $\phi_j^l = (\phi_k^l)_{k=1}^{\infty}$ and the weights of these atoms $\beta^l_j = (\beta_k^l)_{k=1}^{\infty}$ are not only depend on their super-parameters $\gamma_j^l$ and $\gamma_k^l$, but also depend on the parameter values of previous NJST models $\{\phi_j^k, \ldots, \phi_j^{t-1}\}$ and $\{\beta_j^k, \ldots, \beta_j^{t-1}\}$. Thus, at epoch $t$, the prior values of $\beta^l_j$ and $\phi_j^l$ are $\beta^l_j$, $\phi_j^l$.

According to Eq. (8), in our dNJST model, at epoch $t$, for each sentiment label $l$, the stick-breaking representation of the corpus level DP is

$$\beta^l_{jk} \sim \text{Beta}(1, \gamma_j^l), \quad \beta^l_{jk} = \beta^l_{j0} + \prod_{m=1}^{k-1} (1 - \beta^l_{jm})$$

$$\phi^l_{jk} \sim H_l, \quad \phi^l_{jk} = \phi^l_{j0} + \phi^l_{j0} G_0, \quad G_0 = \sum_{k=1}^{\infty} \beta^l_{jk} \phi^l_{jk}$$

where $G_0$ is discrete and has support at the atoms $\phi^l_j = (\phi_k^l)_{k=1}^{\infty}$ with weights $\beta^l_j = (\beta_k^l)_{k=1}^{\infty}$.

In the document level, at each epoch, the stick-breaking representation of each $G'_d$ in dNJST is the same with that in NJST.

4.2. Online variational inference algorithms of dNJST Model

We implement dNJST model based on the above online variational inference algorithm of NJST model. According to Eqs. (12) and (13), the stick parameter of $\beta^l_{jk}$ is $(\bar{u}^l_{jk}, \bar{v}^l_{jk})$, where $\bar{u}^l_{jk}$ is the stick strength of the atom $\phi^l_{jk}$, and $\bar{v}^l_{jk}$ is the stick strength of all the topics $\phi^l_{ik}$ ($i > k$), and then $\bar{u}^l_{jk} \sim u^l_{jk}$. In the online variational algorithm of NJST model, we can see that at the end of epoch $t$, for the sentiment label $l$, the stick parameters $\bar{u}^l = (\bar{u}^l_{jk})_{k=1}^{\infty}$ record the relative strength of all the topics, and the parameter $\lambda^l = (\lambda^l_{jk})_{k=1}^{\infty}$ record all the sentiment-word topic distributions. Therefore, to implement dNJST model, we can directly add the time dependencies of previous NJST models to their topic level sticks $(\bar{u}^l, \bar{v}^l)$ and the topics distribution $\lambda^l$.

At the end of epoch $t$, for the sentiment label $l$, the stick parameters of the NJST model of this epoch is denoted by $(\bar{u}^l, \bar{v}^l)$, and the topic is denoted by $\lambda^l$. Then the cumulative time decay dependencies of the epoch $t + 1$ can be denoted as

$$\bar{u}^l_{jk} = \sum_{k=1}^{\infty} \exp\gamma_{\beta}^d \beta^{-\delta}$$

$$\bar{v}^l_{jk} = \sum_{k=1}^{\infty} \exp\gamma_{\beta,0}^d \beta^{-\delta}$$

$$\lambda^l_{jk} = \sum_{k=1}^{\infty} \exp\gamma_{\beta}^l \beta^{-\delta}$$

According to Eqs. (20)–(22), the corpus level updating equations of the online inference algorithm of dNJST model with time decay dependencies are

$$u_{jk}^l \leftarrow \bar{u}_{jk}^l + \rho_u \partial u_{jk}(j)$$

$$v_{jk}^l \leftarrow \bar{v}_{jk}^l + \rho_v \partial v_{jk}(j)$$

$$\lambda_{jk}^l \leftarrow \lambda_{jk}^l + \rho_\lambda \partial \lambda_{jk}(j)$$

For simplicity, we let $\bar{u}_{jk}^l = 0$, $\bar{v}_{jk}^l = 0$, $\lambda_{jk}^l = 0$ for newly-born topics $k$ at epoch $t + 1$, and $\bar{u}_{jk}^l - \bar{u}_{jk}^l = 0$ denotes the topics that are available to be used but not yet used in any document at epoch $t + 1$. Since we do not know the document number of each epoch beforehand, we set $N$ in Eqs. (23)–(25), the default value of which is 100. When the document number of the corpus become larger than the default number, we set $N$ equal to the actual number.
Based on the online inference algorithm of NJST, for each sentiment label \( l \), our dNJST online algorithm at the epoch \( t + 1 \) is described in Fig. 5.

5. Experiment and evaluation

5.1. Experiment setting

In this section, we present our experiment setting of dynamic topic evolutionary discovery with our TianYa-12261 dataset, which crawled from Chinese social media site.

(1) TianYa-12261 dataset

Since there are no open social media datasets, we crawl our experimental datasets from TianYa By-talk of TianYa Forum,\(^1\) which is the most famous Chinese online Bulletin Board System community, from 6 January 2011 to 8 March 2011, a total of 8 weeks. For each post, the crawled content includes text, title, link address, date, replies and other information. The total number of posts is 72,585. As the topic evolutionary analysis of social media mainly focuses on hot topics, some posts of which the replies number is smaller than 10 are discarded. Furthermore, some posts may be replied after long time, to simplify the data processing, we merge the replies and their original post into an article. Through the data preprocessing, including Chinese word segmentation, and filtering out low/high frequency words, stop words, improper characters, and single words, we get a dataset which contains 12,261 articles. The common threshold values to filter out low/high frequency words are set according to the experience. In our experiments, if the occurrence number of a word is smaller than 5, it is a low frequent word. While if the occurrence frequency of a word in a document is larger than 20%, it is a high frequent word. The Chinese word segmentation system ICTCLAS2013\(^2\) is applied to the word segmentation, which supports the user dictionary, and can find new words. We call this dataset as TianYa-12261 dataset. Then we divided the whole dataset into 8 corpus according to each epoch with a week. The document numbers of 8 epochs are respectively shown in Fig. 6. In fact, the TianYa-12261 dataset is a subset of our bigger dataset, called TianYa-80299 dataset\([10]\).

(2) Evaluation metric

10-fold cross-validation method is used in our experiment. For each corpus, according to time order we divided it into ten subsets randomly. In each experiment, we take 90% as the training data \( D_{train} \), and the remainder as the test data \( D_{test} \). We use the widely adopted perplexity as the evaluation metric of the performance of topic discovery. The perplexity is a appropriate metric to evaluate the generation capability of a topic model, which was proposed in the literature\([4]\). The formal definition of the perplexity is shown in Eq. (37).

\[
\text{perplexity}(D_{test}) = \exp \left \{ \sum_{j=1}^{D_{test}} \frac{\log(p(d_j))}{\sum_{w=1}^{N_j} \log(p(w|d_j))} \right \} \tag{37}
\]

where \( D \) is the document number, \( N_j \) represents the number of words in the document \( d_j \), and \( p(d_j) = \prod_{i=1}^{N_j} p(z_i|d_j)p(w|z_i) \). A smaller value of perplexity represents a better effect of clustering and generalization capability.

(3) Incorporating sentiment priors

In JST and dJST model, they incorporated the subjectivity lexicon as model priors to identify the sentiment polarity of words\([11,18]\), where the prior information was only utilized during the initialization of posterior distribution \( z \), i.e., the assignment of word token to sentiment label and topic. The initialization starts by comparing each word token in the corpus with the words in the sentiment word list. If there is a match, the word token is assigned to the corresponding sentiment label. Otherwise, a sentiment label is randomly chosen for a word token. In fact, the sentiment initialization of JST and dJST model is not good enough, so that we can improve it by integrating some prior knowledge. In our previous work on discovering multi-aspect topic and sentiment, HowNet lexicon was used to improve the sentiment analysis of topics\([9]\). HowNet is an on-line common-sense knowledgebase unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents\([7]\). Therefore, in order to promote the sentiment orientation identification of each token, we also introduce HowNet lexicon into our NJST and dNJST model to assign the prior sentiment label to each word token. Given a sentiment word \( w \), its sentiment orientation \( O(w) \) could be calculated as Eq. (38), where \( kp_0 \) denotes the positive benchmark words with the polarity +1, \( kq_0 \) denotes the negative benchmark words with polarity −1, \( m \) is the number of benchmark word pairs, and \( \text{sim}_{uw}() \) is the similarity function with HowNet lexicon\([9]\).

\[
O(w) = \sum_{i=1}^{m} \left( \text{sim}_{uw}(kp_i, w) - \text{sim}_{uw}(kq_i, w) \right) \tag{38}
\]

These sentiment benchmark words are selected according to the ordered Hits values returned by Google search engine, and the words which have similar semantics are merged. The forty pairs of sentiment benchmark words are list in Table 1.
5.2. Comparing with other sentiment-topic mixture models

In our previous work [9], we proposed the Multi-aspect Sentiment Analysis for Chinese Online Social Reviews (MSA-COSR) method, which integrated HowNet lexicon and LDA model. Through a dataset in which there are 300 Chinese social reviews and each paragraph has some manual sentiment labels, we compare MSA-COSR with JST model and some supervised methods such as Naive Bayesian (NB), K-nearest neighbor (KNN), support vector machine (SVM). The experiment results show that using HowNet lexicon is effective to identify the topic specified sentiment orientation [9].

However, since the topic model assumes the topic is a probabilistic distribution of all the words, it is infeasible to manually label the sentiments of each topic in large dataset. In the sentiment-topic models, we are more concerned about the extracted topics and the effectiveness of topic sentiment captured by these models. We do not evaluate the sentiment accuracy of each topic with labeled data, but evaluate the extracted topics of each sentiment labels by these sentiment-topic models. Only negative sentiment and positive sentiment are used.

We compare dNJST with other models including JST, dJST, NJST, and NJST-WH (NJST without HowNet lexicon). We implement NJST and dNJST methods based on Wang’s online HDP algorithm. The JST algorithm and dJST algorithms are from Chen and He’s implementation. NJST and dNJST initialize the sentiment labels of the word tokens with HowNet lexicon, and NJST-WH initializes the sentiment orientation of the word tokens in the corpus like JST and dJST. In the initialization of NJST-WH, when a word token matches the words in the sentiment benchmark list, it is assigned by the corresponding sentiment label. Otherwise, a sentiment label is randomly chosen for the word token.

Since both JST and dJST are based on LDA model, they need to specify the topic number beforehand. In fact, it is not easy to specify the topic number of social media by artificial setting. We run the online LDA model on Tianya-12261 with different topic number $k$($k=4,6,8,10,12,16,20$) [10]. The perplexity values of the online LDA model in different weeks are shown in Fig. 7.

From the experiment results in Fig. 7, we can see that $k=8$ is a good topic number value. So the topic numbers of both JST and dJST are set as $k=8$. We only consider the sliding window dependencies mode of dJST, which is similar with our time decay dependencies of dNJST. The time dependencies of continuous epochs are determined by the decay width $\lambda$ and the decay factor $\alpha$. In our previous work, we conducted some experiments to select appropriate decay parameter values of our dynamic online HDP model. The experimental results show that the dynamic online model can achieve the best performance when the value of $(\alpha, \lambda)$ closes to $(3,6)$ [10]. So the decay parameters values of JST and dNJST are respectively set as $\lambda=6$ and $\alpha=3$, and the decay weights of each epoch are calculated in Table 2. In addition, there are two level truncation parameters of our NJST, dNJST and NJST-WH, including the corpus level truncation parameter $K$ and the document level truncation parameter $R$. In our experiment, we set $K=150$ and $R=15$, which are the same with Wang’s work [44].

After training these sentiment-topic mixture models with Tianya-12261 dataset, we get the topics with positive sentiment label and the topics of negative sentiment label. Some typical topic words are listed in Table 3. The average perplexity value of each sentiment-topic model is shown in Fig. 8.

According to the experimental results in Fig. 8, we can approximately obtain the following relationships of different sentiment-topic models, where perplexity() denotes the perplexity value of a model.

$$perplexity(NJST) < perplexity(JST)$$

$$perplexity(dNJST) < perplexity(dJST)$$

$$perplexity(dNJST) < perplexity(JST)$$

$$perplexity(NJST) < perplexity(JST)$$

$$perplexity(dNJST) < perplexity(dJST)$$

$$perplexity(NJST, dNJST) < perplexity(JST, dJST, NJST – WH)$$

Eqs. (39) and (40) show that the perplexity values of NJST and dNJST respectively are smaller than that of JST and dJST, which demonstrates that the topic discovery effectiveness and generalization ability of non-parametric sentiment-topic mixture model are better than those of the parametric sentiment-topic mixture models in our Tianya-12261 dataset. We think the reason lies in the topics of social media naturally are not fixed, and the non-parametric can automatically determine the number of topics according to the corpora.

Eqs. (41) and (42) show that the perplexity values of dJST and dNJST respectively are better than that of JST and NJST, which demonstrates that the topic discovery effects and generalization ability of dynamic sentiment-topic mixture model are better than those of the non-dynamic sentiment-topic mixture models Tianya-12261 dataset. Furthermore, we can see that the perplexity

---

Table 1: The forty pairs of sentiment benchmark words.

<table>
<thead>
<tr>
<th>Benchmark words with sentiment value +1</th>
</tr>
</thead>
<tbody>
<tr>
<td>健康</td>
</tr>
<tr>
<td>天使</td>
</tr>
<tr>
<td>欢喜</td>
</tr>
<tr>
<td>著名</td>
</tr>
<tr>
<td>真实</td>
</tr>
<tr>
<td>茂盛</td>
</tr>
<tr>
<td>完整</td>
</tr>
<tr>
<td>not cooperate</td>
</tr>
<tr>
<td>沾染</td>
</tr>
<tr>
<td>污染</td>
</tr>
<tr>
<td>花样</td>
</tr>
<tr>
<td>虚假</td>
</tr>
<tr>
<td>荒诞</td>
</tr>
<tr>
<td>言语</td>
</tr>
<tr>
<td>缺</td>
</tr>
<tr>
<td>善</td>
</tr>
<tr>
<td>勤</td>
</tr>
<tr>
<td>变</td>
</tr>
<tr>
<td>资</td>
</tr>
<tr>
<td>谣</td>
</tr>
</tbody>
</table>

---

4 http://chenghualin.wordpress.com/code/.
5 https://sites.google.com/site/yulanhe8.
value of dNJST is almost the smallest one over the whole time. The experiment results show that it is reasonable to consider the time dependencies of adjacent epochs. That is to say, the model parameters (corpus-level parameters) in epoch \( t \) not only depend on that of the directly previous \( \Delta t \) epoch, but also depend on that of previous epochs.

Table 2
The time decay weights when \( \Delta t = 6 \) and \( C = 3 \).

<table>
<thead>
<tr>
<th>Weeks</th>
<th>t-1</th>
<th>t-2</th>
<th>t-3</th>
<th>t-4</th>
<th>t-5</th>
<th>t-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weights</td>
<td>0.305</td>
<td>0.218</td>
<td>0.156</td>
<td>0.112</td>
<td>0.080</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Eq. (43) show the perplexity values of NJST and dNJST are smaller than that of JST, dJST, and NJST-WH, which demonstrates that the sentiment-topic models with HowNet lexicon are more effective to extract sentiment topics than those models without HowNet lexicon. We think that is a reasonable result. Because during the process of sentiment topic discovery, HowNet lexicon can help the sentiment-topic models to get more information of context than the simple match method.

5.3. Evolutionary analysis of the emotional topic

To discover the sentiment-topic evolutionary with our dNJST, we calculate the weekly topic strength of positive and negative

Table 3
The top 18 words and their weights of some typical topic with positive and negative sentiment labels.

<table>
<thead>
<tr>
<th>Word</th>
<th>Weight</th>
<th>Word</th>
<th>Weight</th>
<th>Word</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical topics with positive sentiment label</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 1</td>
<td>经济/economic 0.022</td>
<td>城市/city 0.020</td>
<td>国家/countries 0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>农业/agricultural 0.012</td>
<td>社会/society 0.012</td>
<td>中国/China 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>工作/work 0.010</td>
<td>收入/income 0.010</td>
<td>全国/nationwide 0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>医疗/medical 0.009</td>
<td>资源/resources 0.009</td>
<td>香港/Hong Kong 0.008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>市场/market 0.007</td>
<td>文化/culture 0.007</td>
<td>地区/region 0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>政府/the government 0.007</td>
<td>改革/reform 0.006</td>
<td>广东/Guangdong 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 2</td>
<td>医院/hospital 0.034</td>
<td>孩子/child 0.020</td>
<td>工程/engineering 0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>政府/government 0.013</td>
<td>事物/thing 0.012</td>
<td>电话/phone 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>小时/hours 0.010</td>
<td>工作/work 0.010</td>
<td>东西/things 0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>项目/project 0.008</td>
<td>手机/phone 0.008</td>
<td>工资/wage 0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>儿子/son 0.006</td>
<td>家国/countries 0.006</td>
<td>政策/policy 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>关系/relationship 0.006</td>
<td>母亲/mother 0.006</td>
<td>妈妈/mother 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 3</td>
<td>学校/school 0.025</td>
<td>大学/university 0.016</td>
<td>男人/man 0.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>工作/work 0.013</td>
<td>权力/power 0.011</td>
<td>女人/woman 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>官方/officialdom 0.010</td>
<td>教育/education 0.009</td>
<td>社会/society 0.008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>英语/English 0.008</td>
<td>专业/professional 0.007</td>
<td>国家/country 0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>事件/the event 0.007</td>
<td>案件/case 0.006</td>
<td>成果/results 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>关系/relationship 0.006</td>
<td>中国/China 0.006</td>
<td>思维/thinking 0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical topics with negative sentiment label</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 1</td>
<td>赵红霞/Zhao Gongxiao 0.042</td>
<td>腐败/corruption 0.040</td>
<td>贪官/corrupt official 0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>雷政富/Lei ZhengFu 0.020</td>
<td>官方/official 0.016</td>
<td>法官/the judge 0.014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>宋林/Songlin 0.012</td>
<td>犯罪/crime 0.009</td>
<td>领导/leader 0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>原告/plaintiff 0.008</td>
<td>老百姓/ordinary people 0.008</td>
<td>视频/video 0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>被告/boss 0.007</td>
<td>情妇/mistress 0.007</td>
<td>干部/cadres 0.007</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>女人/woman 0.006</td>
<td>被告/defendant 0.005</td>
<td>被告/defendant 0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 2</td>
<td>村民/the villagers 0.041</td>
<td>人/employee 0.037</td>
<td>受害人/victim 0.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>问题/the problem 0.017</td>
<td>民警/police 0.016</td>
<td>警察/police 0.015</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>犯罪/crime 0.014</td>
<td>领导/leadership 0.013</td>
<td>农民/farmers 0.013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>律师/lawyer 0.011</td>
<td>公民/citizens 0.011</td>
<td>卫生/secretary 0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>公方/other party 0.010</td>
<td>代表/on behalf of 0.010</td>
<td>暴力/violence 0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic 3</td>
<td>医生/doctor 0.043</td>
<td>问题/the problem 0.028</td>
<td>向人/patient 0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>药品/drug 0.016</td>
<td>药品/medicine 0.011</td>
<td>西药/medicine 0.012</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>麻醉/Anesthesia 0.010</td>
<td>医师/doctor 0.010</td>
<td>医生/physician 0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>中医/Chinese medicine 0.010</td>
<td>老板/boss 0.010</td>
<td>西药/medicine 0.009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>反应/reaction 0.006</td>
<td>活动/activity 0.006</td>
<td>错误/fault 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 8. The perplexity values of JST, dJST, NJST, dNJST andNJST-WH in TianYa-12261 dataset from week 1 to week 8.

Fig. 9. The topic strengths of positive and negative trends on TianYa-12261 from week 1 to week 8.

Table 4
A typical topic evolution with positive label from week 1 to week 8.

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic word</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st week</td>
<td>中国/China 经济/economy 社会/society 企业/enterprise  国家/country</td>
</tr>
<tr>
<td>政府/government</td>
<td>媒体/media 世界/world 城市/city 美国/America</td>
</tr>
<tr>
<td>农村/rural area</td>
<td>农业/agricultural 改革/reform 科技/Sci. and Tech. 文化/culture</td>
</tr>
<tr>
<td>机制/mechanism</td>
<td>医院/hospital 全国/nationwide 关系/relationship 生活/life</td>
</tr>
<tr>
<td>2nd week</td>
<td>农业/agriculture 经济/economy 收入/income 国家/country 城市/city</td>
</tr>
<tr>
<td>政府/government</td>
<td>工作/work 全国/nationwide 社会/society 中国/China</td>
</tr>
<tr>
<td>香港/Hong Kong</td>
<td>资源/resources 农村/rural area 地区/regional 改革/reform</td>
</tr>
<tr>
<td>文化/culture</td>
<td>医疗/medical 市场/market 政策/policy 环境/environment</td>
</tr>
<tr>
<td>3rd week</td>
<td>经济/economy 创新/innovation 中国/China 社会/society 医疗/medical</td>
</tr>
<tr>
<td>城市/city</td>
<td>国家/country 政府/government 工作/work 收入/income</td>
</tr>
<tr>
<td>文化/culture</td>
<td>企业/enterprise 农村/rural area 香港/Hong Kong 改革/reform</td>
</tr>
<tr>
<td>广东/Guangdong</td>
<td>市场/market 资源/resources 改革/reform 深圳/Shenzhen</td>
</tr>
<tr>
<td>4th week</td>
<td>农村/rural area 经济/economy 城市/city 国家/country 农业/agriculture</td>
</tr>
<tr>
<td>社会/society</td>
<td>中国/China 工作/work 收入/income 全国/nationwide</td>
</tr>
<tr>
<td>医疗/medical</td>
<td>资源/resources 市场/market 文化/cultural</td>
</tr>
<tr>
<td>地区/regional</td>
<td>政府/government 改革/reform 广东/Guangdong 媒体/media</td>
</tr>
<tr>
<td>5th week</td>
<td>农业/agriculture 经济/economics 收入/income 国家/country 城市/city</td>
</tr>
<tr>
<td>政府/government</td>
<td>工作/work 全国/nationwide 社会/society 中国/China</td>
</tr>
<tr>
<td>香港/Hong Kong</td>
<td>资源/resources 农村/rural area 地区/regional 改革/reform</td>
</tr>
<tr>
<td>文化/culture</td>
<td>医疗/medical 市场/market 政策/policy 环境/environment</td>
</tr>
<tr>
<td>6th week</td>
<td>城市/city 经济/economy 国家/country 政府/government 全国/nationwide</td>
</tr>
<tr>
<td>社会/society</td>
<td>收入/income 农村/rural area 中央/China 地区/region</td>
</tr>
<tr>
<td>改革/reform</td>
<td>改革/reform 市场/market 政策/policy 改革/reform</td>
</tr>
<tr>
<td>政府/government cars</td>
<td>城镇/town 文化/culture 农业/agriculture 生活/life</td>
</tr>
<tr>
<td>7th week</td>
<td>经济/economy 农村/rural area 政府/government 城市/city 国家/country</td>
</tr>
<tr>
<td>中国/China</td>
<td>社会/society 收入/income 小区/well-off society 政策/policy</td>
</tr>
<tr>
<td>地区/regional</td>
<td>市场/market 精神/spirit 土地/land 改革/reform</td>
</tr>
<tr>
<td>技术/technology</td>
<td>农业/agriculture 资源/resources 政策/policy 生活/life</td>
</tr>
<tr>
<td>8th week</td>
<td>经济/economy 农村/rural area 政府/government 国家/national 政策/policy</td>
</tr>
<tr>
<td>中国/China</td>
<td>社会/society 农村/rural area 工作/work 全国/nationwide</td>
</tr>
<tr>
<td>地区/regional</td>
<td>收入/income 改革/reform 土地/land</td>
</tr>
<tr>
<td>城镇/town</td>
<td>房地产/realestate 和平/peace 政策/policy 台湾/Taiwan</td>
</tr>
</tbody>
</table>
trends on TianYa-12261 dataset. The experiment results are shown in Fig. 8. We can see from Fig. 9, negative topics are mainly discussed on TianYa By-talk, the percentage of which is about 75%–80%, and most of them reflect the social phenomena and dissatisfaction, such as “Chongqing officials indecent video event (重庆官员不雅视频事件)”, "Blue Kerr event (蓝可儿事件)", "Li Tianyi rape event (李天一强奸事件)", "North Korean nuclear test (朝鲜核试验)", etc. The strength of positive topics maintain ratio at 25%, mainly discussing the new reform measures of Chinese rural area, economy, culture, politics, and science, etc. The experiment results show many netizens agree and support the implementation of these new reform measures.

The main purpose of the sentiment-topic mixture model is to extract the topic and sentiment-topic from social media reviews. Two kinds of sentiment label values are considered in our experiments, i.e. the positive and the negative. We respectively select three typical sentiment-topic distributions from positive and negative aspects at the fourth week, which are shown in Table 3.

According to Table 3, in the positive sentiment aspect, topic 1 mainly discusses the contents related to the national economic and political reform measures; topic 2 mainly discusses the contents including life and health care reform; topic 3 mainly discusses some common criminal disciplinary incidents, and users showed some indignation for these incidents; topic 3 mainly discusses some insufficiencies in health care, including counterfeit drugs and medical ethics, and many netizens express fears and sadness on these events.

To analyze the sentiment-topic evolutionary discovered by our dNST model, we respectively select a typical positive topic and a typical negative topic which appear over 8 weeks. The twenty words which have biggest weight of the positive topic distributions are listed in Table 4. According to Table 4, we can see that the topic mainly discusses the reform of government policies including culture, politics, and economy, etc. The content of this topic is broad and comparatively stable. Of course, the discussed themes are also affected by the new policies and events, for example, the major theme in week 1–2 is about the conference of health care reform, coming up with "Consolidate the system for basic drugs and grass-roots operation mechanism (关于巩固完善基本药物制度和基层运行新机制的意见)"; the major theme in week 3–4 turns to "Building a new agricultural management system (构建新型农业经营体系)"; then in week 5–6 the major theme is about "Interim Measures on Management of special funds for the development of strategic emerging industries (战略性新兴产业发展专项资金管理暂行办法)"; and in week 7–8 the main theme is "Building a moderately prosperous society (全面建设小康社会)" as well as a meeting of the 12th session of the CPPCC.

The twenty words which have biggest weight of the negative topic distribution are listed in Table 5. According to Table 5, we can see that the topic mainly discusses the phenomenon of social corruption and crime, etc. Official corruption and issues of violating the disciplines are mainly discussed in the week 1–2; in week 3–5, the main discussed contents are related to "Chongqing officials indecent video (重庆官员不雅视频)" incident; in week 6–7, the main discussed contents are "Blue Kerr event (蓝可儿事件)"

Table 5
A typical topic evolution with negative label from week 1 to week 8.

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic word</th>
<th>1st week</th>
<th>2nd week</th>
<th>3rd week</th>
<th>4th week</th>
<th>5th week</th>
<th>6th week</th>
<th>7th week</th>
<th>8th week</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>难以/problem</td>
<td>违法/illegal</td>
<td>视频/video</td>
<td>违法/illegal</td>
<td>违法/illegal</td>
<td>违法/illegal</td>
<td>违法/illegal</td>
<td>违法/illegal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>领导/leader</td>
<td>领导/leader</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
</tr>
<tr>
<td></td>
<td></td>
<td>干预/corruption</td>
<td>干预/corruption</td>
<td>视频/video</td>
<td>视频/video</td>
<td>视频/video</td>
<td>视频/video</td>
<td>视频/video</td>
<td>视频/video</td>
</tr>
<tr>
<td></td>
<td></td>
<td>领导/leader</td>
<td>领导/leader</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
<td>老百姓/civilian</td>
</tr>
</tbody>
</table>
and official corruption issues; in week 7–8, the major discussed theme turns to "Li Tianyi rape event (李天一强奸事件)". Through the comparison with the process of the real social events, we find that the topics discovered by our dNJST are roughly identical to the real social events.

6. Conclusion and future works

To identify sentiment polarity and detect topics from social reviews have been widely and separately researched in recent years. But in many cases, we are concerned about sentiment and topics simultaneously. Some of the recent work have been aware of this limitation and tried to capture sentiments and mixture of topics simultaneously. In this paper, we propose a novel non-parametric joint sentiment topic model (NJST) and dynamic NJST (dNJST) for detecting and tracking dynamic sentiment and topics of Chinese social media. We implement NJST and dNJST with the online variational inference algorithms. Furthermore, to improve the sentiment identification of dNJST, HowNet lexicon is used to determine the sentiment orientation of the words in corpora beforehand. Different from the existing sentiment-topic models based on latent Dirichlet allocation (LDA), NJST and dNJST can determine the topic number automatically. We compare our dNJST method with JST, dJST and NJST on the real Chinese social media data. The experimental results show that dNJST outperform NJST, JST and dJST in extracting topics of specific sentiment orientation, which indicates the effectiveness of our non-parametric sentiment topic mixture model. The details of typical topic analysis also illustrate that dNJST can effectively detect and track dynamic sentiment and topic of Chinese social media.

Despite the success of our proposed method, dNJST still requires to set the time span of each epoch. In the future works, we will consider some other time dependency modes to model the sentiment topic dynamics. Furthermore, we plan compare our method with the dynamic topic and sentiment model such as TSM, the experiments of dJST in other datasets are aslo implemented in the future. In addition, it is difficult task to accurately analyzing the sentiment polarities of irony and figurative language in social media with traditional approaches based on words and their lexical semantics, we will consider to research the sentiment analysis of figurative language in our future work.

Acknowledgments

This research is supported by the National Nature Science Foundation of China under Grants 61472258 and 6140294, Guangdong Natural Science Foundation under Grant 2013010 40012895, Foundation for Distinguished Young Talents in Higher Education of Guangdong, China under Grant 2013LYM_0076, Science and Technology Foundation of Shenzhen City under Grants JCYJ20140509172609162 and JCYJ20130329102032059.

References


