Journal of Web Engineering, Vol.17, No. 1&2 (2018) 073-092


# SEMANTIC EMOTION-TOPIC MODEL IN SOCIAL MEDIA ENVIRONMENT

RUIRONG XUE

*School of Computer Engineering and Science, Shanghai University, China*
*xueruirong@i.shu.edu.cn*

SUBIN HUANG

*School of Computer Engineering and Science, Shanghai University, China*
*Anhui Polytechnic University, China*
*huangsubin@shu.edu.cn*

XIANGFENG LUO*

*Shanghai Institute for Advanced Communication and Data Science, School of Computer Engineering and Science, Shanghai University, China*
**Corresponding author: luoxf@shu.edu.cn*

DANDAN JIANG

*School of Computer Engineering and Science, Shanghai University, China*
*emily921220@163.com*

*YIKE GUO*

*School of Computer Engineering and Science, Shanghai University, China*
*Department of Computing, Imperial College London, British*
*y.guo@imperial.ac.uk*

*YAN PENG*

*School of Mechatronic Engineering and Automation, Shanghai University, China*
*pengyan@shu.edu.cn*

Received July 28, 2017
Revised October 15, 2017

With the booming of social media users, more and more short texts with emotion labels appear in social media environment, which contain users' rich emotions and opinions about social events or enterprise products. Social emotion mining on social media corpus can help government or enterprise make their decisions. Emotion mining models involve statistical-based and graph-based approaches. Among them, the former approaches are more popular, e.g. Latent Dirichlet Allocation (LDA)-based Emotion Topic Model. However, they are suffering from bad retrieval performance, such as the bad accuracy and the poor interpretability, due to them only considering the bag-of-words or the emotion labels in social media media

environment. In this paper, we propose a LDA-based Semantic Emotion-Topic Model (SETM) combining emotion labels and inter-word relations to enhance the retrieval performance in social media media environment. The performance influence of four factors on SETM are considered, i.e., association relations, computing time, topic number and semantic interpretability. Experimental results show that the accuracy of our proposed model is 0.750, compared with 0.606, 0.663 and 0.680 of Emotion Topic Model (ETM), Multi-label Supervised Topic Model (MSTM) and Sentiment Latent Topic Model (SLTM) respectively. Besides, the computing time of our model is reduced by 87.81% through limiting word frequency, and its accuracy is 0.703, compared with 0.501, 0.648 and 0.642 of the above baseline methods. Thus, the proposed model has broad prospects in social media media environment.

*Key words*: Social emotion mining; Semantic discovery; Social emotion classification; Topic Model; Semantic Emotion Topic Model

*Communicated by*: M. Gaedke & Q. Li

## 1 Introduction

In recent years, with the rapid change of social media environment (e.g. Sina Microblog and Twitter), more and more users tend to share their opinions, experiences or emotions in the above social media environment. Users are increasingly interested in using more emotion labels as well as short texts to express their emotions and opinions. Thus, the mixtures of emotion labels and short texts carry users' rich emotions and opinions. Figure 1 shows an example of short texts with emotion labels. Many news websites, e.g., Sina Society Channel[a] have provided a news service for users to express their emotions and opinions after browsing news [1, 9]. In such websites, each article is shown with ratings by users who have read the article and voted over a set of predefined emotion labels/emoticons, as Figure 2 shows.


Figure 1 Short texts with emotion labels

Social emotion mining has been widely used, including opinion summarization [**Error! Reference source not found.**] and sentiment retrieval [**Error! Reference source not found.**], and has attracted lots of attention from researchers of natural language processing and machine learning [5, 6]. By mining

[a] http://news.sina.com.cn/society/

social emotion, government can find out the emotions and opinions of people towards specified social events. Enterprises can assess customers' satisfaction to help promote their products by analyzing emotions of comments.


Figure 2 An example of emotion labels and user ratings

Existing mainstream approaches to social emotion mining are based on statistical model. However, most of them are suffering from low accuracy and poor interpretability, since they only consider words and emotion labels in short texts. Besides, individual words' emotion is ambiguity [**Error! Reference source not found.**], which may lead to a quite biased prediction of social affective texts. Thus, many researchers start to change Latent Dirichlet Allocation [**Error! Reference source not found.**] involving emotion labels or emoticons [9-11] to correct words emotions. It has enhanced the accuracy of social emotion computing to some extent. But, all these LDA–based models are bag-of-words models, which carry less semantics in social media corpus. As shown in Table 1, the emotion that is represented by the distribution of words has bad semantic interpretability. It is hard to reveal the knowledge association and help researchers find out semantic context. Figure 3 shows the association relations we may extract from social media corpus. For example, negative words like "corrupt" or "arrested" without context may lead to misunderstanding of documents, while semantic context like "corrupt—arrested" express real positive emotion of readers.

Table 1 The word distribution of emotion "surprise" is shown as bag-of-words with weak semantics.

| word | probability |
|---|---|
| rich | 0.0246 |
| children | 0.0212 |
| Samsung | 0.0211 |
| single | 0.0113 |
| witness | 0.0018 |
| hurt | 0.0015 |
| female anchors | 0.0013 |
| corruption | 0.0011 |

In this paper, we propose a LDA-based Semantic Emotion-Topic Model (SETM) involving relations and emotion labels in social media environment. The model follows a several-step generation process for affective terms, which first generates a word's latent topic from a document-specific topical distribution and a word's emotion, label's emotion together with a relation's emotion from a document-specific emotional distribution, then generates a word from Multinomial distribution based on latent topics and emotions, an emotion label and a relation from Multinomial distribution based on their respective emotions.

According to psychology[**Error! Reference source not found.**], we find it best to define six-dimension emotions(love、 fear、 joy、 anger、 sad、 surprise) to describe human emotions. We

evaluate the proposed model on an online collection collected from the Sina Society Channel. Since the website established a one-to-one relationship between each of the emotion labels and users' emotions shown in Figure 2, the emotion labels matrix and emotions matrix of the same document are fixed to be the same. Experimental results show that the performance of proposed model is affected by relations obviously. Mining results of the proposed model can also be interpreted more semantically by combining words and sentence environments.

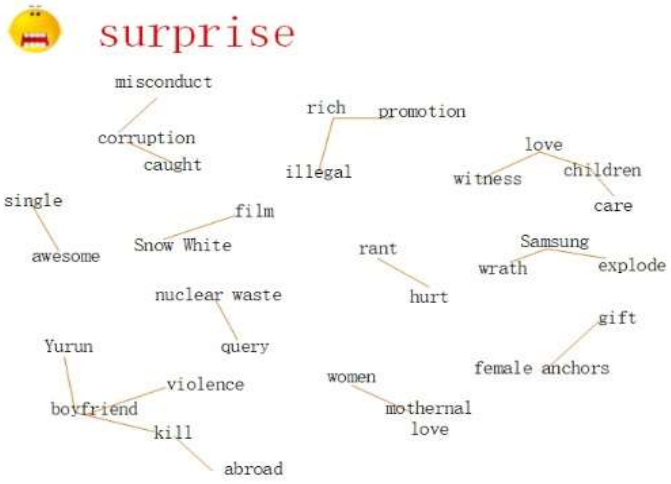


Figure 3 An example of "surprise" relations network extracted from social events, which has strong semantic context

The remainder of this paper is organized as follows. We describe related work in Section 2 and present our model for social emotion mining in Section 3. Data set, results, and discussion are illustrated in Section 4. Finally, conclusion is given in Section 5.

## 2 Related Work

In this section, we firstly review the related work on sentiment classification and analysis, and then introduce the related topic models used in the area of affective text mining.

Previous work on sentiment classification and analysis can be classified into three levels: document-level [13-17], sentence-level [18, 19] and word-level [20, 21].

Document-level sentiment computing can fall into two parts: supervised learning [13, 14] and unsupervised learning [15-17]. Document-level emotion computing is also a text classification problem, so all existed supervised learning methods can be used to solve it. Supervised learning features, like terms and their frequency, part of speech information, opinion words, negations, and syntactic dependence [13, 14], all have been applied in computing document sentiment. Since the supervised learning depends more on subjective factors, it costs more manpower and time to choose and evaluate the training corpus. Besides, only the categories defined in training samples can be recognized. So that the classification result may be influenced by some unknown categories. In unsupervised learning, Turney [**Error! Reference source not found.**] extracted phrasal by relations and proposed an algorithm to calculate one phrase polarity according to the Pointwise Mutual Information (PMI) and Information Retrieval (IR) algorithm and search results of searching engine. Taboada [**Error! Reference source not found.**] used the lexicon involving emotion words and phrases to compute each document's emotion score. Hu [**Error! Reference source not found.**] computed emotion with emotion signals involved in

social media. Li X [**Error! Reference source not found.**] presented a Bayesian-based model named WMCM to learn document-level semantic features. Li X [**Error! Reference source not found.**] leveraged unsupervised teaching models to incorporate semantic domain knowledge into the neural network to bootstrap its inference power and interpretability. But, reliable classification results can be available just after massive analysis, post-processing and labeled dataset.

Sentence-level sentiment computing can fall into two parts: First, subjectivity classification, which distinguishes objective sentences from subjective sentences. Second, sentiment classification on subjective sentences. If the sentence is judged as subjective, then we can identify its emotion orientation. All document-level method and lexicon-based method can be involved in sentence emotion computation. Besides, Yamamoto [**Error! Reference source not found.**] determined a tweet sentiment based on the emoticon role. Tang [**Error! Reference source not found.**] built an emotion classification framework on sentence level text. But, some objective sentences still contain opinion tendency. For example, production descended 0.3%, compared with last year. Just judging the sentence subjectivity and then classifying sentiment orientation may leak some objective sentences with opinions.

Word-level sentiment computing is the basis of sentiment computing in both sentence-level and document-level. Word-level computing methods can be applied in compiling sentiment lexicon. Kiritchenko [**Error! Reference source not found.**] used the method proposed by Mohammad [**Error! Reference source not found.**] that emotional words labeled by hashtag (#) in tweet implying the whole tweet express the same emotion, to construct a word-emotion association lexicon. But, word emotion computing relies on context so much that it is hard to judge words' emotion in different context.

In addition to the above methods, there are also numerous approaches for modeling, e.g., probabilistic Latent Semantic Indexing (PLSI) [**Error! Reference source not found.**] and LDA [**Error! Reference source not found.**]. Later researchers introduce other factors into these topic model. A model called emotion-topic model (ETM) [**Error! Reference source not found.**] followed the Naive Bayes method by assuming words are independently generated from social emotion labels. It introduces an intermediate emotion layer into LDA and assumes each topic being an important component of an emotion. Rao [**Error! Reference source not found.**] proposed another two topic models called Multi-label Supervised Topic Model (MSTM) and Sentiment Latent Topic Model(SLTM) to associate latent topics with evoked emotions of readers.

Most previous works only distinguish the polarity orientation (positive/ negative) of documents. Recent years also spring up many researches focus on multidimensional emotions. Yamamoto [**Error! Reference source not found.**] described ten sentiment dimensions. Plutchik [23, 24] clustering eight basic emotions into four-dimensional sentiment vectors: "Acceptance – Disgust", "Anticipation – Surprise", "Joy – Sadness" and "Anger – Fear". Takaoka [**Error! Reference source not found.**] proposed a method for extracting six-dimensional sentiment. Kumamoto [**Error! Reference source not found.**] used another six sentiment dimensions to represent readers' emotions. However, our method use "love", "fear", "joy", "sad", "surprise" and "anger" as six emotions which considering the nature of human emotions[**Error! Reference source not found.**].

Apart from the traditional features – words, many literatures [27-29] have included emoticons or emotion labels into sentiment computing process. Based on words and emotion labels, we find it more related with the language nature when we consider inter-word relations into the generation process since relations show more semantic relations between words. The details of extracting relations will be presented in section 3.

## 3 Proposed Model

In this section, we present an emotion topic model with more semantics in social media environment. We name the model as Semantic Emotion-Topic Model (SETM).In the following part, we will present the Semantic Emotion-Topic Model in detail.

### *3.1 Semantic Emotion-Topic Model*

In this subsection, we will briefly introduce the Semantic Emotion-Topic Model. Figure 4 presents the graphical model of the proposed SETM in social media environment. SETM generates each word conditioning on topics and emotions simultaneously. But for an emotion label in a document, it is influenced by just the writers' emotion. As relations are composed by words and the process of generating words has already involved topics, so we don't consider the topics in generating relations to simplify our model.

As a complete generative model, SETM allows us to associate each emotion with word tokens and relation tokens jointly, and to predict the probabilities of emotions conditioned to unlabeled documents that contain word tokens and relation tokens (without emotion labels). Here, we define a social texts collection consists of D documents $\{d_1, d_2, \dots, d_D\}$ with word tokens, relation tokens and user ratings. Word tokens are selected from a vocabulary containing V distinct terms. Relation tokens are selected from a relation list containing U distinct terms and a set of emotion user ratings are chosen from a predefined list of T emotion labels. The list of emotions is denoted by $e = \{e_1, e_2, \dots, e_E\}$. In this paper, we define the instances of emotions as "love", "fear", "joy", "sad", "surprise" and "anger". Similarly, a document $d$ consists of a sequence of N word tokens $\{w_{d,1}, w_{d,2}, \dots, w_{d,N}\}$, a sequence of M emotion ratings over T emotion labels denoted by $\{l_{d,1}, l_{d,2}, \dots, l_{d,M}\}$ and a sequence of Q relation tokens $\{r_{d,1}, r_{d,2}, \dots, r_{d,Q}\}$.In the $d^{\text{th}}$ document, $w_{d,n}$ represents the $n^{\text{th}}$ word, $l_{d,m} \in e$ represents the $t^{\text{th}}$ emotion label and $r_{d,q}$ represents the $q^{\text{th}}$ relation. It's worth noting that emotion labels are different from the emotions. It means that one emotion label or emoticon may belong to several emotions in the shape of distribution.

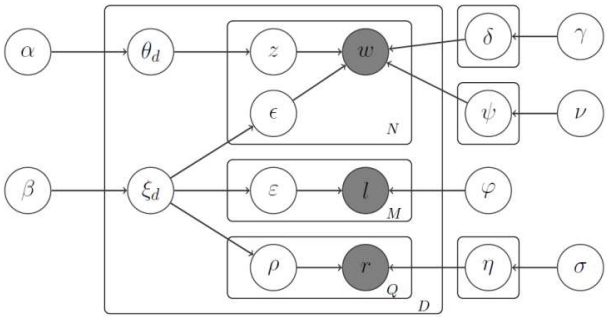


Figure 4 The graphical model of Semantic Emotion-Topic Model (SETM)

In our model, label token $l_{d,m}$ is generated form emotion-emotion distribution $\varphi_e$ and $\varphi$ is related with human true emotion matrix, thus we use $\xi_d$ to represent the multinomial distribution of emotions specific to document d. $w$ is generated from topic z and emotion $\epsilon$. Since words can reflect the latent topics in documents, we use $\theta_d$ to represent the multinomial distribution of topics specific to document d.

Table 2 Notations of variables in our model

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| K | Number of topics | E | Number of emotions |
| D | Number of documents | U | Number of unique relation tokens |
| V | Number of unique word tokens | T | Number of unique predefined emotion labels |
| α | Dirichlet prior of θ | Γ | Dirichlet prior of δ |
| B | Dirichlet prior of ξ | ν | Dirichlet prior of ψ |

| | | | |
|---|---|---|---|
| $\Sigma$ | Dirichlet prior of $\eta$ | N | Number of word token in each document |
| Q | Number of relation token in each document | M | Number of emotion labels in each document |
| $e_t$ | The $t^{th}$ emotion | $w_{d,n}$ | The $n^{th}$ word token in document $d$ |
| $z_{d,n}$ | The topic assigned to word token $w_{d,n}$ | $\epsilon_{d,n}$ | The emotion assigned to word token $w_{d,n}$ |
| $l_{d,m}$ | The $m^{th}$ emotion label in document $d$ | $\varepsilon_{d,m}$ | The emotion assigned to emotion label $l_{d,m}$ |
| $r_{d,q}$ | The $q^{th}$ relation token in document d | $\rho_{d,q}$ | The emotion assigned to relation token $r_{d,q}$ |
| $\theta_d$ | The multinomial distribution of topics specific to document d | | |
| $\delta_k$ | The multinomial distribution of words specific to topic k | | |
| $\xi_d$ | The multinomial distribution of emotions specific to document d | | |
| $\psi_e$ | The multinomial distribution of words specific to emotion e | | |
| $\varphi_e$ | The multinomial distribution of emotion labels specific to emotion e | | |
| $\eta_e$ | The multinomial distribution of relation token specific to emotion e | | |

In our model, label token $l_{d,m}$ is generated form emotion-emotion distribution $\varphi_e$ and $\varphi$ is related with human true emotion matrix, thus we use $\xi_d$ to represent the multinomial distribution of emotions specific to document d. $w$ is generated from topic z and emotion $\epsilon$. Since words can reflect the latent topics in documents, we use $\theta_d$ to represent the multinomial distribution of topics specific to document d.

Table 2 lists the notations of frequently used variables in this paper. In the graphical model as shown in Figure 4, shaded nodes are observed data, blank nodes are latent parameters, and arrows indicate dependence. The parameterization of the latent data in this model is shown as follows:

$$\theta_d|\alpha \sim Dir(\alpha)$$
$$\xi_d|\beta \sim Dir(\beta)$$
$$\delta_k|\gamma \sim Dir(\gamma)$$
$$\psi_e|\nu \sim Dir(\nu)$$
$$\eta_e|\sigma \sim Dir(\sigma)$$
$$z_{dn}|\theta_d \sim Mult(\theta_d)$$
$$\epsilon_{d,n}|\xi_d \sim Mult(\xi_d)$$
$$\varepsilon_{d,m}|\xi_d \sim Mult(\xi_d)$$
$$\rho_{d,q}|\xi_d \sim Mult(\xi_d)$$
$$w_{d,n}|\delta_{z,n}, \psi_{\epsilon,n} \sim Mult((\delta_{z,n} + \psi_{\epsilon,n})|2)$$
$$l_{d,m}|\varphi_{\varepsilon,m} \sim Mult(\varphi_{\varepsilon,m})$$
$$r_{d,q}|\eta_{\rho,q} \sim Mult(\eta_{\rho,q})$$

The generation process of SETM can be described as:

1. Choose $\delta_k \sim Dir(\gamma)$, $\psi_e|\nu \sim Dir(\nu)$, $\eta_e|\sigma \sim Dir(\sigma)$
2. For each document d, the word tokens, emotion labels and relation tokens are generated as follows:
   1) Choose $\theta_d \sim Dir(\alpha)$, $\xi_d \sim Dir(\beta)$
   2) For each of the $n^{th}$ word tokens $w_{d,n}$:
      (1) Choose a topic $z_{d,n} \sim Mult(\theta_d)$.
      (2) Choose a word emotion $\epsilon_{d,n} \sim Mult(\xi_d)$.
      (3) Choose a word token $w_{d,n} \sim Mult((\delta_{z,n} + \psi_{\epsilon,n})|2)$

3) For each of the $m^{\text{th}}$ emotion label $l_{d,\text{m}}$:
   (1) Choose a label emotion $\varepsilon_{d,m} \sim Mult(\xi_d)$.
   (2) Choose an emotion label $l_{d,\text{m}}$ from $p(l_{d,m}|\varepsilon_{d,m}, \varphi)$.
4) For each of the $q^{\text{th}}$ relation tokens $r_{d,\text{q}}$:
   (1) Choose a relation emotion $\rho_{d,q} \sim Mult(\xi_d)$.
   (2) Choose an emotion label $r_{d,\text{q}}$ from $p(r_{d,q}|\rho_{d,q}, \eta)$.

After generating D documents by the process above, the parameter $\psi, \eta$ are used to predict the documents without emotion labels.