Dynamic Topic Detection Model by Fusing Sentiment Polarity

Xi Ding[1], Lanshan Zhang[2], Ye Tian[1], Xiangyang Gong[1] and Wendong Wang[1]
[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications Beijing, 100876, China
[2]School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications Beijing, 100876, China

Email: dingxi515@163.com,zls326@sina.com,{yetian,xygong,wdwang}@bupt.edu.cn

Abstract

Traditional static topic models mainly focus on the statistical correlation between words, but ignore the sentiment tendency and the temporal properties which may have great effects on topic detection results. This paper proposed an LDA-based dynamic sentiment-topic (DST) model, which could not only detect and track topics but could also analyse the shift of general's sentiment tendency towards certain topic. This model combines the data with the sentiment and dynamic properties of time by maximum likelihood estimation and the sliding window. We use Gibbs sampling method to estimate and update model parameters, and use random EM algorithm for model reasoning. Experiments on real dataset demonstrate that DST model outperforms the existing algorithms.

Keywords: dynamic sentiment topic model, sentiment analysis, Gibbs sampling.

1 Introduction

With the rapid development of Web 2.0 technology, social media, represented by Facebook, Twitter, Blog and Weibo, has really taken off in the last few years. More and more people choose social media as main information exchange platform to publish and access real-time information. In addition to daily communication functions, the huge amount of user generated content (UGC) often also carries people's attitudes towards certain public events or products/services. Existing researches show that compared with traditional media channels, social media data which conveys public sentiment has more important social and economic value for community management agency, enterprise and the general public. Affected by this, the sentiment analysis and topic detection of unstructured social media data has emerged as a hot research hotspot among the various tasks of social networking analysis area.

Researchers who are concentrating on the area of sentiment analysis mainly focus on studying the sentiment polarity (positive, neutral, negative) classification methods of the social media data. Although a host of research achievements have been disclosed [Pang, B. and Lee. L][Aue, A. and Gamon, M][Turney, P, D], they are mostly based on supervised learning methods, and there still existed two limitations. First, a lot of manually annotated samples are needed for parameters adjustment process. Second, sentiment classifier trained from certain topic area often does not apply to other topic areas since the sentiment distribution and topic content are closely related. Besides, an unavoidable problem is that the sentiment word dictionary that was trained from traditional corpus could not be applied to social data that are flexibly expressed with much more emotion icons and disjunctive questions used, considering the huge difference of language style between social media and traditional media. Topic detection derives from TDT (Topic Detection and Tracking) technology, which mainly focuses on detecting and organizing unknown topics from traditional formal expressed text stream. Topic detection task consists of two branches: historical topic detection and online topic detection. For the former, its objective is to dig out the hidden topics from a given corpus with unsupervised clustering means. Each cluster corresponds to a certain topic. While online topic detection determines whether newly arrived text stream belongs to an existing topic or a new topic according to historic information. Compared with newspaper, periodicals and academic report, social media are most non-standard expressed unstructured short texts in real-time formalism. These features make the task of topic detection which takes social media as study object more challenging.

Sentiment analysis and topic detection are taken as two independent research tasks in current social networking analysis fields. However, "sentiment" and "topic" are two highly associated concepts. On one hand, the generation and spread process of sentiment must rely on a certain body, i.e., a specified topic. On the other hand, the change of sentiment would react on its carrier, i.e., certain specified topic, and consequently affect the evolution of topic. Take the event "the loss of communication of flight M370" as example. At the beginning stage, anxiety and trepidation are circulating widely among the people who express concern about this event. As time goes on, the mood gradually evolves into sadness and discontent. Under the influence of this mood, M370 event evolves into a new stage of "relatives' doubt about Malaysian Airline's emergency treatment" from "the lost of aircraft" and "the search and rescue" .Thus it could be seen, the probability distribution of topic is affected by sentiment, and these two concepts are highly relevant.

By fusing the polarity of sentiment, a dynamic sentiment-topic (DST) model is proposed based on Latent Dirichlet Allocation (LDA) in this paper. LDA model assumes that, a document is composed of various topics with different probabilistic combination, within which each topic itself is also a probability distribution of a series of words. For a given corpus, LDA detects all hidden

topics by computing the posterior probability of each word that it belongs to a certain topic. LDA model gains tremendous success in topic analysis area. It is widely applied to social networking analysis areas [Lacoste, S, Sha, F and Jordan, M.][Ramage, D., Hall, D., Nallapati, R., and Manning, C.][Wallach, H., Mimno, D., McCallum, A.]. LDA model divides the document generation process into two stages: "document-topic" generative process and "topic-word" generative process. However, in "document-topic-word" three layer architecture, "topic" is affected by the polarity of "sentiment". The influence is mainly embodied in three aspects. Firstly, "document-topic" probability distribution would change as time passes. The probability distribution of topics within a document at time slot *t* is different with that at time slot *t*-1, however the probability distribution at time slot *t* depends on the probability distribution at time slot *t*-1. Secondly, the probability distribution of the sentiment polarity would change as time passes. The probability distribution of the sentiment polarity at time slot *t* depends on the probability distribution at time slot *t*-1. Thirdly, "topic-word" probability distribution would change as time passes. The probability distribution of word within a topic at time slot *t* is different with that at time slot *t*-1. However the probability distribution at time slot *t* depends on the probability distribution at time slot *t*-1. Based on the above considerations, the DST model proposed in this paper takes the influence that sentiment polarity exerts on the probability distribution of topic and the probability distribution of word into account. Also, DST model adjusts the probability distributions dynamically according to the dependencies in temporal dimension. Thus, for the large amount of social media data, this model could not only detect the hidden topics, but can also classify the topics according to their sentiment polarity. Furthermore, the evolution process of topic and its corresponding sentiment would be deduced.

The main contribution of this paper embodies in the following aspects. Firstly, by fusing sentiment polarity into the topic modelling process, DST model could properly reflect the impact that sentiment exerts on topic detection. Secondly, DST model enhances LDA model by introducing the dynamic characteristics, such that it is more appropriate for those social media data with strong real-time characteristics. Thirdly, experiments on a large real dataset demonstrate that DST model is superior to existing algorithms.

The rest part of this paper is organized as follow. Section 2 gives a brief introduction of the related work. In Section 3, we describe the proposed dynamic sentiment-topic model in detail. In the next section, we conduct an empirical analysis of DST model on a large real dataset. Section 5 concludes the whole paper and discusses the future work.

2 Related Work

Vast majority of research work has been conducted on topic analysis fields. Traditional topic detection method, such as TDT, aims to segment text stream into different news articles. TDT deals principally with formal text stream. Since the huge gap in language style, expressing way between formal text and social media text, TDT cannot work well. Statistical learning theory is the basis of most traditional topic detection algorithms. Latent Semantic Indexing (LST) [Hofmann, T] is a typical statistical learning based topic model, within which Single Value Decomposition (SVD) method is utilized to decompose the document-word matrix, and the subspace with the largest degree of distinction in TF-IDF (Term Frequency-Inverse Document Frequency) feature space then corresponds to the hidden topic. PLSA (Probabilistic Latent Semantic Analysis) which is also called Aspect Model derives from LSI. PLSA enhanced LSI from linear algebra level to probability space. Though PLSA is significant for topic modelling, the disadvantages are also obvious. The number of parameters in PLSA model grows linearly with the size of corpus, such that it may cause over-fitting problem. LDA (Latent Dirichlet Allocation) expands PLSA by introducing Dirichlet as prior distribution. LDA greatly reduces the number of required parameters, meanwhile the over-fitting problem get well solved. However, multinomial distribution based topic model (LDA) still face the over-fitting problem when it comes to short texts since the sparse vocabulary.

In recent years, plenty of improved topic models are proposed based on LDA. DTM (Dynamic Topic Model)[David, M., John, D.] models the natural parameter space with state-space model. It constructs a LDA-like topic model for corpus within each time slot, and these topic models locate in a common markov chain. That means dependency exists between the adjacent topic models, and the topic model within time slot *t* evolves from that within time slot *t*-1. Unlike DTM, TOT [Xueru, W.,McCallum, A.](Topic over Time) model could depict the long term evolution of a certain topic, and it can also predict the topic distribution within a given time slot. However, this model only takes the relationship between word and time into account, but ignores the influence that sentiment polarity imposes on topic modelling. A newly proposed JST (Joint Sentiment-Topic) model [He, Y. and Lin, C.] is weakly supervised. This model integrates sentiment factor into topic modelling process, and it assumes all words in a document embody sentiment tendency, so each topic also has its own sentiment attributes. Strong sentiment would influent people's concern about the topic, and consequently change the topic distribution and word distribution. Based on this assumption, JST model could not only detect the topic hidden in text streams, but could also classify the topics according to its corresponding sentiment polarity. However, the evolution of sentiment polarity of a specified topic is not taken into account in JST model. This limitation leads to the inadaptability when deal with social media data which are timeliness and expressed with rich affections.

Seldom researches consider dynamic characteristics of topics and the influence of sentiment polarity simultaneously in topic modelling process. The dJST (dynamic joint sentiment-topic) model [He, Y., Lin, C., Gao,W., and Wong, K.-F] is most close to our work. It derives from LDA model, but combines both dynamic temporal characteristic of topic and author's sentiment tendency in their model. dJST model estimates the hyper parameters of topic distribution and word distribution. The drawback is that dJST overlooks the fact that the sentiment polarity probability distribution in a document is not

constant, but always changing according to the evolution process of topic.

3 DST Model

3.1 Overview

In this section we would detail the DST Model. It is a generative model, and belongs to a form of Bayesian probability model which is a predictive model that learns by the joint probability density distribution of the data, and then finds the conditional probability distribution. The model uses reasonable mathematical methods, such as maximum likelihood estimation and sliding window that makes the sentimental with the dynamic characteristics and time properties reasonably integrate into the existing traditional static text analysis model—LDA model, so we can get the public's sentimental tendencies of certain topic which is occurred in a period of time and the changes of the sentimental tendencies of certain topic in the trend. Because of the actual situation, the text file generated by social network often goes with the author's sentiment and sentimental tendencies vary among people which go with dynamic changes over time. The framework of our model has five layers where the document layer contains many types of sentimental tendencies, and topic layer is associated with sentiment layer, and the word layer is associated with sentiment and topic layers. Sentimental tendencies layer is the connecting bridge and plays an important role in the five layers. A graphical model of DST model is represented in Figure 1.

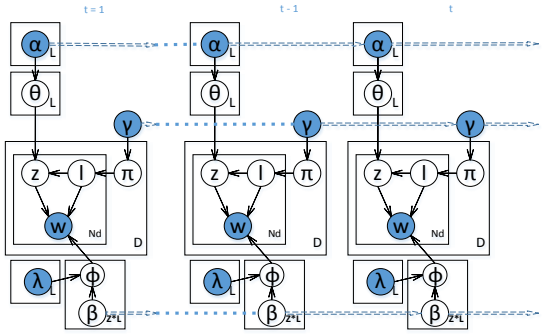

Figure 1: DST MODEL

Model assumes that at each epoch *t*, each of text documents set contains sentimental tendencies, and all the sentimental tendencies is divided into *L* Classes which express as $L=\{l_1,l_2,...,l_L\}$. The object processed by the model is text document collections, which are received with their order of publication timestamps preserved and are represented as $D_t=\{d_{t_1},d_{t_2},...,d_{t_D}\}$ and the text document collections' size is D_t . Appearing in the text document collections at epoch t, all words constitute a vector of word token whose size is V_t , and the word token can be expressed as $V_t=\{W_{t_1},W_{t_2},...,W_{t_V}\}$. Our notations are summarized in Table 1.

Model defines that for each epoch *t*, a document d_t contains *L* classes of sentimental tendencies, and the sentimental tendency obeys the sentiment distribution π_d^t . Each sentimental tendencies l can contain *Z* topics, and

Table 1: Notations used in the paper

Symbol	Description
D_t	number of documents in epoch t
L	number of sentiment labels
Z	number of topics
V_t	number of unique words in the current epoch t
S	number of time slices
γ^t	matrix of $D_t \times L$ dimension , row d represents the priors of the mixing proportion of sentiment in document d_t
α_l^t	matrix of $L \times Z$ dimension, row l represents the priors of the mixing proportion of topics in sentiment label l
$\beta_{l,z}^t$	matrix of $L \times Z \times V_t$ dimension, priors for the word distribution conditioned on sentiment labels and topics
λ_l^t	matrix of $L \times V_t$ dimension which encodes the word prior sentiment polarity information
π_d^t	parameter notation for the sentiment label mixture proportion for document d_t . $\pi^t=\{\pi_d^t\}_{d=1}^{D_t}$ ($D_t \times L$ matrix)
θ_l^t	multinomial distribution over topics for the l*th* sentiment label for document d_t . $\theta^t=\{\{\theta_{d,l}^t\}_{l=1}^{L}\}_{d=1}^{D_t}$ ($D_t \times L \times Z$ matrix)
$\varphi_{l,z}^t$	multinomial distribution over words for the l*th* sentiment label and z*th* topic at epoch $\varphi^t=\{\{\varphi_{l,z}^t\}_{z=1}^{Z}\}_{l=1}^{L}$ ($L \times Z \times V_t$ matrix)
$\delta_{l,z,s}^t$	multinomial word distribution of sentiment label l and topic z with time slice s at epoch t, $\delta_{l,z,s}^t=\{\delta_{l,z,s,w}^t\}_{w=1}^{V_t}$
$\sigma_{l,z,s}^t$	weight vector, $\sigma_{l,z,s}^t=\{\sigma_{l,z,1}^t,...,\sigma_{l,z,S}^t\}$, each of which determines the contribution of time slice s in computing the priors of $\varphi_{l,z}^t$

these topics all obey the topic distribution θ_l^t , and each of these topics is composed of V_t words which obey the word distributed $\varphi_{l,z}^t$. In order to study the dynamic variable data, we assume that the distribution of current sentiment at the current epoch are influenced by sentiment in the past (so the Dirichlet hyper parameter γ^t corresponding to the distribution π_d^t obeys the gamma distribution, ie $\gamma^t \sim \text{gamma}(\xi\gamma^{t-1},\xi)$); and at current epoch *t*, the topic distribution corresponding to the document *d* and the sentimental tendencies *l* is affected by the topic in the past (ie the parameters α_l^t with the corresponding Dirichlet distribution θ_l^t obeys gamma distribution, ie $\alpha_l^t \sim \text{gamma}(\mu\alpha_l^{t-1},\mu)$); same time the current sentiment-topic specific word distributions $\varphi_{l,z}^t$ at epoch *t* are generated according to the word distributions at previous epochs (the word distributions at previous epochs is reflected in two parameters introduced $\delta_{l,z}^t$ 、 $\sigma_{l,z,s}^t$. These

two parameters take different values at different time scales. So our stipulate

$$\beta^{t}_{l,z} = \Sigma\delta^{t}_{l,z} * \sigma^{t}_{l,z,s} \quad (1)$$

Document generation process is mainly determined by the topic and emotional tendencies mark. Formal definition of document process in the model shown in Figure 1 is as follows:

(1) For each sentiment label $l = 1,...,L$

For each topics $z = 1,...,Z$

At the scale S, Compute $\beta^{t}_{l,z} = \Sigma\delta^{t}_{l,z} * \sigma^{t}_{l,z,s}$;

Select the matrix of $L \times Z$ word distribution $\varphi^{t}_{l,z} \sim Dir(\beta^{t}_{l,z})$;

(2) For each document $d_t = 1,...,D_t$

According to $\gamma^{t} \sim \text{gamma}(\xi\gamma^{t-1},\xi)$, compute γ^{t}, Choose a sentiment distribution $\pi^{t}_{d} \sim Dir(\overrightarrow{\gamma^{t}_{l}})$;

For each sentiment label l under document d, according to $\alpha^{t}_{l} \sim \text{gamma}(\mu\alpha^{t-1}_{l},\mu)$, compute α^{t}_{l}, choose a topic distribution $\theta^{t}_{l} \sim Dir(\overrightarrow{\alpha^{t}_{l,z}})$;

For each word $W = 1,...,V$ in document d_t

Choose a sentiment label $l_w \sim Mult(\pi^{t}_{d})$;

Choose a topic $z_w \sim Mult(\theta^{t}_{l,w})$;

Choose a word $W_w \sim Mult(\varphi^{t}_{l_w,z_w})$;

(3) Repeat the above process can get the whole document

Figure 2: Document generation process

3.2 Inference Learning

We present a stochastic EM algorithm to sequentially update model parameters at each epoch using the newly obtained document set and the derived evolutionary parameters. At each iteration step, we use the collapsed Gibbs sampling to estimate potential distribution of sentiment label l and topics z, and use the maximum likelihood estimation method to estimate the hyper parameters.

The total probability of the model for the document set V_t at epoch t given the evolutionary parameters $\delta^{t}_{l,z}$, $\sigma^{t}_{l,z,s}$ and the previous model parameter is:

$$P(w^t, l^t, z^t \mid \overrightarrow{\gamma^t}, \overrightarrow{\alpha^t_l}, \overrightarrow{\delta^t_{l,z}}, \overrightarrow{\sigma^t_{l,z,s}}) = P(l^t \mid \overrightarrow{\gamma^t}) P(z^t \mid l^t, \overrightarrow{\alpha_{t,l}}) P(w^t \mid l^t, z^t, \overrightarrow{\delta^t_{l,z}}, \overrightarrow{\sigma^t_{l,z,s}}) \quad (2)$$

For the first term on the right-hand side of Equation (2), by integrating out π, we obtain

$$p(\vec{l^t} \mid \overrightarrow{\gamma^t}) = \prod_d \frac{\Gamma(\sum_l \gamma^t_l)}{\prod_l \Gamma(\gamma^t_l)} \cdot \frac{\prod_l \Gamma(N^t_{d,l} + \gamma^t_l)}{\Gamma\left(N^t_{d,l} + \sum_l \gamma^t_l\right)} \quad (3)$$

That is equal to $p(\vec{l^t} \mid \overrightarrow{\gamma^t}) = \prod_{d=1}^{D_t} \frac{\Delta(\overrightarrow{n^t_d} + \overrightarrow{\gamma^t_d})}{\Delta(\overrightarrow{\gamma^t_d})}$. Where D_t is the total number of documents in epoch t, $N^t_{d,l}$ is the number of times sentiment label l being assigned to some word tokens in document d at epoch t, $N^t_d = \sum_l N^t_{d,l}$. It is a constant which can be derived from statistical data input and Γ is the gamma function.

For the second term, by integrating out θ, we obtain:

$$p(\overrightarrow{z^t} \mid \overrightarrow{\alpha^t_k}, \overrightarrow{l^t}) = \prod_d \prod_k \frac{\Gamma(\sum_z \alpha^t_{k,z})}{\prod_z \Gamma(\alpha^t_{k,z})} \cdot \frac{\prod_z \Gamma(N^t_{d,k,z} + \alpha^t_{k,z})}{\Gamma\left(N^t_{d,k,z} + \sum_z \alpha^t_{k,z}\right)} \quad (4)$$

That is equal to $\mathrm{p}(\overrightarrow{z^t} \mid \overrightarrow{\alpha^t_k}, \overrightarrow{l^t}) = \prod_{d=1}^{D_t} \prod_{k=1}^{L} \frac{\Delta(N^t_{d,k,z} + \alpha^t_{k,z})}{\Delta(\alpha^t_{k,z})}$.

Where $N^t_{d,k,z}$ is the number of times a word from document d being associated with topic z and sentiment label l at epoch t. It is a constant which can be derived from statistical data input and $N^t_{d,k} = \sum_z N^t_{d,k,z}$.

For the last term, by integrating out φ, we obtain:

$$\mathrm{p}(\overrightarrow{w^t} \mid \overrightarrow{l^t}, \overrightarrow{z^t_k}) = \prod_k \prod_z \frac{\Gamma(\sum_s \delta^t_{k,z,s})}{\prod_w \Gamma(\sum_s \delta^t_{k,z,s} \sigma^t_{k,z,s,w})} \cdot \frac{\prod_w \Gamma(N^t_{k,z,w} + \sum_s \delta^t_{k,z,s} \sigma^t_{k,z,s,w})}{\Gamma\left(N^t_{k,z} + \sum_s \delta^t_{k,z,s}\right)} \quad (5)$$

That is equal to $\mathrm{p}(\overrightarrow{w^t} \mid \overrightarrow{l^t}, \overrightarrow{z^t_k}) = \prod_{k=1}^{L} \prod_{z=1}^{Z} \frac{\Delta(\overrightarrow{n^t_{k,z}} + \delta^t_{k,z,s} \sigma^t_{k,z,s,w})}{\Delta(\delta^t_{k,z,s} \sigma^t_{k,z,s,w})}$. Where $N^t_{k,z,w}$ is the number of times word w appeared in topic z and with sentiment label l at epoch t. It is a constant which can be derived from statistical data input and $N^t_{k,z} = \sum_w N^t_{k,z,w}$.

Gibbs sampling is the use of Markov Chain Monte Carlo method. According to the prior distribution of model parameters and statistical data which is given on the pre-S slots, it sequentially samples each variable of interest, sentiment label l_t and topic z_t through the case that it must meet the detailed balance condition. To meet this condition, let the index $x = (d, n, t)$ and the subscript $\backslash x$ denote a quantity that excludes counts in word position n of document d in epoch t, the conditional posterior for z_x and l_x by marginalizing out the random variables φ, θ, and π is:

$$p(z_x = j, l_x = k \mid \overrightarrow{w^t}, \overrightarrow{z^t_{-x}}, \overrightarrow{l^t_{-x}}, \overrightarrow{\alpha^t}, \overrightarrow{\gamma^t}, \overrightarrow{\delta^t}, \overrightarrow{\sigma^t}) \propto \frac{N^t_{k,j,w_j \backslash x} + \sum_s \delta^t_{k,j,s,w_j} \sigma^t_{k,j,w_j}}{N^t_{k,j \backslash x} + \sum_s \delta^t_{k,j,s}} \cdot \frac{N^t_{d,k,j \backslash x} + \alpha^t_{k,k}}{N^t_{d,k \backslash x} + \sum_j \alpha^t_{k,j}} \cdot \frac{N^t_{d,k \backslash x} + \gamma^t_k}{N^t_{d \backslash x} + \sum_k \gamma^t_k} \quad (6)$$

There are two types of parameters to be estimation which are evolutionary parameters and hyper parameters.

A. The evolutionary parameters estimation

We use the fixed-point iteration method to estimate the weight vector $\delta^{t}_{l,z}$ directly from data by maximizing the joint distribution in Equation (2). The update formula is:

$$(\delta^t_{k,j,s})^{new} \leftarrow \frac{\delta^t_{k,j,s} \sum_w \sigma^{t}_{d,k,j,w} A}{B} \quad (7)$$

Where,

$$A = \phi(N^t_{d,k,j,w} + \sum_{s'} \delta^t_{k,j,s'} \sigma^{t}_{d,k,j,w}) - \phi(\sum_{s'} \delta^t_{k,j,s'} \sigma^{t}_{d,k,j,w}) \quad (8)$$

$$B = \phi(N^t_{d,k,j} + \sum_{s'} \delta^t_{k,j,s'}) - \phi(\sum_{s'} \delta^t_{k,j,s'}) \quad (9)$$

And $\phi(.)$ is the digamma function defined by $\phi(x)=\frac{\partial \log\Gamma(x)}{\partial x}$.

The evolutionary parameter $\sigma^{t}_{d,k,j,w}$ accounts for the historical word distributions at different time slices. So we estimate $\{\sigma^{t}_{d,k,j,w}\}_{w=1}^{V_t}$ the word distribution in topic *j* and sentiment label *k* at time slice *s*, which can be calculated as follows:

$$\sigma^{t}_{d,k,j,w}=\frac{C^{t}_{k,j,s,w}}{\sum_{w}C^{t}_{k,j,s,w}} \quad (10)$$

Where $C^{t}_{k,j,s,w}$ is the expected number of times word *w* is assigned to sentiment label *k* and topic *j* at time slice *s*. At different time scales, its value is different. Thus $C^{t}_{k,j,s,w}$ can be obtained directly from the count $\hat{N}^{t'}_{k,j,w}$,that is $C^{t}_{k,j,s,w}=\hat{N}^{t'=t-s}_{k,j,w}$, the expected number of times word *w* is associated with sentiment label *l* and topic *z* at epoch t', which can be calculated by:

$$\hat{N}^{t'}_{k,j,w}=N^{t'}_{k,j,w}\cdot\frac{N^{t'}_{k,j,w}+\sum_{s'}\delta^{t'}_{k,j,s'}\sigma^{t'}_{d,k,j,w}}{N^{t'}_{k,j}+\sum_{s'}\delta^{t'}_{k,j,s'}} \quad (11)$$

Where $N^{t'}_{k,j,w}$ is the observed count for the number of times word *w* is associated with sentiment label *l* and topic *z* at epoch t'.

B. Hyper parameters (α^{t} , γ^{t}) estimation

We estimate α^{t} and γ^{t} from data using maximum-likelihood as part of the online stochastic EM algorithm.

A common practice for the implementations of topic models is to use symmetric Dirichlet hyper parameters. However, it has been found that an asymmetric Dirichlet prior over the per-document topic proportions has substantial advantages over a symmetric prior. So when first entering a new epoch, we initialize the asymmetric α^{t}_{1}=0.01， γ^{t}=0.01, Afterwards according that γ^{t} obeys gamma distribution $\gamma^{t}\sim\text{gamma}(\xi\gamma^{t-1},\xi)$ which lead to the probability function of γ^{t} is $p(\overrightarrow{\gamma^{t}}\mid\overrightarrow{\gamma^{t-1}},\xi)=\prod_{l}\frac{\xi^{\xi\gamma_l^{t-1}}(\gamma_l^{t})^{\xi\gamma_l^{t-1}-1}\exp(-\xi\gamma_l^{t})}{\Gamma(\xi\gamma_l^{t-1})}$ and α^{t}_{1} obeys gamma distribution $\alpha^{t}_{1}\sim\text{gamma}(\mu\alpha^{t-1}_{1},\mu)$ which lead to the probability function of γ^{t} is $p(\overrightarrow{\alpha_k^{t}}\mid\overrightarrow{\alpha_k^{t-1}},\mu)=\prod_{z}\frac{\mu^{\mu\alpha_{k,z}^{t-1}}(\alpha_{k,z}^{t})^{\mu\alpha_{k,z}^{t-1}-1}\exp(-\mu\alpha_{k,z}^{t})}{\Gamma(\mu\alpha_{k,z}^{t-1})}$

For every 40 Gibbs sampling iterations, α^{t}_{1} and γ^{t} are learned directly from data using maximum-likelihood estimation.

$$(\alpha_{k,z}^{t})^{new}\leftarrow\frac{\mu\alpha_{k,z}^{t-1}-1+\alpha_{k,z}^{t}\sum_{d}(\phi(N_{d,k,z}^{t}+\alpha_{k,z}^{t})-\phi(\alpha_{k,z}^{t}))}{\mu+\sum_{d}(\phi(N_{d,k}^{t}+\sum_{z'}\alpha_{k,z'}^{t})-\phi(\sum_{z'}\alpha_{k,z'}^{t}))} \quad (12)$$

ALGORITHM: Gibbs sampling procedure for DST MODEL

Input: Number of topics Z、number of sentiment labels L, number of time slices S, word prior polarity transformation matrix λ , epoch $t\in\{1,2,\ldots,\text{maxEpochs}\}$, a stream of documents $D_t=\{d_{t_1},d_{t_2},...,d_{t_D}\}$;

Output: DST model;

Sort documents according to their time stamps;
FOR $t = 1$ to maxEpochs **do**
  IF $t == 1$ **THEN**
    Set $\beta^{t}_{l,z}=\lambda\times 0.01$;
  End
  ELSE $\delta^{t}_{l,z}$, $\sigma^{t}_{l,z,s}$
    Set $\delta^{t}_{l,z}=\delta^{t-1}_{l,z}$;
    Set $\sigma^{t}_{l,z,s}=1/S$;
    Set $\beta^{t}_{l,z}=\Sigma\delta^{t}_{l,z}*\sigma^{t}_{l,z,s}$;
  END
    Set γ^{t} = (0.05 × Average document length)/*L* ;
    Set α^{t} = (0.05 × Average document length)/($L\times Z$) ;
    Initialize π^{t} , θ^{t} , φ^{t} , and all count variables;
    Initialize sentiment label and topic assignmentrandomly for all word tokens in D_t;
  FOR i = 1 *to* max *Gibbs Sampling Iterations* **DO**
    [π^{t} , θ^{t} , φ^{t} , l_t, z_t] = GibbsSampling (D_t, α^{t} , $\beta^{t}_{l,z}$, γ^{t});
    FOR *every 40 Gibbs sampling iterations* **DO**
      Update α^{t} using Equation（**12**）;
      Update γ^{t} using Equation（**13**）;
      Update $\delta^{t}_{l,z}$ using Equation（**7**）;
      Set $\beta^{t}_{l,z}=\Sigma\delta^{t}_{l,z}*\sigma^{t}_{l,z,s}$;
    END
    FOR *every 200 Gibbs sampling iterations* **DO**
      Update π^{t} , θ^{t} , φ^{t} with the new sampling results;
    END
  END
    Update $\beta^{t}_{l,z}$, using Equation（**10**）;
END

Figure 3: Online Stochastic EM Algorithm

$$(\gamma_l^{t})^{new}\leftarrow\frac{\xi\gamma_l^{t-1}-1+\gamma_l^{t}\sum_{d}(\phi(N_{d,l}^{t}+\gamma_l^{t})-\phi(\gamma_l^{t}))}{\xi+\sum_{d}(\phi(N_{d}^{t}+\sum_{l'}\gamma_{l'}^{t})-\phi(\sum_{l'}\gamma_{l'}^{t}))} \quad (13)$$

The detail description of the online stochastic EM algorithm for the DST model is given in Algorithm 1.

4 Experiment

To evaluate the proposed solution, we conducted experiments using a real dataset collected from Sina Weibo.

4.1 Dataset Preparation

We implement a Sina Weibo crawler using the APIs provided by Sina Corp. (http://open.weibo.com/) to collect the experiment dataset.

From January 23, 2013 to February 23, 2013, the data crawling process continued for 31 days. Table 1 illustrates the statistics of the dataset. Considered that many public

events happed in the period of December 19, 2011 to December 26, 2011, we select the Weibo data generated in that time period as our test dataset.

Table 2: Statistics of Collected Dataset

Statistical Measure	Value
Number of Users	531,935
Number of Posts	25,838,961
Number of Comments	14,498,416

4.2 DST Model V.S Existing Models

In order to evaluate the effectiveness of our proposed model in topic detection, we compared the performance of DST with LDA and DST with metric of perplexity.

Perplexity is a metric which was mostly used in language modelling. It could measure a language model's prediction ability when applied on an uncharted dataset. Perplexity is defined as the reciprocal geometric mean of the likelihood of a test corpus given a trained model's Markov Chain state *M*. In this paper, we adopted the definition of perplexity as below [He, Y., Lin, C., Gao,W., and Wong, K.-F]:

$$Perplexity = P(D_t \mid M) = \exp\left(-\frac{\sum_{d=1}^{D_t} \log p(\tilde{w}_d^t \mid M)}{\sum_{d=1}^{D_t} \tilde{N}_d^t}\right) \quad (14)$$

Here, this metric means the per-word predictive perplexity of the unseen test set D_t at each epoch t based on the previously trained model M .

In the experiments we studied the influence of the topic number settings on the DST model performance. With the number of time slices fixed at S = 4, we vary the topic number $T \in \{1, 5, 10, 15, 20\}$.

Figure 4 shows the average per-word perplexity over epochs with different number of topics. DST has lower perplexities than all the other models with the increased number of topics.

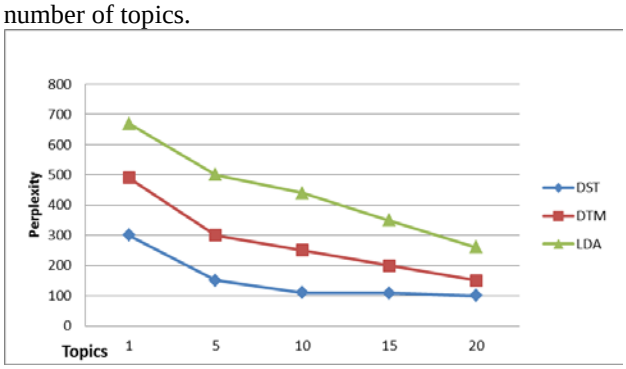

Figure 4: Perplexity VS. umber of topics

The average perplexity for each epoch with the number of time slices set to 4 and the number of topics set to 15 for the DST models is shown in Figure 5. In addition, we also plot the perplexity results of LDA and DTM. LDA only uses the data in the previous epoch for training and hence it does not model dynamics while DTM uses all past data for model learning. We set the number of topics to 15 for both DTM and DST. For LDA, the number of topics was set to 3 corresponding to positive, negative, and neutral sentiment labels.

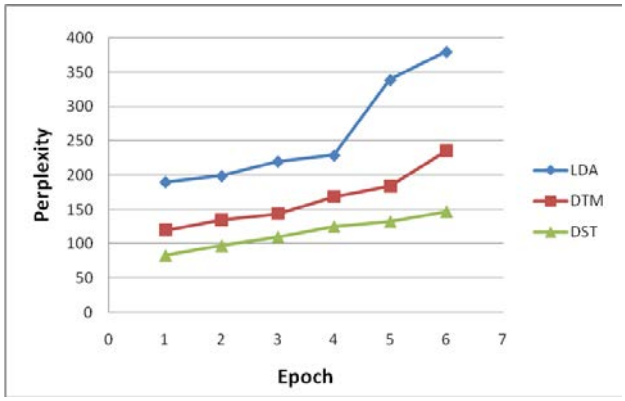

Figure 5: Perplexity vs. Number of epochs

Figure 5 shows that LDA has the highest perplexity values followed by DTM and DST. The perplexity gap between DTM and the DST models increases with the increasing number of epochs. The variants of DST models have quite similar perplexities.

5 Conclusion

In this paper, we proposed a dynamic Sentiment Topic (DST) model, which could models the topics and its corresponding sentiment polarities from huge amount of user generated content data. By taking the sentiment effects into account, DST could properly reflect the impact that sentiment exerts on topic detection. Also, it enhances LDA model by introducing the dynamic characteristics, such that it is more appropriate for those social media data with strong real-time characteristics. Experiments results on a large real dataset show that compared with existing models, DST performances better in terms of perplexity. In future work, we may work on reducing the temporal and spatial overhead of this model.

Acknowledgements

This research was partially supported by National Natural Science Foundation of China under Grant No. 61402045, Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20130005110011, and Fundamental Research Funds for the Central Universities under Grant No. 2014RC1301.

References

Pang, B., Lee. L.(2008): Opinion mining and sentiment analysis. Found. *Trends Inf. Retr*. **2**(2): 1–135.

Aue, A., Gamon, M. (2005): Customizing sentiment classifiers to new domains: a case study. *Proc. of RANLP*.

Turney, P, D. (2001): Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proc. of ACL*, 417–424.

Lacoste, S, Sha, F and Jordan, M. (2008): DiscLDA: Discriminative learning for dimensionality reduction and classification. *Proc. of NIPS.*

Ramage, D., Hall, D., Nallapati, R., and Manning, C. (2009): Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proc. of EMNLP:248-256.*

Wallach, H., Mimno, D., McCallum, A. (2009): Rethinking LDA: Why priors matter. *Proc. of Topic Models: Text and Beyond Workshop in Neural Information Processing Systems Conference.*

Hofmann, T. (1999): *Probabilistic latent semantic indexing. Proc. of SIGIR*:50-57.

David, M., John, D.,(2006): Dynamic Topic Models. *Proc. of the 23rd International Conference on Machine Learning.*

Xueru, W.,McCallum, A.(2005): Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. *Proc. of ACM SIGKDD.*

He, Y. and Lin, C. (2012): Online sentiment and topic dynamics tracking over the streaming data. *Proc. of the ASE/IEEE International Conference on Social Computing (SocialCom'12).*

He, Y., Lin, C., Gao,W., and Wong, K.-F.(2013): Dynamic joint sentiment-topic model. *ACM Trans. Intell. Syst.Technol. 5, 1, Article 6.*