Emotional Contagion-Based Social Sentiment Mining in Social Networks by Introducing Network Communities

Xiaobao Wang
College of Intelligence and Computing, Tianjin University
Tianjin, China
wxbxmt@tju.edu.cn

Di Jin*
College of Intelligence and Computing, Tianjin University
Tianjin, China
jindi@tju.edu.cn

Mengquan Liu
College of Intelligence and Computing, Tianjin University
Tianjin, China
liumengquan@tju.edu.cn

Dongxiao He
College of Intelligence and Computing, Tianjin University
Tianjin, China
hedongxiao@tju.edu.cn

Katarzyna Musial
University of Technology Sydney
Sydney, Australia
Katarzyna.Musial-Gabrys@uts.edu.au

Jianwu Dang
Japan Advanced Institute of Science and Technology
Japan
jdang@jaist.ac.jp

ABSTRACT

The rapid development of social media services has facilitated the communication of opinions through online news, blogs, microblogs, instant-messages, and so on. This article concentrates on the mining of readers' social sentiments evoked by social media materials. Existing methods are only applicable to a minority of social media like news portals with emotional voting information, while ignore the emotional contagion between writers and readers. However, incorporating such factors is challenging since the learned hidden variables would be very fuzzy (because of the short and noisy text in social networks). In this paper, we try to solve this problem by introducing a high-order network structure, i.e. communities. We first propose a new generative model called Community-Enhanced Social Sentiment Mining (CESSM), which 1) considers the emotional contagion between writers and readers to capture precise social sentiment, and 2) incorporates network communities to capture coherent topics. We then derive an inference algorithm based on Gibbs sampling. Empirical results show that, CESSM achieves significantly superior performance against the state-of-the-art techniques for text sentiment classification and interestingness in social sentiment mining.

CCS CONCEPTS

• **Information systems → Data mining**; **World Wide Web**.

KEYWORDS

social sentiment; social network; emotional contagion; community

*Corresponding author: Di Jin, jindi@tju.edu.cn

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357941

ACM Reference Format:
Xiaobao Wang, Di Jin, Mengquan Liu, Dongxiao He, Katarzyna Musial, and Jianwu Dang. 2019. Emotional Contagion-Based Social Sentiment Mining in Social Networks by Introducing Network Communities. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3357384.3357941

1 INTRODUCTION

Measuring public opinions about social events, political movements, company strategies, marketing campaigns, and product preferences is challenging [5]. With the transformation in Web usage, from information consumption to information production and sharing, numerous social media services have emerged. Online users can now conveniently express their opinions through news portals, reviews, and microblogs/tweets. As an important media to report events happening around the world, news portals become increasingly popular in conveying positive or negative sentiments underlying an opinion. In such news portals, each article is associated with ratings shared by readers who vote over a set of predefined emotion labels/tags. Such response from readers is known as "social emotion", shown in Figure 1(a), the amount of which has kept increasing rapidly [3, 21]. Mining social emotion can help authors understand how their work will emotionally influence their readers. Social emotion mining has therefore attracted a large amount of attention from researchers in natural language processing and machine learning.

Recently, numerous sentiment topic models [23, 24], which borrow the machinery of latent topic models such as the Latent Dirichlet Allocation (LDA) [4], have been proposed for mining social emotion in social media like news portals. While enjoying immense success, these methods still have large drawbacks. First, existing methods treat the readers' votes as observed variables. However, most social networks such as Reddit and Twitter do not have the emotion voting information after reading microblogs/tweets. Instead, they only have textual comments of readers, as shown in Figure 1(b). The first thing we need to make clear is that, "emotion" refers to a complex psychological state, while "sentiment", on the other hand, can be defined as a mental attitude or an opinion that has been influenced by emotion and expressed in the text. Since

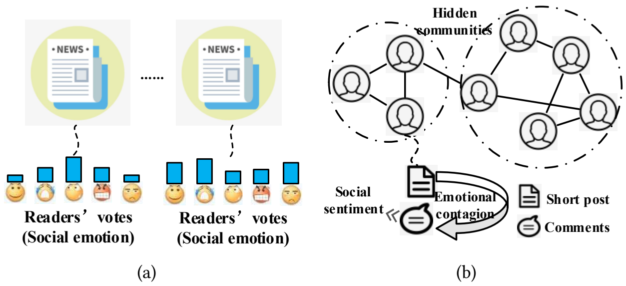


Figure 1: Comparison of applicable social medias of existing methods and our approach. (a) Existing methods treat the content of news and readers' votes as input variables to forecast social emotion of unvoted News. (b) Our approach can be applied to more general social medias, e.g. Reddit and Twitter, where each user can submit a post and other users may comment on it with certain social sentiments.

it is too hard to detect emotion from only short and noisy textual comments, we transform the task of "social emotion" mining into "social sentiment" mining. Here social sentiment of readers are hidden variables implicit in the textual comments, which require further exploitation, as shown in Figure 1(b). Further, Social Psychology studies have shown that people exhibit emotional contagion in the process of social interaction [9]. Recently, in the field of Social Media analysis [28, 32] , it has been found that the characteristics of emotional contagion also exist in the interaction between Internet users' microblogs. That is, sentiment expressed on a post written by a user can influence emotion of readers, which in turn influences social sentiments of readers expressed in the content of comments, as shown in Figure 1(b).

However, to the best of our knowledge, no existing method has considered the characteristics of emotional contagion when mining social sentiment. This is mainly because considering emotional contagion in social networks without voting information is challenging. Not only topics, but also sentiments of authors and readers are all hidden in the texts, which require further exploitation simultaneously. Nevertheless, compared with long formal text, microblogs and their comments are much shorter, noisier and have various expression styles, e.g. 'lol' or 'It is so cooooooooool!', which exacerbates the problem of vocabulary sparsity. The hidden variables exploited from such text would be very fuzzy.

Fortunately, social networks provide different types of metadata, such as user relationships, which can complement text information and can be leveraged to improve the accuracy of hidden variables exploitation: according to the homophily theory a pair of linked users have similar interest. By doing so, networks may be able to help mitigate the semantic fuzziness of topics, and improve the detection accuracy of authors and readers' sentiments. However, we also find that using links directly may not perform well on improving hidden variables exploitation since this low-order network information is often sparse and noisy [15, 19, 22]. That is to say, some users having similar interests may not be connected due to the sparseness of the network. In special cases, this may even destroy the quality of these hidden variables.

On the other hand, community structure, as a higher-order network information, as shown in Figure 1(b), may be suitable to address this problem. Users in the same community of a social network are connected more tightly than those in different communities. In addition, it is believed that users in the same community are often topic-related [12, 13, 27]. Therefore, instead of using links directly, it may be better to use a community, to make users in the same community own similar topic distributions. By doing so, the problem brought by sparsity and noise of networks may be alleviated.

In this paper, we present a generative social sentiment mining model, namely Community-Enhanced Social Sentiment Mining (CESSM), which considers the emotional contagion between writers and readers, and incorporates network structure. In addition, it can be applied to a majority of social networks and is very general. CESSM consists of two main components connected by a probability transition mechanism. The first is the text component, including microblogs and their comments. The text component contains sentiments of writers and social sentiments of readers as hidden variables (with respect to the topics) and incorporates the influences between them into social sentiment modeling. The second is the network component, which plays the role of alleviating the problem of fuzziness of topics (with the help of topic-related network communities) and thus improves the accuracy of social sentiment.

To summarize, we make the following contributions:

- We identify the problem of social sentiment mining in social networks without users voting information. To the best of our knowledge, such a new angle has not been studied before.
- We propose a latent model to uncover the hidden communities, topics, sentiments and social sentiments as well as consider emotional contagion and incorporate network communities. It exhibits improved capacity to model social sentiment in social networks.
- We decouple the model into several components, based on which an efficient inference algorithm is developed.
- An effective text sentiment classification is exhibited which shows significant superiority over the mining accuracy of social sentiment. A verification experiment demonstrates the effect of introducing network communities. We also study the real-world application of CESSM on Reddtit.

The rest of the paper is organized as follows. We introduce the generative model in Section 2, present the basic inference algorithm in Section 3, present experimental results in Section 4 and then give the related work in Section 5. We conclude with some discussions in Section 6.

2 THE JOINT LATENT MODEL

The central task of this paper is to extract social sentiment patterns from social networks such as e.g. Reddit and Twitter,. In this section, we first formulate the social sentiment modeling problem from the emotional contagion perspective. We then propose CESSM (Community-Enhanced Social Sentiment Mining), a comprehensive latent variable model, to address the problem.

2.1 Problem Formulation

The notations used in this paper are listed in Table 1.

Definition 1 (Social Network). Consider a social network G = (U, E). U is a set of n users. The link set E denotes interactions between users and can be derived from various types of user interactions such as following, retweeting and commenting. A link (i, i') ∈ E represents the existence of an interaction from user i to i', e.g. i' once commented i.

Each user $i \in U$ is associated with a set of posts, denoted as D_i, where each post d_{ij} contains a bag of words, along with comment d'_{ij} also represented by a bag of words from a given vocabulary. We use E_i to denote the set of links of user i.

A community is a collection of users with more intense interactions (than those in the rest of the global network). It can be characterized not only by interaction link structures, but also the content (i.e. posts) generated by its members who have similar interest. While existing works on community modeling generally assume one community corresponding to one interest/topic, we associate each community with a topic distribution representing its different topical interests. Based on this idea, here we propose a new definition of community.

Definition 2 (Community). A community $c \in \{1, \ldots, |C|\}$ has two components: a multinomial distribution over topics $\boldsymbol{\eta}_c$, where each component η_{ck} represents the probability of a post from a member of the community related to the corresponding topic k; and a multinomial distribution over nodes $\boldsymbol{\theta}_c$ where each component $\theta_{ci'}$ denotes the probability of the node in the community c connected to node i'.

Classically, most existing works like LDA characterize a topic by the words that co-occur frequently in the same documents. To explore sentiments of writers and social sentiments of readers implicit in posts and comments respectively, we extend topics with two new layers, which creates topic-sentiment correlations and topic-social sentiment correlations, one for modeling sentiments of writers and another for modeling social sentiments of readers. Roughly speaking, topic-sentiment correlation is characterized by the co-occurrence frequency of topic words with sentiment-bearing words (the same as topic-social sentiment correlation). Topic is defined as follows.

Definition 3 (Topic). A topic $z \in \{1, \ldots, |Z|\}$ has two dual distributions over words, one under each sentiment polarity s $\in \{1, \ldots, |S|\}$ of writers, denoted as $\boldsymbol{\varphi}_{zs}$, where each component denotes the probability of a word in one post generated by topic k and sentiment s, and another under each social sentiment l $\in \{1, \ldots, |L|\}$ of readers denoted as $\boldsymbol{\mu}_{zl}$, where each component denotes the probability of a word in comments generated by topic k and social sentiment l. Here we set sentiment polarity $|S|$ =$|L|$ = 3 (i.e. neutral, negative and positive).

Definition 4 (Social Sentiment). Social sentiment is the sentimental responses of readers expressed in the content of comments after reading posts submitted by writers, denoted as hidden variables l.

Note that we aim to uncover social sentiment in social networks without users voting information based on topic models, which considers emotional contagion between writers and readers and incorporates network communities. The communities, topics, sentiments of writers and social sentiments of readers are all latent factors to be extracted, and we also need to uncover the dependencies between them.

Table 1: Notations used in this paper

Symbol	Description
	Observed variables
U, E, D, D'	Set of users, links, posts and comments
d_{ij}, d'_{ij}	The jth post by user i and its comments
$\lvert d_{ij}\rvert, \lvert d'_{ij}\rvert$	Number of words in the post d_{ij} and comment d'_{ij}
$e_{ii'}$	Indicator of the existence of link (i, i')
w_{iju}	The uth word in the post d_{ij}
m_{ijp}	The pth word in the comments d'_{ij}
	Latent variables
C, Z	Set of latent communities and topics
S, L	Set of latent sentiments and social sentiments
c_i	Community associated with user i
z_{ij}	Topic associated with post d_{ij}
s_{ij}	Sentiment associated with post d_{ij}
l_{ijp}	Social sentiment associated with word m_{ijp} in d'_{ij}
	Model parameters
$\boldsymbol{\pi}$	Multinomial distribution over communities
$\boldsymbol{\theta}_c$	Multinomial distribution over users specific to community c
$\boldsymbol{\eta}_c$	Multinomial distribution over topics specific to community c
$\boldsymbol{\sigma}_z$	Multinomial distribution over sentiments specific to topic z
$\boldsymbol{\varphi}_{zs}$	Multinomial distribution over words specific to topic z and sentiment s
$\boldsymbol{\lambda}_{zs}$	Multinomial distribution over sentiments specific to topic z and sentiment of writers s
$\boldsymbol{\mu}_{zl}$	Multinomial distribution over words specific to topic z and social sentiment of readers l
	Hyper-parameters
$\alpha, \beta, \varepsilon, \omega$	Dirichlet priors
ρ, γ, ζ	Dirichlet priors

Our model is different from prior works which need voting information for each document and ignore the emotional contagion between writers and readers. Next, we develop our new model by describing its general structure as well as five properly decoupled components.

2.2 Model Structure

CESSM is a generative model jointly over posts, comments, and network. It uncovers latent communities, topics, sentiments of writers and social sentiments of readers in a unified way.

Although some of its building blocks are inspired by recent successful attempts, including Mixture Model [20] over networks, and Topic-Sentiment Model [6] over text, CESSM significantly goes beyond those by having more comprehensive input features and

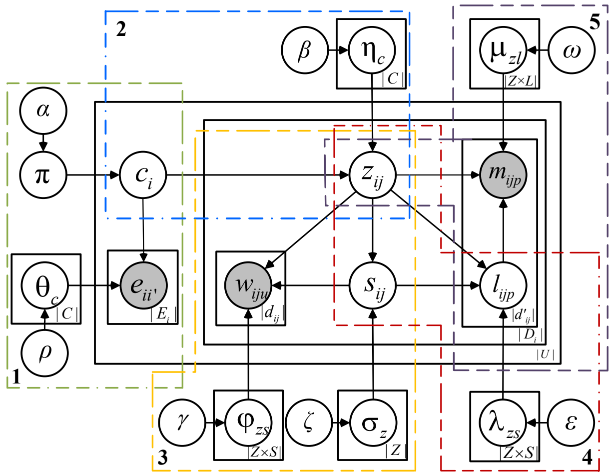


Figure 2: Graphical Model Representation of CESSM. Part 1 (green box) denotes network component describing communities. Part 2 (blue box) denotes probabilistic transition mechanism between communities and topics. Part 3 (yellow box) describes posting. Part 4 (red box) describes emotional contagion. Part 5 (purple box) describes commenting.

powerful modeling ability. Compared with previous mining social emotion models, CESSM better fits the social networks setting by recovering accurate community interest and refining user posts and comments treatment.

CESSM aims to model three basic types of user behaviors, i.e. posting (where writers generate text), commenting (where readers generate text) and social interactions (which form links). By jointly considering the three types of user behaviors with properly separated generative process, CESSM naturally combines content and network data while still keeping the model tractable (we will discuss the time complexity later).

2.3 Individual Components

Figure 2 shows the five components from CESSM together: the *network component* accounts for the link structure; the *probabilistic transition* captures the correlation between communities and topics; the *posting component* and *commenting component* uncovers topic-sentiment correlation on posts and topic-social sentiment correlation on comments; the *emotional contagion* component models the influences between sentiments of writers and social sentiments of readers.

Network component. We associate each user i with a community assignment c_i. We define $\theta_{c_i,i'}$ as the probability of user i in community c_i connected to node i'. In essence $\boldsymbol{\theta}_{c_i}$ represents the "preferences" of vertices in community c_i about which other nodes they link to.

Probabilistic transition component. After associating each user i with a community assignment c_i, for each post $d_{ij} \in D_i$, we use a multinomial distribution $\boldsymbol{\eta}_{c_i}$ to sample z_{ij}. By doing so, the users in the same community will have the same distribution over topics indicating the similar interest, thus helping to get coherent topics.

Posting component. Each post d_{ij} contains a bag of words $\{w_{ij_1}, \ldots, w_{ij|d_{ij}|}\}$, where $|d_{ij}|$ denotes the length of the post. In traditional topic-sentiment models such as TS [6], a document is associated with a mixture of topics and sentiments, and each word has a topic label and a sentiment label. This is reasonable for long documents such as academic papers. However, in social media like microblogging, a post is usually short, and thus is most likely to be about a single topic [11], and the same idea can be applied for sentiment, i.e., a single sentiment. We therefore associate d_{ij} with, 1) a single topic variable z_{ij} drawn from $\boldsymbol{\eta}_{c_i}$ to indicate its topic and 2) a single sentiment variable s_{ij} drawn from $\boldsymbol{\sigma}_{z_{ij}}$ to indicate its sentiment. The words of posts are then sampled from the corresponding word distribution $\boldsymbol{\varphi}_{z_{ij},s_{ij}}$.

Emotional contagion component. Based on emotional contagion, the emotions of readers would be affected by the sentiment expressed on the post, which in turn will influence the sentiments expressed in the content of comments. In addition, readers have different opinions on different topics. We therefore draw the sentiment label l_{ijp} from a multinomial distribution $\boldsymbol{\lambda}_{z_{ij},s_{ij}}$ under topic z_{ij} and sentiment s_{ij}.

Commenting component. Each post d_{ij} has several comments of different users gathered in d'_{ij}, which also contains a bag of words $\{m_{ij1}, \cdots, m_{ij|d'_{ij}|}\}$ where $|d'_{ij}|$ denotes the length of d'_{ij}. Note that d'_{ij} gathers all comments of different readers on this post to a single document, thus it is a relatively long text and contains multiple social sentiments from different readers. It is reasonable to assume that each word in d'_{ij} has a social sentiment label l_{ijp}. Besides, the topic of the comment d'_{ij} and the post d_{ij} should be consistent, because readers always talk about the same topic as the post written by the writer. Then the words are drawn from $\boldsymbol{\mu}_{z_{ij},l_{ijp}}$.

The generative process is summarized in Alg. 1.

3 MODEL INFERENCE

This section develops an efficient inference method for CESSM. We first present the basic inference algorithm using a sampling method, then analyze the time complexity.

3.1 Approximate Inference

Below we describe the inference algorithm for CESSM based on collapsed Gibbs Sampling.

Given a social network $G = (U, E)$ with a set of posts D, a set of comments D' and the pre-defined hyper-parameters $\alpha, \rho, \beta, \zeta, \gamma, \varepsilon$ and ω, CESSM specifies the following full posterior distribution:

$$\begin{aligned} P\,(\mathrm{c}, \mathrm{z}, \mathrm{s}, \mathrm{l}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\mu} | E, D, D' \rho, \alpha, \beta, \gamma, \zeta, \varepsilon, \omega) \propto \\ P\,(\boldsymbol{\pi}|\alpha)\, P\,(\boldsymbol{\theta}|\rho)\, P\,(\boldsymbol{\eta}|\beta)\, P\,(\boldsymbol{\sigma}|\zeta)\, P\,(\boldsymbol{\varphi}|\gamma)\, P\,(\boldsymbol{\lambda}|\varepsilon) \\ \cdot P\,(\boldsymbol{\mu}|\omega)\, P\,(\mathrm{c}|\boldsymbol{\pi})\, P\,(E|\mathrm{c}, \boldsymbol{\theta})\, P\,(\mathrm{z}|\boldsymbol{\eta}, \mathrm{c})\, P\,(\mathrm{s}|\boldsymbol{\sigma}, \mathrm{z}) \\ \cdot P\,(w_D|\mathrm{z}, \mathrm{s}, \boldsymbol{\varphi})\, P\,(\mathrm{l}|\mathrm{z}, \mathrm{s}, \boldsymbol{\lambda})\, P\,(m_{D'}|\mathrm{z}, \mathrm{l}, \boldsymbol{\mu})\,, \end{aligned} \tag{1}$$

where w_D is the words in the post set, $m_{D'}$ is the words in the comments set, and the constant of proportionality is the marginal likelihood of the observed data.

The task of posterior inference for CESSM is to determine the probability distribution of the hidden variables given the observed words, time stamps and network. However, exact inference is intractable due to the complexity of calculating the normalizing constant in the above posterior distribution.

Algorithm 1 Generative Process for CESSM

1. Sample the distribution over communities, $\boldsymbol{\pi}|\alpha \sim Dir(\alpha)$
2. For each community c = 1, 2, ..., $|C|$,
   (a) Sample the distribution over nodes, $\boldsymbol{\theta}_c|\rho \sim Dir(\rho)$.
   (b) Sample the distribution over topics, $\boldsymbol{\eta}_c|\beta \sim Dir(\beta)$.
3. For each topic z= 1, 2, ..., $|Z|$,
   (a) Sample the distribution over sentiments, $\boldsymbol{\sigma}_z|\zeta \sim Dir(\zeta)$.
   (b) For each sentiment of writers s = 1, 2, ..., $|S|$,
      i. Sample the distribution over words, $\boldsymbol{\varphi}_{zs}|\gamma \sim Dir(\gamma)$.
      ii. Sample the distribution over social sentiments, $\boldsymbol{\lambda}_{zs}|\varepsilon \sim Dir(\varepsilon)$.
   (c) For each social sentiment of readers l = 1, 2, ..., $|L|$,
      i. Sample the distribution over words, $\boldsymbol{\mu}_{zl}|\omega \sim Dir(\omega)$.
4. For each user i = 1, 2, ..., $|U|$,
   (a) Sample community indicator, $c_i|\boldsymbol{\pi} \sim Mul(\boldsymbol{\pi})$.
   (b) For each link $(i, i') \in E_i$,
      Sample link, $e_{ii'}|\boldsymbol{\theta}_{c_i} \sim Mul(\boldsymbol{\theta}_{c_i})$.
   (c) For each post d_{ij}, j = 1, 2,...,
      i. Sample topic indicator, $z_{ij}|\boldsymbol{\eta}_{c_i} \sim Mul(\boldsymbol{\eta}_{c_i})$.
      ii. Sample sentiment indicator, $s_{ij}|\boldsymbol{\sigma}_{z_{ij}} \sim Mul(\boldsymbol{\sigma}_{z_{ij}})$.
      iii. For each word in the post u = 1, 2,...,
         A. Sample word, $w_{iju}|\boldsymbol{\varphi}_{z_{ij},s_{ij}} \sim Mul(\boldsymbol{\varphi}_{z_{ij},s_{ij}})$.
      iv. For each word in the comments, p = 1, 2,...,
         A. Sample social sentiment indicator, $l_{ijp}|\boldsymbol{\lambda}_{z_{ij},s_{ij}} \sim Mul(\boldsymbol{\lambda}_{z_{ij},s_{ij}})$.
         B. Sample word, $m_{ijp}|\boldsymbol{\mu}_{z_{ij},l_{ijp}} \sim Mul(\boldsymbol{\mu}_{z_{ij},l_{ijp}})$.

We use collapsed Gibbs Sampling [7], a well-established Markov chain Monte Carlo (MCMC) technique for approximate inference. In collapsed Gibbs Sampling, the multinomial distribution $\{\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\mu}\}$ are first marginalized (collapsed), a Markov chain over the latent indicator {c, z, s, l} is then constructed, for which stationary distribution is the posterior. We obtain samples of latent variables from the Markov chain. Point estimates for the collapsed distributions can then be computed given the samples and predictive distributions are computed by averaging over multiple samples.

Sampling Procedure. Gibbs Sampler repeatedly samples each latent variable conditioned on the current states of other hidden variables and observations. A configuration of latent states of the system is then obtained. Next we provide the derivation of the sampling formulas.

By marginalizing out $\{\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\mu}\}$ in Eq.(1), we obtain:

$$\begin{aligned}
&P(\mathrm{c}, \mathrm{z}, \mathrm{s}, \mathrm{l}|E, D, D', \rho, \alpha, \beta, \gamma, \zeta, \varepsilon, \omega) \\
&\quad \propto \int P(\boldsymbol{\pi}|\alpha) P(\mathrm{c}|\boldsymbol{\pi})\, d\boldsymbol{\pi} \int P(\boldsymbol{\theta}|\rho) P(E|\mathrm{c}, \boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&\quad \cdot \int P(\boldsymbol{\eta}|\beta) P(\mathrm{z}|\boldsymbol{\eta}, \mathrm{c})\, d\boldsymbol{\eta} \int P(\boldsymbol{\sigma}|\zeta) P(\mathrm{s}|\boldsymbol{\sigma}, \mathrm{z})\, d\boldsymbol{\sigma} \\
&\quad \cdot \int P(\boldsymbol{\varphi}|\gamma) P(w_D|\mathrm{z}, \mathrm{s}, \boldsymbol{\varphi})\, d\boldsymbol{\varphi} \int P(\boldsymbol{\lambda}|\varepsilon) P(\mathrm{z}, \mathrm{s}, \boldsymbol{\lambda})\, d\boldsymbol{\lambda} \\
&\quad \cdot \int P(\boldsymbol{\mu}|\omega) P(m_{D'}|\mathrm{z}, \mathrm{l}, \boldsymbol{\mu})\, d\boldsymbol{\mu} \\
&\quad = P(\mathrm{c}|\alpha)\, P(E|\mathrm{c}, \rho)\, P(\mathrm{z}|\beta, \mathrm{c})\, P(\mathrm{s}|\zeta, \mathrm{z})\, P(w_D|\mathrm{z}, \mathrm{s}, \gamma) \\
&\quad \cdot P(\mathrm{l}|\mathrm{z}, \mathrm{s}, \varepsilon)\, P(m_{D'}|\mathrm{z}, \mathrm{l}, \omega).
\end{aligned} \tag{2}$$

The conditional of c_i can be computed by dividing the joint distribution of all variables by the joint of all variables except c_i (denoted as c_{-i}):

$$\begin{aligned}
&P(c_i = c|\mathrm{c}_{-i}, \mathrm{z}, \mathrm{s}, \mathrm{l}, \cdots) \\
&= \frac{P(\mathrm{c}, \mathrm{z}, \mathrm{s}, \mathrm{l}|\ )}{P(\mathrm{c}_{-i}, \mathrm{z}, \mathrm{s}, \mathrm{l}|\ )} \\
&= \frac{P(\mathrm{c}|\alpha)}{P(\mathrm{c}_{-i}|\alpha)} \frac{P(E|\mathrm{c}, \rho)}{P(E|\mathrm{c}_{-i}, \rho)} \frac{P(\mathrm{z}|\beta, \mathrm{c})}{P(\mathrm{z}|\beta, \mathrm{c}_{-i})}.
\end{aligned} \tag{3}$$

We now derive the first fraction of Eq. (3), i.e.,

$$\frac{P(\mathrm{c}|\alpha)}{P(\mathrm{c}_{-i}|\alpha)} = \frac{\int P(\boldsymbol{\pi}|\alpha) P(\mathrm{c}|\boldsymbol{\pi})\, d\boldsymbol{\pi}}{\int P(\boldsymbol{\pi}|\alpha) P(\mathrm{c}_{-i}|\boldsymbol{\pi})\, d\boldsymbol{\pi}}. \tag{4}$$

As we assume each c is generated from a multinomial distribution $\boldsymbol{\pi}$, and the hyper-parameter for conjugate Dirichlet prior is α, we have:

$$\begin{aligned}
&\int P(\boldsymbol{\pi}|\alpha) P(\mathrm{c}|\boldsymbol{\pi})\, d\boldsymbol{\pi} \\
&= \int \frac{\Gamma(|C|\alpha)}{\prod_c \Gamma(\alpha)} \prod_c \pi_c^{\alpha-1} \prod_c \pi_c^{n^{(c)}}\, d\boldsymbol{\pi} \\
&= \frac{\Gamma(|C|\alpha)}{\prod_c \Gamma(\alpha)} \frac{\prod_c \Gamma\left(\alpha + n^{(c)}\right)}{\Gamma\left(|C|\alpha + n^{(\cdot)}\right)}.
\end{aligned} \tag{5}$$

Combining the above equation with Eq.(4) leads to :

$$\frac{P(\mathrm{c}|\alpha)}{P(\mathrm{c}_{-i}|\alpha)} = \frac{n_{-i}^{(c)} + \alpha}{n_{-i}^{(\cdot)} + C\alpha}, \tag{6}$$

where the subscript $-i$ denotes a vector (e.g. c_{-i}) or a count (e.g. $n_{-i}^{(c)}$) excluding current assignment (the same hereafter), $n^{(c)}$denotes the number of nodes assigned to community c. Marginal counts are represented with dots (the same hereafter), e.g., $n^{(\cdot)}$ denotes the total number of nodes. Here we use the identity $\Gamma(x+1) = x\Gamma(x)$. The second and third fractions of Eq. (3) can be derived analogously. The Dirichlet-Multinomial conjugates ensure the tractability of the integrals. Specifically, the second fraction can be written as:

$$\frac{P(E|c, \rho)}{P(E|c_{-i}, \rho)} = \frac{\prod_{i'=1}^{|U|} \prod_{q=0}^{n_i^{(i')}-1} \left(\rho + n_{c,-i}^{(i')} + q\right)}{\prod_{q=0}^{n_i^{(\cdot)}-1} \left(|U|\rho + n_{c,-i}^{(\cdot)} + q\right)}, \tag{7}$$

where $n_i^{(i')}$ denotes the number of links between nodes i and i' with = 1 if an edge exists, or 0 otherwise; $n_c^{(i')}$ denotes the number of times node i' generated by community c.

While the third fraction as:

$$\frac{P(z|\beta, c)}{P(z|\beta, c_{-i})} = \frac{\prod_{z=1}^{|Z|} \prod_{q=0}^{n_i^{'(z)}-1} \left(\beta + n_{c,-i}^{'(z)} + q\right)}{\prod_{q=0}^{n_i^{'(\cdot)}-1} \left(|Z|\beta + n_{c,-i}^{'(\cdot)} + q\right)} \tag{8}$$

where $n_i^{'(z)}$denotes the number of posts of user$i$ assigned to topic z; $n_c^{'(z)}$ denotes the number of posts assigned to community c and generated by topic z.

Finally, by combining Eqs. (6-8) we obtain the sampling formulas for community indicator c_i for user i:

$$\begin{aligned}
P(\mathrm{c}_i = c|\mathrm{c}_{-i}, \mathrm{z}, \mathrm{s}, \mathrm{l}, \cdots) \propto \frac{n_{-i}^{(c)}+\alpha}{n_{-i}^{(\cdot)}+|C|\alpha} \cdot \frac{\prod_{i'=1}^{|U|} \prod_{q=0}^{n_i^{(i')}-1} \left(\rho+n_{c,-i}^{(i')}+q\right)}{\prod_{q=0}^{n_i^{(\cdot)}-1} \left(|U|\rho+n_{c,-i}^{(\cdot)}+q\right)} \\
\cdot \frac{\prod_{z=1}^{|Z|} \prod_{q=0}^{n_i^{'(z)}-1} \left(\beta+n_{c,-i}^{'(z)}+q\right)}{\prod_{q=0}^{n_i^{'(\cdot)}-1} \left(|Z|\beta+n_{c,-i}^{'(\cdot)}+q\right)}.
\end{aligned} \tag{9}$$

The hidden variables z, s and l can be derived in a similar manner.

Sampling topic indicator z_{ij} according to

$$P\left(z_{ij}=z|c_i=c,s_{ij}=s,\mathrm{c}_{-i},\mathrm{z}_{-ij},\mathrm{s}_{-ij},\mathrm{l},\cdots\right)$$
$$\propto \frac{n'^{(z)}_{c,-ij}+\beta}{n'^{(\cdot)}_{c,-ij}+|Z|\beta}\cdot\frac{n^{(s)}_{z,-ij}+\zeta}{n^{(\cdot)}_{z,-ij}+|S|\zeta}\cdot\frac{\prod_{v=1}^{|V|}\prod_{q=0}^{n^{(v)}_{ij}-1}\left(\gamma+n^{(v)}_{zs,-ij}+q\right)}{\prod_{q=0}^{n^{(\cdot)}_{ij}-1}\left(|V|\gamma+n^{(\cdot)}_{zs,-ij}+q\right)}$$
$$\cdot\frac{\prod_{l=1}^{|L|}\prod_{q=0}^{n'^{(l)}_{ij}-1}\left(\varepsilon+n'^{(l)}_{ks,-ij}+q\right)}{\prod_{q=0}^{n'^{(\cdot)}_{ij}-1}\left(|L|\varepsilon+n'^{(\cdot)}_{zs,-ij}+q\right)}\cdot\frac{\prod_{l=1}^{|L|}\prod_{v=1}^{|V|}\prod_{q=0}^{n^{(v)}_{ij,l}-1}\left(\omega+n^{(v)}_{kl,-ij}+q\right)}{\prod_{l=1}^{|L|}\prod_{q=0}^{n^{(\cdot)}_{ij,l}-1}\left(|V|\omega+n^{(\cdot)}_{zl,-ij}+q\right)},\tag{10}$$

where $n_z^{(s)}$ denotes the number of posts assigned to topic z and generated by sentiment s; $n_{ij}^{(v)}$ denotes the number of times word v occurs in the post d_{ij}; $n_{zs}^{(v)}$ denotes the number of times word v generated by topic z and sentiment s; $n_{ij}^{'(l)}$ denotes the number of times words in comments d'_{ij} assigned to social sentiment l; $n_{zs}^{'(l)}$ denotes the number of words in comments assigned to social sentiment l (which were generated by topic z and sentiment s); $n_{ij,l}^{(v)}$ denotes the number of times word v in the comments d'_{ij} assigned to social sentiment l; $n_{zl}^{(v)}$ denotes the number of times word v generated by topic z and social sentiment l.

Sampling sentiment indicator s_{ij} according to:

$$P\left(s_{ij}=s|,z_{ij}=z,\mathrm{z}_{-ij},\mathrm{s}_{-ij},\mathrm{c},\mathrm{l},\cdots\right)$$
$$\propto\frac{n^{(s)}_{z,-ij}+\zeta}{n^{(\cdot)}_{z,-ij}+|S|\zeta}\cdot\frac{\prod_{v=1}^{|V|}\prod_{q=0}^{n^{(v)}_{ij}-1}\left(\gamma+n^{(v)}_{zs,-ij}+q\right)}{\prod_{q=0}^{n^{(\cdot)}_{ij}-1}\left(|V|\gamma+n^{(\cdot)}_{zs,-ij}+q\right)}\tag{11}$$
$$\cdot\frac{\prod_{l=1}^{|L|}\prod_{q=0}^{n'^{(l)}_{ij}-1}\left(\varepsilon+n'^{(l)}_{zs,-ij}+q\right)}{\prod_{q=0}^{n'^{(\cdot)}_{ij}-1}\left(|L|\varepsilon+n'^{(\cdot)}_{zs,-ij}+q\right)}.$$

Sampling social sentiment indicator l_{ijp} according to:

$$P\left(l_{ij,p}=l|,z_{ij}=z,s_{ij}=s,\mathrm{z}_{-ij},\mathrm{s}_{-ij},\mathrm{c},\mathrm{l}_{-(ij,p)},\cdots\right)$$
$$\propto\frac{n'^{(l)}_{zs,-(ij,p)}+\varepsilon}{n'^{(\cdot)}_{zs,-(ij,p)}+|L|\varepsilon}\cdot\frac{n^{(v)}_{zl,-(ij,p)}+\omega}{n^{(\cdot)}_{zl,-(ij,p)}+|V|\omega}.\tag{12}$$

At each iteration of our Gibbs sampler, for each user i, we sample the corresponding community indicator c_i; for each post d_{ij} by user i, we sample both the corresponding topic indicator z_{ij} and sentiment indicator s_{ij}; for each word p in the d'_{ij}, we sample the corresponding social sentiment indicator l_{ijp}. After a sufficient number of sampling iterations described above, we obtain a set of samples. The unknown distributions can then be computed by integrating across the samples [7] as follows:

$$\pi_c=\frac{n^{(c)}+\alpha}{n^{(\cdot)}+|C|\alpha},\theta_{cj}=\frac{n_c^{(j)}+\rho}{n_c^{(\cdot)}+|U|\rho},\eta_{cz}=\frac{n_c^{'(z)}+\beta}{n_c^{'(\cdot)}+|Z|\beta},\tag{13}$$

$$\sigma_{zs}=\frac{n_z^{(s)}+\zeta}{n_z^{(\cdot)}+|S|\zeta},\varphi_{zs,v}=\frac{n_{zs}^{(v)}+\gamma}{n_{zs}^{(\cdot)}+|V|\gamma},\tag{14}$$

$$\lambda_{zs,l}=\frac{n_{zs}^{'(l)}+\varepsilon}{n_{zs}^{'(\cdot)}+|L|\varepsilon},\mu_{zl,v}=\frac{n_{zl}^{(v)}+\omega}{n_{zl}^{(\cdot)}+|V|\omega}.\tag{15}$$

Table 2: Statistics of the four data sets

	Reddit1	Reddit1	Reddit2	RedditAll
No. of users	11,273	8,361	29,011	46,595
No. of links	18,691	10,535	50,135	79,361
No. of posts	1,295	675	3,228	5,198
No. of topics	4	4	4	4
Date (Aug.)	01-10	11-20	21-31	01-31

3.2 Time Complexity

Let $|E|$ be the number of social links, $|Y|$ the number of words in the posts, $|D|$ the number of posts and $|Y'|$ the number of words in the comments. In each iteration, the total time complexity is $O(|C|(R|+|D|)+|Z|(|D|+|Y'|)+|S|(|Y|+|Y'|))$, since all the numbers (e.g. $n_{-i}^{(c)}$) can be cached and updated in constant time. Because $|E|$, $|D|$, $|Y|$ and $|Y'|$ are usually much larger than $|C|$, $|Z|$ and $|S|$, the time complexity is linear with respect to the data size (i.e. social links, posts and comments), which ensures that CESSM is scalable to large-scale data sets.

4 EXPERIMENTS

In this section, we conduct intensive experiments to evaluate the performance of CESSM, compared with state-of-the-art techniques. We present our experimental setting in Section 4.1 and analyze experimental results in Section 4.2.

4.1 Experimental Setting

4.1.1 Real Data Sets. Although there are some publicly available social emotion analysis experimental data sets, such as news portals (e.g. news.sina.com.cn), these data sets only contain text data of writers, misssing comments of readers and topology information, thus cannot be applied to social sentiment mining. We therefore use four Reddit datasets corresponding to four time slices in August 2012 [26]. Each Reddit dataset contains the threads of four sub-forums (i.e. Movies, Politics, Science and Olympic) in *www.reddit.com.* Table 2 shows the detailed statistics of the four data sets. Each user can submit a post and other users may comment on it by replying to the post. If user i writes a post on a sub-forum, and user i' comments on this post, there will be a link between user i and user i' and the post will be labeled by this sub-forum as topic category. We then gather all comments on a post into a single document.

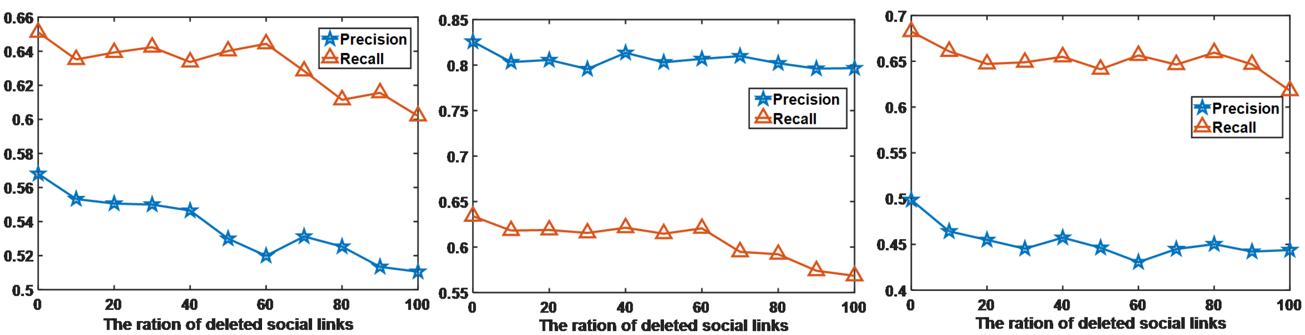
However, there is a lack of sentiment polarity of posts in these data sets, which we will need later in experiments. We employ a third party to label the sentiment polarity manually: we ask three people to label the sentiment polarity of all posts. For posts with inconsistent sentiment tagging, we determine their sentiment polarity according to the majority principle. We will make the labeled datasets publicly available upon publicaton

4.1.2 Performance Metrics. For sentiment classification, given a class (positive, neutral, or negative), we use the precision and recall [30] as performance metrics.

4.1.3 Parameter Settings. Since all methods compared (see below) need $|Z|$ and $|S|$ as input parameters, as well as our method need

Table 3: Comparisons in terms of precision and recall on the four data sets. R1, R2, R3 and RA represent the Reddit1, Reddit2, Reddit3 and RedditAll data sets respectively. All values in the table are percentages. Bold denotes the best result.

Class	Method	Precision				Recall			
		R1	R2	R3	RA	R1	R2	R3	RA
Negative	JST	16.10	18.48	37.00	32.18	45.60	49.37	47.25	60.39
	ASUM	16.67	14.67	33.83	38.18	48.90	**62.02**	67.86	51.98
	TS	34.23	24.44	42.42	45.44	46.23	55.22	57.43	55.43
	MSTM	33.34	22.55	55.44	55.85	**54.24**	58.23	60.75	62.85
	CESSM	**43.50**	**33.57**	**63.68**	**57.80**	53.29	60.76	**71.35**	**65.11**
Neutral	JST	64.24	66.76	59.07	68.54	40.19	44.08	37.76	37.53
	ASUM	70.63	75.41	57.46	72.32	43.15	42.92	**56.35**	42.39
	TS	68.23	78.35	65.63	75.63	48.23	40.32	51.33	52.45
	MSTM	72.36	77.66	66.54	78.64	55.45	52.62	54.54	56.64
	CESSM	**81.91**	**82.30**	**76.97**	**82.58**	**64.52**	**61.48**	55.86	**63.40**
Positive	JST	26.67	35.07	13.91	29.96	29.71	45.12	28.05	42.62
	ASUM	28.36	20.18	15.23	35.59	33.53	26.82	34.64	36.56
	TS	34.23	43.12	26.34	36.55	43.23	47.23	45.20	49.32
	MSTM	30.12	50.22	33.92	38.44	51.12	46.22	40.22	55.42
	CESSM	**52.52**	**52.63**	**39.57**	**49.86**	**65.59**	**67.07**	**50.82**	**68.24**



(a) (b) (c)

Figure 3: Precision and recall accuracies of (a) "negative", (b) "neutral" and (c) "positive" class for 11 different data sets where some portions of edges are randomly removed in the RedditAll data set. Each point in the figure is averaged over 20 problem inyiyustances.

$|C|$, $|Z|$, $|S|$ and $|L|$ as input parameters, for all experiments, we set the number of communities and topics to $|C| = |Z| = 4$ the same as the number of topic labels in the data sets (i.e. Movies, Politics, Science and Olympic), and the number of sentiments and social sentiments to $|S| = |L| = 3$ (i.e. neutral, negative and positive).

It is worth pointing out that Dirichlet hyper-parameters have low impact on model performance, and can be set as fixed values following the common strategy [11, 29] (i.e. $\alpha = 50/|C|$, $\beta = 50/|Z|$, $\varepsilon = 50/|L|$, $\zeta = 50/|S|$, $\omega = \rho = \gamma = 0.01$).

4.2 Experimental Results

4.2.1 Text Sentiment Classification. Since no existing methods can be applied to social networks without user votes, it is almost impossible to do quantitative analysis directly for mining social sentiment. As an acceptable alternative, we chose sentiment classification on posts, which can also validate the strong performance of our model for mining social sentiment, since it is believed that social sentiments of readers depend on sentiment expressed on posts of writers.

We tested four state-of-the-art methods based on topic model, which are more relevant as baselines to compare with our method, for text sentiment classification, including JST [17], ASUM [14] TS [6] and MSTM [1]. JST is an earlier topic-sentiment model that extends LDA with a new sentiment layer. Thus, the generation of a word depends not only on the topic but also on the sentiment polarity. ASUM is similar to JST but it works at sentence level (all words of a sentence are generated under the same topic). TS is also a variant of JST where the topic and sentiment layers are inverted and propose overall topic-specific distribution over sentiment polarities. MSTM is also a sentiment topic model suitable for short-length text like microblogs. Note that the traditional supervised classification methods for long documents are not suitable due to their poor performance on short posts with only a few sentences [31].

The code of the methods compared were obtained from their authors, and we used their default parameters. Besides, because all of these algorithms converge to local minima, we ran each algorithm 20 times and report the result with highest value of log-likelihood. The number of topics, which all methods require as an input parameter, was set to the ground-truth of the number of topics.

Our method almost always outperformed all of the methods compared in terms of the two performance metrics (Table 3). According to Table 3, we have the following four findings: (1) In general, ASUM outperforms JST a little because ASUM considers that all words in a sentence generated under the same topic, which is more suitable than in JST for short text in social networks. (2) TS outperforms JST and ASUM a little because TS models the sentiment distribution for each topic whereas JST and ASUM uncover the topic distribution for each sentiment. The sentiment distribution of topics is more reasonable and intuitive than the topic distribution of sentiments, because people usually choose a topic and then express various sentiments conditioned on it, rather than adopting the reverse order. (3) In general, MSTM outperforms the other three rival methods a little because MSTM assumes that each post in the social networks belongs to a single topic and sentiment, which is more reasonable and similar to our methods. (4) In contrast to the compared methods, our proposed method significantly increases the sentiment classification precision and recall for majority of the cases considered. Our explanation is that: CESSM considers not only the dependency of sentiments on topics but also the dependencies of topics on communities. These communities help finding the coherent topics, because communities play an important role in topics creation and maintenance. A user often shares more common topics with users in the same community than those in different communities. Most importantly, CESSM exploits emotional relevance between readers and writers, which indeed exists in real life. By considering the dependencies between sentiments of writers and social sentiments of readers, our method can get more accurate results than other methods. The experimental results validate the superiority of modeling the dependencies among the communities, topics, sentiments of writers and social sentiments of readers together over using this information in isolation.

4.2.2 Influence of Topology. In order to evaluate the impact of user's relationships on the accuracy of text sentiment classification, we randomly remove some portions (i.e. 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100%) of edges in the RedditAll data set to form 11 different data sets respectively. We run our model on these 11 data sets and show results in Figure 3. Here, we can see that as the number of edges decreases, the accuracy of sentiment analysis sometimes increases and sometimes decreases, but the overall trend is decrease. It may be because in social networks, links are denser within a community than across communities, and as the proportion of removed edges increases, communities in the social networks become fuzzier in most cases and clearer in a few cases. The clear community structure helps finding more coherent topics thus getting more accurate sentiments hidden in the text.

From the above analysis, we can conclude that the communities originating from the relationship between users have a great impact on identification of sentiments. Therefore, it proves that introducing the communities of a social network can help improving the detection accuracy of sentiments of writers, thereby affecting mining social sentiment.

Interestingly, it also seems that our model is outperforming almost all baselines also when the network information is completely omitted. This is because, instead of only exploring sentiments implicit on posts, our method mined both the sentiments of writers and readers that expressed on posts and comments, and we took the emotional contagion between writers and readers into account and linked the two through latent associations. We made the results of sentiments on posts and comments complement each other, and thus achieved a better performance even when the network information is completely omitted.

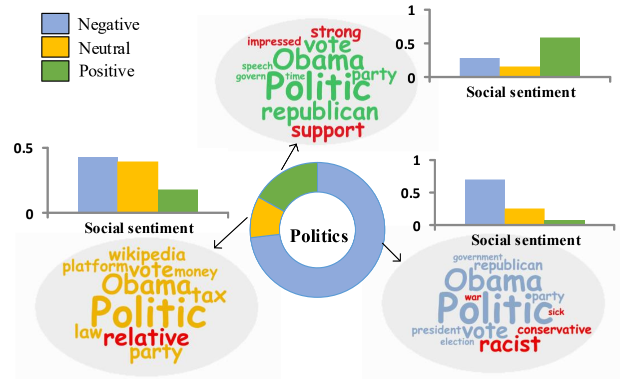


Figure 4: One example of the "Politics" topic. The ring represents the distribution of topic "Politics" over sentiments expressed in posts. The top ranked words of topic "Politics" and sentiment polarity pairs are shown in the three word clouds. More relevant a word, larger it is in the word clouds. The bar charts are the distributions of topic "Politics" and sentiment polarity pairs over social sentiments of readers.

4.2.3 Case Studies. The key application of the proposed method is to derive social sentiments. We carried out two corresponding case studies from two perspectives (i.e. topic-level and post-level) on the datasets of RedditAll to further evaluate the meaningfulness and interestingness of our method. There is no baseline algorithm to contrast with ours, since no other social sentiment mining method can be applied to microblogging networks like Reddit.

Topic-level: One run of our method found four topics, each of which had a distribution over sentiment polarity of posts and social sentiment of comments(i.e. $\boldsymbol{\sigma}_z$ and $\boldsymbol{\lambda}_{zs}$). Figure 4 shows the example of "Politics" topic. For this topic, posts published by users are mainly negative, followed by positive ones. That may be because users on Reddit are more radical and tend to complain . We also show the "Politics" topics under three different sentiment polarities of writers (i.e. negative, neutral and positive), each of which is represented by the top ranked words. For example, under the negative sentiment label category, users complain about various government policies, e.g. war and racial discrimination, and under the positive sentiment label category, users may express support for the government and Obama.

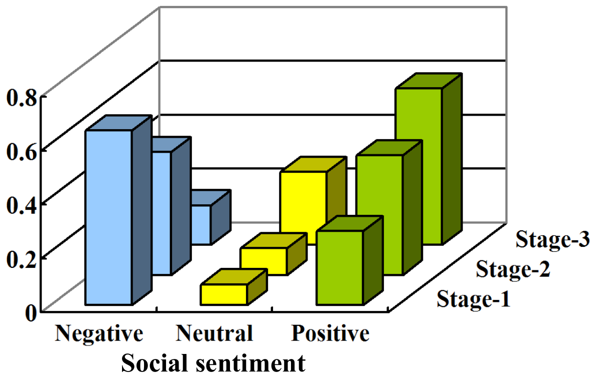


Figure 5: Social sentiments in Obama's presidential election at three different stages.

For each topic and sentiment polarity pair, our method extracted different sentimental responses from reader's perspective, i.e. social sentiment. As shown in Figure 4, the "Politics" topic under the negative sentiment label category mainly causes negative social sentiment because of readers' resonance. Interestingly, for the social sentiments under the other sentiment polarities, negative polarity also accounts for a large part. It is probably because most users on the network do not support the government and would refute in comments against positive or neutral posts.

In general, the topic-level perspective gives a global view of social sentiment distribution of different topics under each sentiment polarity of writers' posts.

Post-level: After running our method, topic label and sentiment polarity of each post can be obtained by latent variable z and s. Similarly, the social sentiment label of every word in each comment, by which we can export post-level social sentiment caused by the corresponding post, can also be obtained by latent variable l.

Through analysis of the posts in the dataset, we found an interesting real-world event about Obama's re-election. In August 2012, Obama first logged on the Reddit website to answer questions from Internet users online. The learned results successfully reveal the social sentiments at different stages of the above event. Figure 5 shows the learned social sentiments of three posts released at different stages. Stage-1 is the first post about news that Obama would come to Reddit and is dominated by negative social sentiment like "doubt" and "not optimistic." Stage-2 is a post saying that Obama is answering questions on Reddit online, and our model reveals that the proportion of positive social sentiment has increased like "surprised" and "support", while negative social sentiment still holds the biggest proportion as many users suspect that the one who is answering questions online is not Obama. Stage-3, which is assigned with positive social sentiment mostly, is the post which says that "Obama won the Internet".

Generally, the post-level perspective helps users find out how each post emotionally influences their readers. And it can be easily extended to the task that foresees how a new post will arouse readers' social sentiments via Bayes' theorem using these learned parameters of CESSM.

5 RELATED WORK

In this section, we firstly review the related work on sentimental topic mining, and then summarize the existing works on social emotion mining. Sentimental topic mining mainly deal with reviews and messages using supervised or unsupervised approaches, which paved the way to the newer research area of social emotion mining.

5.1 Sentimental Topic

Many researchers devote their time and effort to uncover topics and sentiments by separating opinion words from aspect words in texts for sentiment classification [31]. Most studies extend the PLSI [10] or LDA [4] model by placing a latent variable for both topics and sentiments. (1) Some methods assume that the latent sentiments are conditioned on topics, i.e., each topic has a multinomial distribution over sentiments. For instance, TS [6] can get different descriptions of the same topic corresponding to the different sentiment polarities, and overall topic-specific distribution over sentiment polarities. SJASM [8] develop supervised joint aspect and sentiment model to analyze overall and aspect-level sentiments for online user-generated review data, which often come with labeled overall rating information. (2) In contrast, other methods assume that the latent topics are conditioned on sentiments, i.e., each sentiment has a multinomial distribution over topics. Specifically, the work [14] automatically discovers what aspects are evaluated in reviews and how opinions are expressed for different aspects based on the sentence-LDA model [14] (that assumes all words in a sentence generated from a single aspect). JST [17] detect coherent and informative topics and sentiments simultaneously from text for document level sentiment classification by utilizing a domain independent sentiment lexicon. MSTM [1] discover sentiments associated with the topics in short-length microblogs and was based on the assumption that because of the short-length nature of microblogs, all the tokens in these social media posts belong to a single topic.

We used some ideas from the above sentimental topic models to draw the sentiments of writers. For example, we also assume that each post belongs to a single topic and sentiment because of the short-length nature of microblogs. However, our work mainly focuses on the social sentiment mining.

5.2 Social Emotion Mining

The existing approaches to social emotion mining can be classified into two categories: word-level [16] and topic-level [2] models. Earlier works of social emotion mining focused on exploiting the sentiment of individual words [16], by assuming that words are the foundation of annotating user sentiments on public events. For example, a number of relevant words are extracted from the news articles and the word-level features used to assign each article to an appropriate emotion category. In particular, the methods of constructing the word-level emotion lexicon treat each word individually. Thus, many relevant words associated with emotions are usually mixed with background noisy words which do not convey much affective meaning. More recently, topic-level models are proposed to exploit the sentiment of topics. The Emotion-Topic Model (ETM) [2] borrows the machinery of latent topic models like the Latent Dirichlet Allocation (LDA) [4], thereby facilitating to distinguish different meanings of the same word. Unfortunately, it is more suited to emotion annotations of text from the writer's perspective, rather than to online user emotional response after reading specific articles (i.e., the reader's perspective). Sentiment analysis from the reader's perspective has potential applications that differ from those of writer-sentiment analysis [18]. For example, the former can assist authors in foreseeing how their work will influence the readers emotionally, and help users retrieve documents that contain both relevant contents and desired emotions if we integrate emotion scores and rankings into information retrieval. Multi-label Supervised Topic Model (MSTM) [25] connect latent topics with evoked emotions of readers.

However, the above social emotion mining methods from the reader's perspective can only be applied to social medias like news portals with voting information. In addition, since it is too hard to detect social emotions from only short and noisy textual comments, we transform the task of "social emotion" mining into "social sentiment" mining. Meanwhile, existing methods ignore the emotional contagion between writers and readers, which is the indispensable characteristics considered by our approach.

6 CONCLUSION

In this paper, we have proposed a generative probabilistic model CESSM to mine social sentiment in microblogging networks by exploiting the important dependencies between communities topics, sentiments and social sentiments, and devised an approximate learning algorithm based on collapsed Gibbs sampling to estimate the model parameters. The new approach considers the emotional contagion between writers and readers, which is consistent with reality, and utilizes high-order network information, i.e. community structure, to solve the problem of short and noisy text in social networks. The experiment results on three tasks (i.e. sentiment classi-fication, influence of topology and case studies) show that proposed approach outperforms existing state-of-art methods.

As a part of our future work, we plan to extend our CESSM to dynamic social networks, and apply CESSM to new applications.

7 ACKNOWLEDGMENT

The work was supported by National Natural Science Foundation of China (No. 61772361, 61876128).

REFERENCES

[1] Aman Ahuja, Wei Wei, and Kathleen M. Carley. 2017. Microblog Sentiment Topic Model. In *IEEE International Conference on Data Mining Workshops*.

[2] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2011. Mining social emotions from affective text. *IEEE transactions on knowledge and data engineering* 24, 9 (2011), 1658–1670.

[3] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2012. Mining Social Emotions from Affective Text. *IEEE Transactions on Knowledge and Data Engineering* 24, 9 (2012), 1658–1670.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2012. Latent dirichlet allocation. *Journal of Machine Learning Research* 3 (2012), 993–1022.

[5] Erik Cambria, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. 2015. Knowledge-Based Approaches to Concept-Level Sentiment Analysis. *IEEE Intelligent Systems* 28, 2 (2015), 12–14.

[6] Mohamed Dermouche, Leila Kouas, Julien Velcin, and Sabine Loudcher. 2015. A joint model for topic-sentiment modeling from text. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, Salamanca, Spain, 819–824.

[7] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl 1 (2004), 5228–5235.

[8] Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao. 2017. Analyzing sentiments in one go: a supervised joint topic modeling approach. *IEEE Transactions on Knowledge and Data Engineering* 29, 6 (2017), 1172–1185.

[9] Elaine Hatfield, John T. Cacioppo, and Richard L. Rapson. 1993. Emotional Contagion. *Current Directions in Psychological Science* 2, 3 (1993), 96–100.

[10] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Fifteenth Conference on Uncertainty in Artificial Intelligence*.

[11] Zhiting Hu, Junjie Yao, Bin Cui, and Eric P Xing. 2015. Community Level Diffusion Extraction.. In *SIGMOD Conference*. 1555–1569.

[12] Xin Huang and Laks VS Lakshmanan. 2017. Attribute-driven community search. *Proceedings of the VLDB Endowment* 10, 9 (2017), 949–960.

[13] Di Jin, Xiaobao Wang, Ruifang He, Dongxiao He, Jianwu Dang, and Weixiong Zhang. 2018. Robust detection of link communities in large social networks by exploiting link semantics. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[14] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 815–824.

[15] Indika Kahanda and Jennifer Neville. 2009. Using transactional information to predict link strength in online social networks. In *Third International AAAI Conference on Weblogs and Social Media*.

[16] Phil Katz, Matthew Singleton, and Richard Wicentowski. 2007. Swat-mp: the semeval-2007 systems for task 5 and task 14. In *Proceedings of the 4th international workshop on semantic evaluations*. Association for Computational Linguistics, 308–313.

[17] Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. 2011. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering* 24, 6 (2011), 1134–1145.

[18] Kevin Hsin-Yih Lin and Hsin-Hsi Chen. 2008. Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 136–144.

[19] Sharad Nandanwar and M Narasimha Murty. 2016. Structural neighborhood based classification of nodes in a network. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1085–1094.

[20] Mark EJ Newman and Elizabeth A Leicht. 2007. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences* 104, 23 (2007), 9564–9569.

[21] Nie Peng, Zhao Xue, Yu Li, Wang Chao, and Zhang Ying. 2016. Social Emotion Analysis System for Online News. In *Web Information System and Application Conference*.

[22] Guo-Jun Qi, Charu Aggarwal, Qi Tian, Heng Ji, and Thomas Huang. 2011. Exploring context and content links in social media: A latent space method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 5 (2011), 850–862.

[23] Yanghui Rao. 2016. Contextual Sentiment Topic Model for Adaptive Social Emotion Classification. *IEEE Intelligent Systems* 31, 1 (2016), 41–47.

[24] Yanghui Rao, Qing Li, Wenyin Liu, Qingyuan Wu, and Xiaojun Quan. 2014. Affective topic model for social emotion detection. *Neural Networks* 58, 5 (2014), 29–37.

[25] Yanghui Rao, Qing Li, Xudong Mao, and Liu Wenyin. 2014. Sentiment topic models for social emotion mining. *Information Sciences* 266 (2014), 90–100.

[26] Chang-Dong Wang, Jian-Huang Lai, and S Yu Philip. 2013. NEIWalk: community discovery in dynamic content-based networks. *IEEE transactions on knowledge and data engineering* 26, 7 (2013), 1734–1748.

[27] Wenhui Wu, Sam Kwong, Yu Zhou, Yuheng Jia, and Wei Gao. 2018. Nonnegative matrix factorization with mixed hypergraph regularization for community detection. *Information Sciences* 435 (2018), 263–281.

[28] Hu Xia, Tang Lei, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Acm International Conference on Web Search and Data Mining*.

[29] Hongzhi Yin, Yizhou Sun, Bin Cui, Zhiting Hu, and Ling Chen. 2013. LCARS: a location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 221–229.

[30] Jia-Dong Zhang and Chi-Yin Chow. 2016. Crats: An lda-based model for jointly mining latent communities, regions, activities, topics, and sentiments from geosocial network data. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 2895–2909.

[31] Jia-Dong Zhang, Chi-Yin Chow, and Yu Zheng. 2015. ORec: An opinion-based point-of-interest recommendation framework. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1641–1650.

[32] X. Zou, J. Yang, and J. Zhang. 2018. Microblog sentiment analysis using social and topic context. *Plos One* 13, 2 (2018), e0191163.