Expert Systems With Applications 114 (2018) 322–333

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Universal affective model for Readers' emotion classification over short texts

Weiming Liang[a], Haoran Xie[b,*], Yanghui Rao[a], Raymond Y.K. Lau[c], Fu Lee Wang[d]

[a] School of Data and Computer Science, Sun-Yat Sen University, Guangzhou, China
[b] Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong
[c] Department of Information Systems, City University of Hong Kong, Hong Kong
[d] School of Science and Technology, The Open University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 2 May 2018
Revised 7 July 2018
Accepted 11 July 2018
Available online 26 July 2018

Keywords:
Topic model
Emotion classification
Biterm
Short text

ABSTRACT

As the rapid development of Web 2.0 communities, social media service providers offer users a convenient way to share and create their own contents such as online comments, blogs, microblogs/tweets, etc. Understanding the latent emotions of such short texts from social media via the computational model is an important issue as such a model will help us to identify the social events and make better decisions (e.g., investment in stocking market). However, it is always very challenge to detect emotions from above user-generated contents due to the sparsity problem (e.g., a tweet is a short message). In this article, we propose an universal affective model (UAM) to classify readers' emotions over unlabeled short texts. Different from conventional text classification model, the UAM structurally consists of topic-level and term-level sub-models, and detects social emotions from the perspective of readers in social media. Through the evaluation on real-world data sets, the experimental results validate the effectiveness of the proposed model in terms of the effectiveness and accuracy.

1. Introduction

With the rapid development of social media service providers, there is an increasing affective data such as the review comments and/or the vote counts of emotions on news articles, which reflect the emotion tendency and perspective from users (Bosco, Patti, & Bolioli, 2013). As one of the most important medium in the modern world, the Web can effectively not only convey users' positive or negative sentiments but also express more detailed emotions such as happiness, fearfulness, or surprise (Shaikh, Prendinger, & Ishizuka, 2008). The emotional voting service provided by many online news sites and social media communities enables users to express their emotions after reading news articles (Bao et al., 2009). Social emotion mining techniques have drawn greater attention of researchers in the machine learning and natural language processing (Cambria & White, 2014), for that they can be employed in various applications including sentiment retrieval (Ku, Liang, & Chen, 2006) and opinion summarization (Eguchi & Lavrenko, 2006). Early studies of social emotion classification primarily focused on identifying the emotional tendency of each individual word, because it was believed that the words in the natural language played a crucial role in expressing various emotions (Kazemzadeh, Lee, & Narayanan, 2013). The SWAT system (Katz, Singleton, & Wicentowski, 2007) adopted a word-emotion mapping dictionary to judge social emotions of unseen news headlines, which this dictionary was exploited by a supervised method. The emotion-term model (ETM) (Bao et al., 2009; 2011) modeled the associations between emotions and words from the authors perspective. These mentioned term-level models performed well in sentiment polarity classification of labeled short texts (Rao et al., 2016), because words in short texts were more easily and accurately mapped to sentiment polarity than diverse emotion labels.

However, those models, which assumed that a word is a unique key feature for sentiment analysis, are difficult to address the problem of sentiment ambiguity over multi-label texts (Quan & Ren, 2010). The sentiment ambiguity refers that the same word may express different emotions in various contexts. To address this problem, it was suggested to explore emotion distribution under specific topics by the emotion-topic model (ETM) (Bao et al., 2009). Those topics represented the objects, real-world events, or abstract entities which indicated the contexts of the sentiment (Stoyanov & Cardie, 2008). The ETM, which learned from the machinery of latent variable topic models like the Latent Dirichlet Allocation (LDA)

[*] Corresponding author.
E-mail addresses: terryliang020@foxmail.com (W. Liang), hrxie2@gmail.com, hxie@eduhk.hk (H. Xie), raoyangh@mail.sysu.edu.cn (Y. Rao), raylau@cityu.edu.hk (R.Y.K. Lau), pwang@ouhk.edu.hk (F.L. Wang).

https://doi.org/10.1016/j.eswa.2018.07.027
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

model (Blei, Ng, & Jordan, 2003), can distinguish different meanings of a single word. The ETM is developed for sentiment analysis from authors' perspective rather than readers'. The readers' perspective is more natural to understand the emotions of readers after they read the news article, while authors' perspective reflects the emotions of authors when they write the article (Lin & Chen, 2008). The affective topic model for social emotion detection (ATM) (Rao, Li, Wenyin, Wu, & Quan, 2014c) was proposed for detecting reader emotions towards certain topics by introducing an emotional intermediate layer. One limitation of the ATM is difficult to detect emotions from short texts, which are frequently occurred in social documents like tweets.

In light of these considerations, we propose the universal affective model (UAM) to detect social emotions over short texts from readers' perspective. The main contributions of this research are listed as follows.

- To enhance the semantic relationships between biterms, we combine biterms with keywords which are extracted by a new paradigm named Average Term Frequency Inverse Document Frequency (ATF-IDF).
- To differentiate various semantic meanings of the same word in short texts, our proposed the UAM based on the biterm topic model (BTM) (Cheng, Yan, Lan, & Guo, 2014) by adopting an intermediate layer which bridges emotion labels and topics.
- A word-level emotional lexicon is established for background words in the corpus by using SWAT.
- Through conducting experiments on 3 different data sets including a sensibly small and unbalanced news headlines with 6 emotions, a large social network short documents annotated over 2 emotions, and a larger online news articles annotated over 8 emotions, the effectiveness of the proposed model is verified.

The remaining sections of this paper are organized as follows. In Section 2, the related research studies are reviewed. In Section 3, the proposed UAM for readers' emotion classification are elaborated. Experiments are introduced in Section 4. The conclusion and future research directions are discussed in Section 5.

2. Related work

2.1. Sentiment analysis

The sentiment analysis aimed to identify and extract the attitude of a document (i.e., reaction of an online news reader to either a topic or the content of the document (Gangemi, Presutti, & Reforgiato Recupero, 2014)). In some earlier studies, the mission of sentiment analysis was to estimate whether the text is positive or negative by analyzing the entire text and the rating scores in reviews (Cambria, Schuller, Liu, & Wang, 2013a). A classification algorithm (Das & Chen, 2001) was exploited by Das and Chen to capture the latent opinions of markets from stock message boards. These latent opinions were further used as the steering of decision-making in the financial market. Turney (2002) attempted to classify the sentiment orientations of user reviews by using a unsupervised learning method. Pang, Lee, and Vaithyanathan (2002) classified movie reviews as positive or negative with an algorithm which is the combination of maximum entropy, naive Bayes, and support vector machine (SVM)(Adankon & Cheriet, 2009). However, some words may express different sentiments in specific domain applications (e.g., bull shows positive in the financial market) (Bollegala, Weir, & Carroll, 2011), these studies encountered a problem that a classifier trained by data in one domain may achieve poor performance in another one (Pan, Ni, Sun, Yang, & Chen, 2010). To solve this problem, several algorithms for domain-independent sentiment classification have been proposed (Bollegala et al., 2011). Another solution to capture different sentiments of the same word is conducted by introducing a topic layer. For instance, Rao proposed a contextual sentiment topic model for adaptive classification (Rao, 2016). Poddar et al. proposed a model to determine opinions by modeling aspects, topics, and sentiments jointly (Poddar, Hsu, & Lee, 2017).

Considering the remarkable performance in computer vision, deep neural network models have been recently employed for sentiment analysis over documents. In a preliminary study, Kim (2014) employed a convolutional neural network (Collobert, Weston, Karlen, Kavukcuoglu, & Kuksa, 2011) to generate both task-specific and static vectors for sentence classification. Due to the limited contextual information in short messages, Santos and Gattit proposed a deep neural network architecture which jointly uses representation at the character-level, word-level and sentence-level to perform sentiment analysis (Santos & Gattit, 2014). To exploit the information provided by sentiment lexicons, Shin et al. integrated lexicon embeddings and an attention mechanism into convolutional neural networks for sentiment analysis (Shin, Lee, & Choi, 2016).

The above algorithms and models have been applied in sentiment classification at levels of pages or paragraphs primarily. Some limitations have been identified if more fine-grained levels (e.g., sentences or clauses) have been used (Cambria, Schuller, Liu, Wang, & Havasi, 2013b). Sentiment analysis becomes more challenge as the opinion holders are anonymous and noisy data is often mixed with useful information (Moreo, Romero, Castro, & Zurita, 2012). For example, many fake comments, offensive comments or comments for advertisement are always appeared in different e-commerce sites and social media communities. Therefore, there has been another stream of research studies about opinion spam filtration (Jindal & Liu, 2008; Moreo et al., 2012) and noisy label aggregation in social media (Zhan et al., 2017). Some survey studies of sentiment analysis discuss the future research directions in this area (Cambria et al., 2013a; Cambria et al., 2013b).

2.2. Social emotion detection

Social emotion detection, which aimed to identify readers' emotions evoked by news headlines, has attracted increasingly attentions in the research communities since the emergence of the SemEval-2007 Tasks (Katz et al., 2007). This research was based on the hypothesis that all words including neutral ones can effectively express the positive or negative emotions of the author, and then evoke corresponding pleasant or painful reactions from readers (Shaikh et al., 2008).

The SWAT system (Katz et al., 2007) detected social emotion of unlabeled news headlines through a word-emotion lexicon. In this lexicon, each word is associated to multiple emotion, such as fear, anger, joy, surprise, etc., with each label has an emotion score respectively. However, the limited information in news headlines makes it difficult to detect emotions correctly and consistently (Katz et al., 2007; Quan & Ren, 2010). It was considered to make use of all words in the body of a news item in the emotion term (ET) model (Bao et al., 2009; 2011) and word-emotion (WE) method (Rao, Lei, Liu, Li, & Chen, 2014a). ET was designed to establish relationships between words and emotions based on the naive Bayes classifier. The WE method used maximum likelihood estimation to generated a word-level emotional lexicon, and then utilized this lexicon to detect emotion based on all terms in the news content. However, these word-level methods are unable to distinguish the different semantics of the same words under various contexts due to the issue of ambiguity (i.e., the same word may convey positive emotions in one context but negative emotions in another one) (Bollegala et al., 2011; He, Lin, & Alani, 2011; Quan & Ren, 2010). To deal with this, a semantically rich hybrid neural network was proposed for social emotion classification by building word-level embeddings (Li et al., 2017). since words are projected from a sparse and 1-of-V (here V is the vocabulary size) onto a low dimensional vector space via a hidden layer, these embeddings are essentially high-level representations that encode the semantics of words.

2.3. Labeled topic model

To discriminate various semantics of the same word under contexts, the labeled topic models have been extensively studied in social emotion detection. Rao et al. proposed the Emotion Latent Dirichlet Allocation (ELDA) model (Rao et al., 2014a) to predict social emotions at the topic levels. To deal with synonymous and polysemous words, ELDA model (Rao et al., 2014a) firstly used LDA to generate the original word-by-document matrix to a lower dimensionality and the conditional probability of the social emotion under each topic was then estimated by the maximum likelihood estimation. Although ELDA can capture latent topics involving social emotions, it was essentially a method based on the feature reduction. More specifically, ELDA generates topics by employing an unsupervised approach without instruction of emotion labels. Different with ELDA, ETM (Bao et al., 2011) was a joint emotion-topic model. It was an extension of labeled LDA (Ramage, Dumais, & Liebling, 2010; Ramage, Hall, Nallapati, & Manning, 2009) and the joint sentiment-topic model (JSTM) (Lin & He, 2009) in terms of generative processes. The labeled LDA constrains the model to use topics within label sets observed in documents. The generative process of JSTM has good performance on the sentiment analysis of a dataset of movie comments. In ETM, an intermediate layer was introduced into LDA, and topics were regarded as important components of an emotion. Further, the ETM captured and clustered coherent topics under different emotions. Empirical results had illustrated that the ETM model was much more effective than conventional machine learning methods like SVM for social emotions detection.

In recent years, the supervised topic model (STM) (Blei & Mcauliffe, 2010) , the topic-over-time (TOT) model (Wang & Mccallum, 2006) and the user-question-answer (UQA) model (Guo, Xu, Bao, & Yu, 2008) have been proposed, they are all suitable for analysis from authors' perspective. However, these models establish relationships between topics and a single label rather than multi-labels. Since the task of social emotion detection is multi-labels classification problem, these models can not be directly applied to this task. More recently, two topic models from the readers' perspective, named multilabel supervised topic model (MSTM) and sentiment latent topic model (SLTM) (Rao, Li, Mao, & Wenyin, 2014b), were proposed to overcome this limitation. MSTM was an extension of STM, it firstly generated some topics from words, and then obtained emotions from each topic. Meanwhile, SLTM generated topics directly from social emotions. Similarly, the affective topic model (ATM) was proposed to associate each topic with words and emotions jointly (Rao et al., 2014c). However, it is quite difficult for these models to train a universal classifier for varied corpora.

2.4. Topic models over short text

Research studies on short text clustering mainly focus on exploiting external knowledge for enriching short texts. For example, Jin, Liu, Zhao, Yu, and Yang (2011) learned topics on short texts via transfer learning from auxiliary long text data. Phan, Nguyen, and Horiguchi (2008) learned hidden topics from large external resources to enrich the representation of short texts. Sahami and Heilman (2006) suggested a search-snippet-based similarity measure for short texts. Unfortunately, these methods can have an effective topic distribution only if external dataset is very similar with the original corpus.

To alleviate the problem of text sparsity, biterm topic model (BTM) (Cheng et al., 2014) was therefore proposed for short text clustering and topic learning. The key idea of BTM is to learn topics over short texts based on the assumption that two frequently co-occurred words are more likely to belong to a same topic (Cheng et al., 2014). Specifically, the BTM considers that the whole corpus as a mixture of topics, where each biterm is drawn from a specific topic independently. Experimental results have shown that the effectiveness of BTM outperforms LDA in short text clustering and topic learning. To accelerate the sampling process of topics, He et al. proposed a FastBTM by combining the alias method and Metropolis-Hastings sampling (He, Xu, Li, He, & Yu, 2017). However, both BTM and FastBTM are unsuitable for social emotion detection as they are unsupervised learning methods. By consolidating the main advantage of the BTM, the UAM is proposed for social emotion detection over short texts.

3. Universal affective model

The universal affective model (UAM) is proposed for classifying readers' emotion over short texts. UAM structurally consists of two sub-models at the topic-level and term-level respectively. These two sub-models will respectively learn from keyword corpus (i.e., documents only containing keywords) and background word corpus (i.e., documents only containing background words). The method to determine whether a word is a keyword is to calculate the ATF-IDF weight of that over the whole corpus. In this section, firstly, the ATF-IDF paradigm to determine whether a word is a keyword or not is introduced. Secondly, the concept of the biterm and how to extract biterms from a document are discussed. Thirdly, the research problem including the relevant general terms and notations is formally defined, and then graphical model of UAM is presented. Finally, we describe the estimation of parameters and prediction of emotions.

3.1. Average term frequency-Inverse document frequency

In information retrieval, Term Frequency Inverse Document Frequency (TF-IDF) (Salton & Yu, 1974) is an frequently used method to extract keywords in a document. Based on TF-IDF, we propose a new paradigm to evaluate how important a term is in a corpus, which named ATF-IDF (Average Term Frequency Inverse Document Frequency). Typically, the ATF-IDF weight is composed by two components: (i) the first computes the normalized Average Term Frequency (ATF), which means the number of average times a word appears in all documents of a corpus; and (ii) the second part is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. The multiplication of the two components measures for how important a term is in a corpus. It is formally defined as follows.

$$ATF-IDF(t) = aver(t)\log\frac{|D|}{|D_t|} \tag{1}$$

where t denotes the specific term in a corpus, $aver(t)$ is a function to compute the number of average term frequency of t in a corpus, and $|D|$ and $|D_t|$ refers to total number of documents and number of documents containing term t respectively. The advantage of ATF-IDF is indicated from the qualitative analysis that ATF-IDF can evaluate how important a term is in a corpus while TF-IDF can only extract keywords in a document. Thus, ATF-IDF is more suitable to extract the global keywords and background words in a corpus for our UAM than TF-IDF.

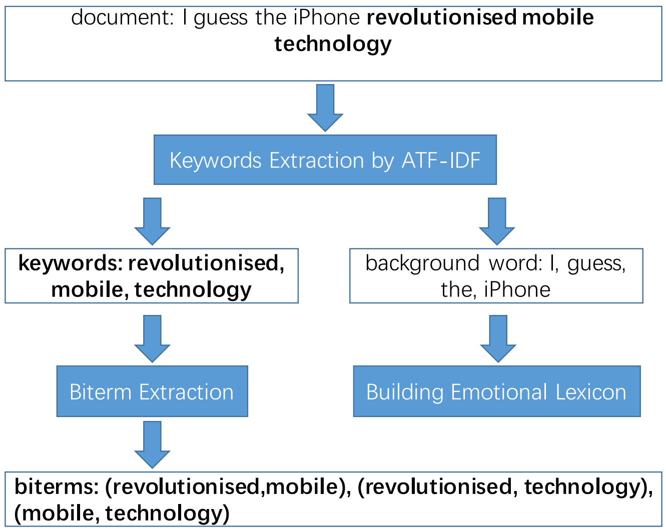

document: I guess the iPhone **revolutionised mobile technology**
Keywords Extraction by ATF-IDF
keywords: revolutionised, mobile, technology
background word: I, guess, the, iPhone
Biterm Extraction
Building Emotional Lexicon
biterms: (revolutionised,mobile), (revolutionised, technology), (mobile, technology)

Fig. 1. The procedure of biterms extraction for a document.

3.2. Biterm extraction based on keywords

Conventional topic models derived from LDA usually focus on the word co-occurrence patterns in documents. In addition, the frequently co-occur words are supposed to belong a same topic in these models. However, the short texts only provide very limited co-occur information for inferring topics.

Unlike other topic models, UAM is extended from BTM by separating all documents into keyword and background corpus. The assumption of UAM is that the whole keyword corpus as a mixture of topics and learn topics based on global biterms extracted from keyword corpus, which may alleviate the problem of text sparsity and enhance the semantic relationships of biterms. A biterm denotes an unordered word-pair co-occurring in a proper text window of the document (i.e., an instance of word co-occurrence pattern) (Cheng et al., 2014). Fig. 1 shows the procedures of biterms extraction from a document. Note that the keywords and biterms in the figure are highlighted in boldface.

To enrich word co-occurrence patterns, UAM extracts each two distinct keywords in a short text as a biterm. For example, there is a text consisted of four distinct words including three keywords w^t and one background word w^b:

$$(w_1^t, w_2^b, w_3^t, w_4^t);$$

biterms extracted from the above text should be:

$$(w_1^t, w_3^t), (w_1^t, w_4^t), (w_3^t, w_4^t).$$

Two words composing a biterm are likely to belong a same topic, which is denoted as groups of correlated words in topic models. To improve the correlation of the word-pair, fitted window size, which limits the distance between the positions of the word-pair co-occurring in a short text, should be applied to the procedure of biterms extraction. In the above example, the window size is set to be 1, so w_1^t is only combined with w_3^t to compose a biterm. The detail procedure of the biterms extraction for each document is presented in Algorithm 1.

The biterms extracted from all the training documents compose the global biterm set. The UAM directly includes all biterms with latent topics in the global biterm set rather than word tokens in a single document. In this way, UAM can fully exploit the rich global word co-occurrence patterns to identify the latent topics.

Algorithm 1 Biterms extraction for each document.

```
Input:
$S$: The window size of the biterms extraction;
$N_d^t$: The number of keywords in the document $d$;
$\mathbf{W}_d^t$: The keyword set in the document $d$;
Output:
$\mathbf{B}_d$: The biterm set of the document $d$;
procedure BITERMS EXTRACTION
    for $i \leftarrow 0; i < N_{w_k,d}; i \leftarrow i + 1$ do
        for $j \leftarrow i + 1; j < \textsc{Min}(j + S, N_d^t); j \leftarrow j + 1$ do
            $w_{k,i}, w_{k,j} \in \mathbf{W}_d^t$ combine to a new biterm $b$;
            add the new biterm $b$ to the biterm set $\mathbf{B}_d$;
        end for
    end for
end procedure
function $\textsc{Min}(X, Y)$
    $result \leftarrow 0$
    if $X <= Y$ then
        $result \leftarrow X$
    else
        $result \leftarrow Y$
    end if
    return $result$
end function
```

3.3. Problem definition

The research question is to predict evoked emotions of unlabeled documents with limited features. Formally, we use the following notations for describing the UAM:

A training data set $(d_1, d_2, d_3, \ldots, d_D)$ consists of D documents containing W^t distinct keywords and W^b distinct background words, and total B biterms are extracted from the training set. After biterms extraction, a document d generates $N_{d,\,b}$ biterms ($B = \sum_d^D N_{d,b}$). Readers can vote the document with predefined a list of E emotion labels. Supposed that there is a corpus of two documents d_1 and d_2, and the predefined five emotion labels are "anger", "guilt", "joy", "shame" and "sadness". In this example, D and E equal to 2 and 5 respectively. Furthermore, we assume that each document has ratings over the five emotions, e.g., $E_{d_1} = (3, 0, 0, 1, 2)$ and $E_{d_2} = (0, 4, 2, 1, 0)$. Accordingly, the document d_1 is voted by 6 readers in total, where three readers voted for "anger". The document d_2 is voted by 7 readers in total, and there is no reader voted for "anger". The whole keyword corpus (i.e., training documents only containing keywords) is depicted by K latent topics. The notations θ, ϕ^t and ϕ^b denotes the keyword corpus-topics distribution, topic-keyword distribution and background corpus-background word distribution, respectively. All above notations are summarized in Table 1.

3.4. Graphical model

As mentioned, UAM structurally consists of two sub-models at the topic-level and term-level respectively. These two sub-models will respectively learn from keyword corpus (i.e., documents only containing keywords) and background word corpus (i.e., documents only containing background words). In particular, the topic-level sub-model of UAM is a supervised topic model which extends BTM for short text classification from reader's perspective by introducing a topic-emotion layer. To overcome the sparisty problem of features in short texts, the topic-level model learns topics by identifying keyword co-occurrence patterns (i,e. biterms) in the keyword corpus. The word-level sub-model of UAM adopts SWAT to establish emotional lexicon with background words. If we assume

Table 1
Notations.

Symbol	Description
K	Number of topics
W^b	Number of distinct background words
N^b	Number of background words in the all training documents
B	Number of biterms in the all training documents
E	Number of emotion labels
E_j	The i-th emotion label
$E_{d,j}$	Rating of the j-th emotion label in the document d
$\mathbf{W}^t$	The distinct keyword set
$\mathbf{W}^b$	The distinct background word set
$\mathbf{B}$	The global biterm set extracted from training documents
b_i	The i-th biterm in the global biterm set $\mathbf{B}$
$w^t_{i,n}$	The n-th keyword of the i-th biterm
w^b_n	The n-th background word in the distinct background word set $\mathbf{W}^b$
d_{b_i}	The document d with biterm b_i in it
d_{w_b}	The document d with background word w_b in it
z_i	The topic associated with the i-th biterm
θ	The $1 \times K$ multinomial distributions of topics specific to the keyword corpus
ϕ^t_z	The $K \times W^t$ multinomial distribution of keywords specific to the topic z
ϕ^b	The $1 \times W^b$ multinomial distribution of background word to the background corpus
$\varepsilon^t_{z,j}$	Rating of the j-th emotion label specific to topic z
$\varepsilon^b_{w^b,j}$	Rating of the j-th emotion label specific to background word w^b
γ	The proportion of background word in all distinct words

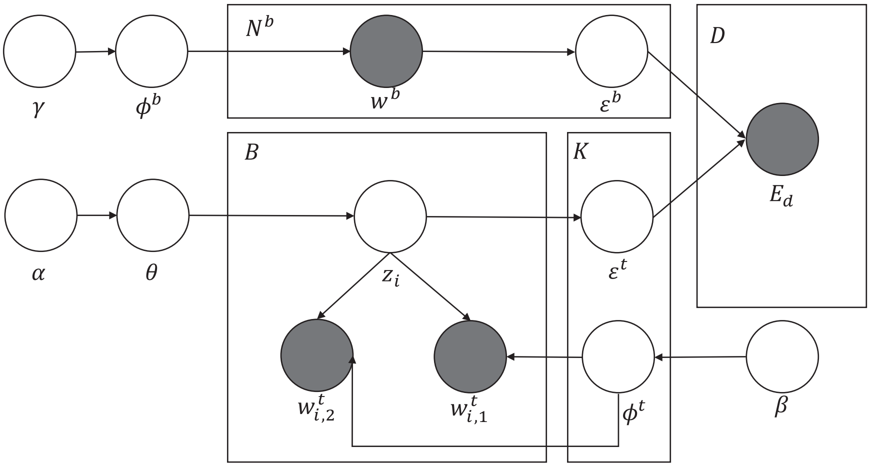


Fig. 2. The graphical model of UAM.

that γ is set to 0 (i.e., if words in all training documents are keywords), UAM becomes a totally supervised topic model. If we assume that γ is set to 1 (i.e., if words in all training documents are background words), UAM becomes a totally word-emotion mapping dictionary. Therefore, UAM can be quite flexible and adapted the learning methods for various tasks.

The graphical representation of UAM is shown in Fig. 2, where shaded nodes are observed data, blank ones are latent (not observed), and arrows indicate dependence. The hyperparameter α is a Dirichlet prior of θ, which can be interpreted as the prior observation counts for the number of times the topic was sampled from document before any words are observed. The hyperparameter β is a Dirichlet prior of ϕ, which can be regarded as the prior observation counts for the frequency of words which were sampled from topic before any actual words have been observed (Lin & He, 2009). Keywords $w^t_{i,1}$ and $w^t_{i,2}$ belong to the i-th biterm in the biterm set $\mathbf{B}$, and each biterm is drawn from a specific topic independently. To classify the social emotions from readers' perspective, each latent topics with biterms are supposed to be modelled with readers' emotion jointly. Since many sites[1] have provided emotion polling services that existing votes are invisible for new readers. In other words, every new reader votes are independent on the old readers. We therefore choose exponential distribution to describe the emotion distribution related to latent topics (Rao et al., 2014c). The memorylessness of exponential variable is an essential characteristic for exponential distribution. Formally, the distribution can be described as

$$P(E > init + new | E > init) = P(E > new),$$

where E, ini and new are exponential variable of the emotion rating, initial votes and new readers' votes respectively. Due to the characteristic of memorylessness, exponential distribution is capable to depict readers voting for different documents (or latent topics). The parameter of the j-th emotion label associated to topic z is denoted by $\lambda_{z,j}$, which is the rate parameter of the corresponding exponential distribution, i.e., the decline rate of reader ratings on the j-th emotion label belonging to topic z. As reviewed in Section 2, many extant methods can offer to directly model background words with emotion labels, while SWAT is one of few studies for establishing the emotional dictionary from the perspective of readers.

For each document d, biterms, background words and reader's ratings are generated as shown in Fig. 3.

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

```
For keyword corpus, draw $\theta \sim \text{Dirichlet}(\alpha)$
For each topic $z \in [1, K]$, draw $\phi^t_z \sim \text{Dirichlet}(\beta)$
For each biterm $b_i \in \mathbf{B}$:
    Generate a topic $z_i \sim \text{Multinomial}(\theta)$
    Generate $w^t_{i,1}, w^t_{i,2} \in b_i \sim \text{Multinomial}(\phi^t_{z_i})$
    For emotion $E_j \in [1, E]$:
        Generate $E_{d_{b_i},j} \sim \text{Exponential } (\lambda_{z_i,j})$
For background word corpus, draw $\phi^b$ according to $\gamma$
For each background word $w^b \in \mathbf{W}^b$
    Generate a $w^b \sim \text{Multinomial}(\phi^b)$
    For emotion $E_j \in [1, E]$:
        Generate $E_{d_{w^b},j} \sim \text{Multinomial}(\varepsilon^b_{w_b})$
```

Fig. 3. Generative processes of UAM.

UAM assumes that each document is composed by keywords and background words, which respectively serve sub-models at the topic-level and word-level of UAM as training data. Therefore, the observed emotion label $E_{d,\,j}$ in the document is generated from $\varepsilon^t_{z,j}$ and $\varepsilon^b_{w_b,j}$.

3.5. Parameter estimation and prediction

Gibbs sampling, a widely used method in topic models (Bao et al., 2009; 2011; Griffiths & Steyvers, 2004; Lau, Xia, & Ye, 2014), is adopted to estimate the parameters of UAM. After the biterms are extracted from the training documents, we randomly initialize topic assignments for all biterms, and then change each biterm's topic with Gibbs sampling method according to the topics of other biterms. The conditional posterior distribution of each biterm b_i can be described as follow:

$$P(z_i = z|\mathbf{z}_{\neg i}, \mathbf{B}, \mathbf{E}; \alpha, \beta, \lambda_{z,1}, \ldots, \lambda_{z,|E|}) \propto (n_{\neg i,z} + \alpha) \frac{(n_{\neg i,w^t_{i,1}|z} + \beta)(n_{\neg i,w^t_{i,2}|z} + \beta)}{(n_{\neg i,\cdot|z} + W^t\beta + 1)(n_{\neg i,\cdot|z} + W^t\beta)} \prod_{e=1}^{|E|} \lambda_{z,j} e^{-\lambda_{z,j} E_{d_{b_i},j}} \tag{2}$$

The Eq. (2) indicates the probability of the biterm b_i assigned to topic z where n_z presents the number of biterms assigned to topic z, $n_{w^t|z}$ is the number of times that keyword w^t assigned to topic z, $n_{\cdot|z}$ is the frequency of all keywords assigned to topic z, and $E_{d_{b_i},j}$ denotes the rating of the j-th emotion label in the document including biterm b_i. The subscript $\neg i$ indicates that the number does not include the current assignment of biterm b_i. In particular, we assume that readers' votes for emotion labels are independent with each other. Based on this assumption, the exponential variable $\lambda_{z,\,j}$ can be estimated by using the method of maximum likelihood as follows.

$$\lambda_{z,j} = \frac{n_z}{\sum_{d_z} E_{d_z,j}} \tag{3}$$

where n_z is the number of all biterms assigned to the topic z, d_z presents the documents including biterms assigned to topic z, and $E_{d_z,j}$ is the documents' proportion of each emotion label.

After Gibbs Sampling of the latent topics for each biterm, we estimate θ, ϕ and ε as follows.

$$\theta_z = \frac{n_z + \alpha}{B + K\alpha}, \tag{4}$$

$$\phi^t_{z,w^t} = \frac{n_{w^t|z} + \beta}{n_{\cdot|z} + W^t\beta}, \tag{5}$$

$$\varepsilon^t_{z,j} = \frac{1}{\lambda_{z,j}} \tag{6}$$

where $\varepsilon^t_{z,j}$ indicates the emotion probability distribution conditioned to relevant topics. The estimation procedures of θ_z, ϕ^t_{z,w^t} and $\varepsilon^t_{z,j}$ are presented in Algorithm 2.

Algorithm 2 Gibbs sampling algorithm for UAM.

```
Input:
$K$: The topic number;
$\mathbf{B}$: The biterms extracted from keywords corpus;
$\mathbf{E}$: The emotion labels of training documents;
$\alpha$: The hyperparameter of $\theta$;
$\beta$: The hyperparameter of $\phi$;
$N_{iter}$: The iteration times of the Gibbs sampling algorithm;
Output:
$P(e|d)$: The emotion proportion of document $d$;
procedure BUILD UAM
    Randomly initialize topic assignments for all biterms;
    Record emotion rating assignments for all topics;
    repeat
        for all $b_i = (w^t_{i,1}, w^t_{i,2}) \in \mathbf{B}$ do
            Sample topic $z$ according to Equation 2;
            Update $n_z$, $n_{z|w^t}$, $\lambda_{z,j}$;
        end for
    until $N_{iter}$ times
    Generate $\theta$, $\phi$, $\varepsilon^t_{z,j}$ according to Equation 4,5,6;
end procedure
```

The parameters estimation of the sub-model at the topic-level of UAM are introduced. For the sub-model at word-levels, we adopt SWAT to create a background word-emotion mapping relationship without adding synonyms and antonyms of the emotion labels (Katz et al., 2007), and the mapping relationship $\varepsilon^b_{w^b,j}$ are shown as follows.

$$\varepsilon^b_{w^b,j} = \sum_{w^b \in d} E_{d,j} \tag{7}$$

Therefore, the emotion probability distribution for an unlabeled document $\tilde{d}$ can be estimated as follows.

$$E_{\tilde{d},j} = \omega_1 \sum_{z}^{K} \varepsilon^t_{z,j} P(z|\tilde{d}) + \omega_2 \frac{\sum_{w^b \in \tilde{d}} \varepsilon^b_{w^b,j}}{\sum_{E_j \in E} \sum_{w^b \in \tilde{d}} \varepsilon^b_{w^b,j}} \tag{8}$$

where weights ω_1 and ω_2 represent the proportion of keywords and background words in $\tilde{d}$ respectively. Assumed that $N_{b,\tilde{d}}$ is the number of biterms in $\tilde{d}$, and $P(z|\tilde{d})$ is the topic probability distribution conditioned to document $\tilde{d}$:

$$P(z|\tilde{d}) = \sum^{N_{b,\tilde{d}}} [P(z|b_i)P(b_i|\tilde{d})] \tag{9}$$

In the Eq. 9, we estimate $P(z|b_i)$ by using Bayes Theorem:

$$P(z|b_i) = \frac{P(z|b_i)}{P(b_i)} = \frac{\theta_z \phi_{z,w^t_{i,1}} \phi_{z,w^t_{i,2}}}{\sum_{\tilde{z}} [\theta_{\tilde{z}} \phi_{\tilde{z},w^t_{i,1}} \phi_{\tilde{z},w^t_{i,2}}]} \tag{10}$$

and $P(b_i|\tilde{d})$ can be estimated as:

$$P(b_i|\tilde{d}) = \frac{n_{\tilde{d}}(b_i)}{N_{b,\tilde{d}}} \tag{11}$$

where $n_{\tilde{d}}(b_i)$ denotes the number of b_i in the document $\tilde{d}$. Given the estimation of $E_{d,\,j}$ for the testing set, we can evaluate the performance of UAM for emotion detection.

4. Experiments

In this section, we design the experiments to evaluate the performance of the proposed model for readers' emotion detection over short texts. The experiments are conducted to achieve the following two goals: (i) to observe the influence of topic number on the task of short text emotion detection; (ii) to compare the performance of UAM with other supervised topic models; and (iii) to compare the performance of UAM with word-level and deep learning models.

4.1. Experiment design

In our experiments, two datasets of English short texts and a dataset of Chinese long texts are selected to verify the effectiveness of the UAM. The detail information about these datasets are introduced as follows.

SemEval: This is a short text collection used in the 14-th task of the 4-th International Workshop on Semantic Evaluations (SemEval-2007) (Katz et al., 2007). The collection includes 1246 valid the news headline and user scores over emotions of “anger”, “disgust”, “fear”, “joy”, “sad” and “surprise”, which are normalized from 0 to 100. As for 1246 short texts, the first 246 of them are used to train the proposed model, and the remaining 1000 are used to be a testing set. In general, the documents in *SemEval* are short messages and almost have the identical number of words.

Six: This is a dataset including 11,813 short texts collected from BBC Forum posts (BBC), Digg.com posts (Digg), MySpace comments (MySpace), Runners World forum posts (Runners World), Twitter posts (Twitter) and YouTube comments (YouTube). Each text was manually labelled by readers with negative and positive sentiment strengths. The positive sentiment strength ranges from 1 (not positive) to 5 (extremely positive), and the negative sentiment strength ranges from −1 (not negative) to −5 (extremely negative). Considering that the number of texts in the dataset is sufficient for model training, we randomly choose 60% of texts as the training set, 20% as the validation set, and the rest as the testing set. In general, although the documents in *Six* are short sentences, the length of texts is quietly unbalanced. That is, some documents contain a few words (3 or 4), while other documents may contain sentences with more than 20 words.

Sinanews: This is a dataset of long texts collected from the Society channel of Sina. *Sinanews* contains 4570 news articles and the attributes of each article include the URL address, new title, publishing date (from January to April of 2012), content and user ratings over 8 emotion labels: “touching”,“empathy”, “boredom”, “anger”, “amusement”, “sadness”, “surprise” and “warmness”. As for 4570 valid news articles, we randomly sample 2342 texts as training set, and the remaining 2228 forms the testing set. Generally, the sentence length of documents in *Sinanews* are long sentences (i.e., more than 20 words).

As a preprocessing step, all datasets are split into keywords and formatted standardly. Unlike English words in the first two datasets, consecutive Chinese words in a document for the third dataset are not separated by spaces. To deal with this, we use a straightforward English tokenizer which exploits spaces and punctuations, and a Chinese lexical analysis system (ICTCLAS, http://www.ictclas.org/) to extract words from English and Chinese documents, respectively. We split the dataset into a training set and a testing set, and evaluate the performance by 5-fold cross-validation (Blei & Mcauliffe, 2010). We also implement following topic-level baselines for comparison:

- Affective Topic Model (ATM) (Rao et al., 2014c). As a multi-labeled topic model, ATM allows us to associate each topic with word tokens and social emotions jointly. To generate user ratings for each emotion label, the exponential distribution is employed for ATM.
- Multi-label Supervised Topic Model (MSTM) (Rao et al., 2014b). MSTM extends Supervised Topic Model (STM) for social emotion mining and models multiple kinds of labels rather than a single type of labels.

Table 2
Performance statistics of different models.

(a) *SemEval*

Models	$AP_{document}$		$AP_{emotion}$		*Accu*@1	
	Mean	Variance	Mean	Variance	Mean	Variance
UAM	**0.3227**	0.0004	**0.2027**	0.0005	**0.3491**	0.0001
ATM	**0.3043**	**7.71E−05**	0.0736	0.0012	**0.3626**	**2.81E−05**
MSTM	0.1887	**4.81E−05**	0.0086	**0.0003**	0.2089	**2.99E−05**
SLTM	0.1697	0.0076	0.041	0.0013	0.2269	0.0022
ETM	0.1971	0.0014	**0.0844**	**0.0008**	0.2674	0.0004

(b) *Six*

Models	$AP_{document}$		$AP_{emotion}$		*Accu*@1	
	Mean	Variance	Mean	Variance	Mean	Variance
UAM	**0.5319**	**0.0001**	**0.3930**	0.0014	**0.7660**	**2.31E−05**
ATM	0.2596	3.18E-04	0.1986	**0.0007**	0.6298	7.94E-05
MSTM	0.4989	3.16E-04	0.1536	0.0015	0.7494	7.91E-05
SLTM	0.4573	**7.04E−05**	0.0481	0.0015	0.7286	**1.76E−05**
ETM	**0.6777**	2.62E-04	**0.4044**	**0.0003**	**0.8388**	6.55E-05

(c) *Sinanews*

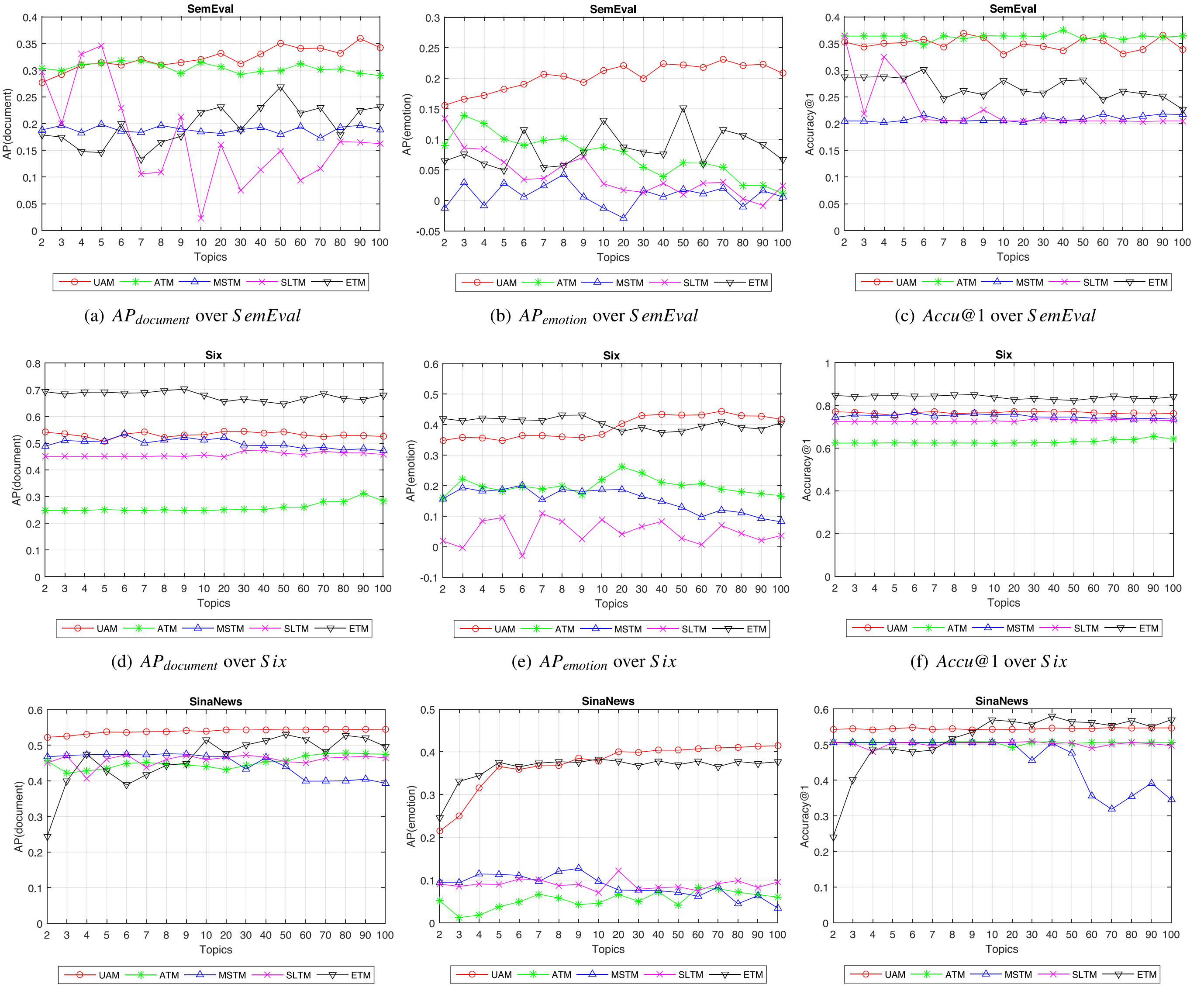
Models	$AP_{document}$		$AP_{emotion}$		*Accu*@1	
	Mean	Variance	Mean	Variance	Mean	Variance
UAM	**0.5392**	**4.35E−05**	**0.3703**	0.0032	**0.5448**	**4.19E−06**
ATM	0.452	0.0003	0.0541	**0.0004**	0.5056	**1.33E−05**
MSTM	0.4477	0.0011	0.0864	0.0007	0.4592	0.0049
SLTM	0.4591	**0.0003**	0.0898	**0.0001**	0.502	4.98E-05
ETM	**0.4625**	0.0050	**0.3631**	0.0010	**0.5152**	0.0069

- Sentiment Latent Topic Model (SLTM) (Rao et al., 2014b). SLTM can associate each topic with words and emotions jointly, and infer the emotion distribution of future unlabeled documents.
- Emotion Topic Model (ETM) (Bao et al., 2011). ETM is proposed by augmenting Latent Dirichlet Allocation with an additional layer for emotion modeling and allows us to infer a number of conditional probabilities for unseen documents, e.g., the probabilities of latent topics given an emotion and terms given in a topic.

Three evaluation metrics are employed as indicators of performance: $AP_{document}$ (Rao, 2016), $AP_{emotion}$ (Katz et al., 2007), and *Accu*@1 (i.e., the Accuracy at top 1) (Bao et al., 2009; 2011). As a fine-grained metric based on average Pearson’s correlation coefficient, *AP* takes emotional distributions into consideration. For each document, $AP_{document}$ measures correlations between predicted probabilities and actual user ratings over all the emotion labels. For each emotion label, $AP_{emotion}$ measures correlations between predicted probabilities and actual votes over all documents. *Accu*@1 is defined as the percentage of future documents whose top ranked emotion label is correctly predicted.

4.2. Influence of topic number

In topic models, the number of latent topics of documents influences the performance. To evaluate the influence of the number of topics K, we vary K from 2 to 100 for all datasets. The hyperparameters α and β are set to symmetric Dirichlet priors with values of $50/K$ and 0.1, respectively; iteration times of Gibbs is set to 300; and UAM’s windows size of biterm extraction and background word proportion γ are optimized from the validation set. After five models predicted the emotion distribution of each unlabeled document, we compute their the mean and variance of $AP_{document}$, $AP_{emotion}$ and *Accu*@1. Fig. 4 illustrates the performance of all the five models on *SemEval, Six* and *Sinanews* under different K values. Table 2 shows the mean and variance of different models in the three metrics. Note that top 2 values of mean and variance in terms of Pearson’s correlation coefficient are highlighted as boldface.



(a) $AP_{document}$ over $SemEval$ (b) $AP_{emotion}$ over $SemEval$ (c) $Accu@1$ over $SemEval$

(d) $AP_{document}$ over Six (e) $AP_{emotion}$ over Six (f) $Accu@1$ over Six

(g) $AP_{document}$ over $SinaNews$ (h) $AP_{emotion}$ over $SinaNews$ (i) $Accu@1$ over $SinaNews$

Fig. 4. Model performance with different topic numbers.

The experimental results over *SemEval* (i.e., the English short texts) show that the $AP_{document}$ and mean $AP_{emotion}$ of UAM tend to be increased with the increasing topic number. The *Accu@*1 of UAM tend to be stable with small topic number but unstable with the large one. Since *SemEval* is a very sparse dataset, all methods including UAM and baselines are not quite stable with varying of the topic number.

UAM outperforms all others baselines on $AP_{emotion}$ under all pre-defined topic numbers and on $AP_{document}$ when the K is not less than 7. For metric *Accu@*1, the mean *Accu@*1 of UAM is the second best performance. Specifically, compared with ATM, MSTM, SLTM and ETM,

1. the mean $AP_{document}$ of UAM improves all baselines with 1.8%, 13.4%, 15.3%, and 12.6%;
2. the mean $AP_{emotion}$ of UAM improves all baselines with 12.9%, 19.4%, 16.2%, and 11.8%;
3. and the mean *Accu@*1 of UAM is similar to with ATM (less than 1.3%), and still improves MSTM, SLTM and ETM with 14%, 12.2%, and 8.2%.

Furthermore, the UAM's variances, which are the top-3 least for all metrics, indicates that the performance of UAM is quite stable. The experimental results of UAM is consistent with the assumptions that the whole keyword corpus as a mixture of topics and learn topics based on global biterms extracted from keyword corpus, which may alleviate the problem of text sparsity and enhance the semantic relationships of biterms.

The experimental results over *Six* (i.e., English texts) show that the $AP_{document}$ and *Accu@*1 of UAM tend to be stable and the $AP_{emotion}$ becomes larger with the increasing topic numbers. Specifically, the mean values of three metrics of UAM has the second best performance, while ETM has the best performance. However, it is worth to point out that UAM's performance on $AP_{emotion}$ starts to catch up with ETM when the topic number is larger than 20. Compared with ATM, MSTM, and SLTM,

1. the mean $AP_{document}$ of UAM improves above three baselines with 27.2%, 3.3% and 7.5%;
2. the mean $AP_{emotion}$ of UAM improves above three baselines with 19.4%, 23.9% and 34.5%;

3. and the mean *Accu*@1 of UAM improves above three baselines with 13.6%, 1.7%, and 3.7%.

In addition, the UAM's variances, the second least in both $AP_{document}$ and *Accu*@1 and the third least in $AP_{emotion}$, indicates that UAM also has quite stable performance in a different dataset. The performance of UAM is not the best but still better than most baselines in the dataset. The reason could be the unbalanced length of documents in this dataset. After extracting biterms from all the documents, the distance of the number of biterms between short and long texts are enlarged. The frequency of the co-occurred word-pairs belonging to long texts is also unreasonably increased in the global biterm set, so topics learnt by UAM is overfitting.

The experimental results over *Sinanews* (i.e., the Chinese long text dataset) show that the $AP_{document}$ and *Accu*@1 of UAM tend to be stable and the $AP_{emotion}$ becomes larger with the increasing topic numbers. UAM yields competitive performance on the three metrics and achieves the best $AP_{document}$ and $AP_{emotion}$. Compared with ATM, MSTM, SLTM and ETM,

1. the mean $AP_{document}$ of UAM improves all baselines with 8.7%, 9.2%, 8.0%, and 7.7%;
2. the mean $AP_{emotion}$ of UAM improves all baselines with 31.6%, 28.4%, 28.0% and 0.7%;
3. and the mean *Accu*@1 of UAM improves all baselines with 3.9%, 8.6%, 4.3%, and 3%.

In addition, the variances of UAM, has the least values in both $AP_{document}$ and *Accu*@1, indicate that the performance is also stable in this dataset. In general, UAM has better performance compared with baselines in *Sinanews*. Since *Sinanews* is a long text dataset, the results also support our assumption that emotion labels with semantics biterms are better than models with single word tokens over long texts. From the experimental results on above three datasets, UAM has more stable and better performance than baseline methods in most cases. These results support that the proposed UAM is an accurate and stable supervised model for reader's emotion detection for both short texts and long texts.

To statistically evaluate the differences between UAM and baselines, we perform two statistical tests on paired models in terms of three metrics. Firstly, the analysis of variance *F*-test is conducted on UAM with ATM, MSTM, SLTM and ETM respectively. We then employ *t*-test to evaluate the differences of these paired models in mean performance according to the *F*-test results. We use the conventional significance level (i.e., *p*-value) of 0.05. Table 3 shows the statistical analysis of results between UAM and baselines (*p*-values larger than 0.05 are highlighted in boldface).

As shown in Table 3(a), the *F*-test results on $AP_{document}$ and *Accu*@1 over *SemEval* between UAM and all baselines are less than 0.05, yet results on $AP_{emotion}$ between the UAM and baselines of MSTM, ETM are larger than 0.05. These results indicate that the UAM is significantly more stable than baselines statistically. The *p*-values of *t*-test are almost less than 0.05, and further implies that the proposed UAM outperforms baselines. As shown in Table 3(b), the *F*-test results on between the UAM and baselines over *Six* are almost less than 0.05 except for SLTM. The *t*-test results are all less than 0.05 except for ETM, and further support that the UAM is the most effective model. As shown in Table 3(c), the *F*-test and *t*-test results on between the UAM and baselines over *Sinanews* are less than 0.05, and further support the UAM is a stable and effective model for short text emotion detection.

To sum up, the UAM performs well over *SemEval, Six* and *Sinanews*. As discussed in Section 3, UAM enhances the semantic relationships of biterms by extracting keywords and bridging topics of biterms with emotion labels from readers' perspective. Therefore, biterm's topic probability distribution is strongly connected with emotion labels.

To explore the detailed model performance over each emotion label, the mean correlation coefficient with 2 to 100 topics is summarized in Table 4. Note that top 2 values are highlighted in boldface. As shown in Table 4(a), UAM outperforms all baselines for all emotions over *SemEval*. Particularly, UAM's result on the "fear" emotion reaches the global highest value of 0.3005, which improves ATM, MSTM, SLTM, and ETM by 20.22%, 28.67%, 26.23%, and 21.01%, respectively. As shown in Table 4(b), UAM achieves the second best performance over *Six*, which is very close to the best-performing baseline of ETM in each emotion (less than 1.14%). As shown in Table 4(c), UAM achieves the second best performance on "boredom" and "sadness" emotions while ranks top 1 on other six emotions over *SinaNews*. Specifically, UAM's performance on the "touching" emotion reaches the global highest value of 0.695, which improves ATM, MSTM, SLTM, and ETM by 52.2%, 47.85%, 45.83%, and 3.87%, respectively. These results indicate that the proposed UAM also performs quite competitive in each emotion over given datasets.

4.3. Compared with word-level and deep learning baselines

In this subsection, we further compare UAM with some state-of-the-art methods which do not exploit latent topics. As mentioned in Section 2, SWAT is one of the top-performing systems on the affective text task in SemEval-2007 (Katz et al., 2007), which uses emotional ratings of news headlines and scored emotions of each word with a unigram model. The ET adopted the naive Bayes method to model word-emotion associations, and considered emotion ratings when estimating parameters (Bao et al., 2011). CharSCNN is a deep neural network architecture which jointly uses representation at the character-level, word-level and sentence-level to perform sentiment analysis (Santos & Gattit, 2014). To learn word-level embeddings, we adopt English twitter corpus[2] and Chinese Wiki-pedia corpus[3] as sources of unlabeled data for CharSCNN. We then employed 5-fold cross-validation to tune parameters of UAM and CharSCNN. Table 5 shows the experimental results with baselines at non-topic levels (the best values are highlighted in boldface).

As shown in Table 5(a), the $AP_{emotion}$ and *Accu*@1 of UAM outperform other baselines. $AP_{document}$ of UAM also has the second best performance. Therefore, UAM is capable to distinguish different meanings of the same word in various contexts. That is, the model is suitable for multi-emotion label classification from readers' perspective. As shown in Table 5(b), the word-level models (i.e., SWAT and ET) perform better than UAM and CharSCNN on all three metrics. Since *Six* is a two labelled dataset which contains sparse words and information at other levels (e.g., the topic level) are inapplicable, the word-level models such as SWAT and ET are more effective to make use of these limited word features. Another reason is that biterms generated from UAM are overfitting in long texts as discussed in Section 4.2. As shown in Table 5(c), $AP_{emotion}$ and *Accu*@1 of UAM outperform other baselines. $AP_{document}$ of UAM also has the second best performance. The results indicate that extracting keywords as biterms enables the learning of meaningful topics over long texts in *Sinanews*.

To sum up, UAM has gained competitive performance for social emotion detection compared to word-level models and deep learning models.

4.4. Discussions

Suitable scenario of different models: In terms of the accuracy of emotion classification, it is more suitable for UAM to clas-

[2] https://nlp.stanford.edu/projects/glove/.

[3] https://dumps.wikimedia.org/zhwiki/latest/.

Table 3
The *p* values of statistical tests on the UAM and baselines.

(a) *SemEval*

Models	$AP_{document}$		$AP_{emotion}$		*Accu*@1	
	F-test	*t*-test	*F*-test	*t*-test	*F*-test	*t*-test
ATM	5.31E−04	9.58E−04	3.06E−02	4.00E−14	1.44E−03	6.21E−05
MSTM	2.37E−05	7.83E−18	**2.16E−01**	7.40E−26	2.06E−03	1.96E−25
SLTM	1.12E−07	3.65E−07	2.73E−02	3.27E−16	1.91E−07	8.40E−10
ETM	1.01E−02	4.01E−13	**1.33E−01**	8.04E−16	1.51E−02	4.58E−15

(b) *Six*

Models	$AP_{document}$		$AP_{emotion}$		*Accu*@1	
	F-test	*t*-test	*F*-test	*t*-test	*F*-test	*t*-test
ATM	7.46E−03	3.75E−29	8.13E−02	1.86E−19	7.46E−03	3.75E−29
MSTM	7.60E−03	1.16E−07	**4.20E−01**	5.61E−20	7.60E−03	1.16E−07
SLTM	**2.90E−01**	1.14E−23	**4.27E−01**	4.54E−25	**2.90E−01**	1.14E−23
ETM	1.92E−02	3.30E−24	2.69E−03	**1.25E−01**	1.92E−02	3.30E−24

(c) *Sinanews*

Models	$AP_{document}$		$AP_{emotion}$		*Accu*@1	
	F-test	*t*-test	*F*-test	*t*-test	*F*-test	*t*-test
ATM	9.51E−05	9.61E−16	2.50E−05	1.88E−16	1.09E−02	8.40E−26
MSTM	9.20E−09	4.63E−10	1.18E−03	1.91E−16	0	3.69E−05
SLTM	3.21E−04	3.90E−16	1.30E−08	2.95E−14	2.68E−06	9.07E−17
ETM	3.35E−14	1.34E−04	1.19E−02	**3.22E−01**	0	7.37E−02

Table 4
Model performance on the correlation coefficient over each emotion label.

(a) *SemEval*

models/labels	anger	disgust	fear	joy	sad	surprise
UAM	**0.1637**	**0.1125**	**0.3005**	**0.2451**	**0.2115**	**0.1829**
ATM	**0.0664**	**0.0494**	**0.0983**	0.0953	0.0595	**0.0727**
MSTM	−0.0060	0.0094	0.0138	0.0229	0.0095	0.0053
SLTM	0.0349	0.0375	0.0382	0.0690	0.0306	0.0360
ETM	0.0606	0.0363	0.0904	**0.1861**	**0.0739**	0.0593

(b) *Six*

models/labels	positive	negative
UAM	**0.3930**	**0.3930**
ATM	0.1986	0.1986
MSTM	0.1536	0.1536
SLTM	0.0463	0.0562
ETM	**0.4044**	**0.4044**

(c) *Sinanews*

models/labels	touching	empathy	boredom	anger	amusement	sadness	surprise	warmness
UAM	**0.6590**	**0.2493**	**0.2735**	**0.5212**	**0.3641**	**0.3055**	**0.3752**	**0.2148**
ATM	0.1370	0.0400	0.0257	0.1120	0.0389	0.0646	-0.0020	0.0175
MSTM	0.1805	0.0594	0.0382	0.1743	0.0407	0.0896	0.0476	0.0605
SLTM	0.2007	0.0326	0.0526	0.1715	0.0640	0.0924	0.0552	0.0497
ETM	**0.6203**	**0.2409**	**0.2887**	**0.5175**	**0.3172**	**0.3542**	**0.3629**	**0.2035**

sify datasets with multiple emotion labels (usually more than two) over both short messages and long documents, because synonymous and polysemous words are frequently appeared in a document with hybrid emotions and UAM can address the emotion ambiguity by strongly associating topics with emotion labels. Besides, UAM learns topics via global biterms extracting from keywords, which can alleviate the problem of feature sparsity. The experimental results in Sections 4.2 and dummyTXdummy- 4.3 over *SemEval* and *Sinanews* validated the characteristics of UAM. After analyzing the result in Section 4.3, we observe that when compared with topic-level and deep learning models, word-level methods such as SWAT and ET is not good at classifying datasets with many emotion labels like *SemEval* and *SinaNews*. However, the word-level method is suitable for classifying datasets with only two labels. The reason may be that synonymous and polysemous words seldom appeared in such a kind of dataset.

Factors affecting classification performance: For the proposed UAM, its classification performance may be affected by the number of topics, the proportion of background words among all distinct words, the window size for biterm extraction, the number of emotion labels, and the length of each document. As shown in Section 4.2, our UAM performs stably over *Six* and *SinaNews* under varied parameter values. However, the fluctuation of classification accuracy becomes obvious with different numbers of topics over *SemEval*. This is because short messages with an extremely sparse feature space may result in overfitting. Furthermore, the main difference when applying our model and baseline models of ATM, MSTM, SLTM, ETM, SWAT, and ET to documents written in different languages is the segmentation of words. For the deep neural network model CharSCNN, however, the language of documents has an additional effect on learning word-level embeddings.

Table 5
Compared with word-level and deep learning baselines .

model	$AP_{document}$	$AP_{emotion}$	$Accu@1(\%)$
(a) *SemEval*			
UAM	0.357	**0.241**	**36.7**
SWAT	0.262	0.223	34.1
ET	0.236	0.22	31
CharSCNN	**0.38**	0.014	36.2
(b) *Six*			
model	$AP_{document}$	$AP_{emotion}$	$Accu@1(\%)$
UAM	0.54	0.444	77.2
SWAT	0.636	0.285	81.9
ET	**0.714**	**0.45**	**85.7**
CharSCNN	0.54	0.289	79.8
(c) *Sinanews*			
model	$AP_{document}$	$AP_{emotion}$	$Accu@1(\%)$
UAM	0.545	**0.413**	**54.7**
SWAT	0.472	0.335	50.6
ET	0.427	0.379	42.9
CharSCNN	**0.57**	0.018	53.2

5. Conclusion

In this article, we focus on the problem of how to relieve the text sparsity for readers' emotion detection. To tackle this problem, the Universal Affective Model (UAM) is proposed by composing a brand new supervised topic model and word-level emotional dictionary. Rather than directly modeling emotion labels with the single word token, the topic-level model of UAM is a supervised topic model which extends BTM for short text classification from readers' perspective by introducing a topic-emotion layer. To overcome the sparsity problem of limited features in short texts, the topic-level model learn topics by directly modeling the generation of keyword co-occurrence patterns (i.e., biterms) in the keyword corpus. For our future research, we plan to consolidate UAM with a deep neural networks architecture.

Acknowledgment

The work described in this paper was fully supported by the Innovation and Technology Fund (Project No. GHP/022/17GD) from the Innovation and Technology Commission of the Government of the Hong Kong Special Administrative Region, a grant from Research Grants Council of Hong Kong Special Administrative Region, China (UGC/FDS11/E03/16), Guangdong Science and Technology Program grant 2017A050506025, and the National Natural Science Foundation of China (61502545).

Appendix A. Summary of Acronyms

Acronyms	Details	Reference	Section
ETM	emotion term model	Bao et al. (2009)	Section 1/2/4
ET	emotion term model	Bao et al. (2009)	Section 2.4
LDA	latent dirichlet allocation	Blei et al. (2003)	Section 1/2
ATM	affective topic model	Rao et al. (2014c)	Section 1/2/4
TF-IDF	term frequency inverse document frequency	Salton and Yu (1974)	Section 3
BTM	biterm topic model	Cheng et al. (2014)	Section 1/2
SWAT	semeval-2007 systems	Katz et al. (2007)	Section 1/2/4
SVM	support vector machine	Adankon and Cheriet (2009)	Section 2
WE	word-emotion	Rao et al. (2014a)	Section 2
ELDA	emotion latent dirichlet allocation	Rao et al. (2014a)	Section 2
JSTM	joint sentiment-topic model	Lin and He (2009)	Section 2
STM	supervised topic model	Blei and Mcauliffe (2010)	Section 2
TOT	topic-over-time model	Wang and Mccallum (2006)	Section 2
UQA	user-question-answer model	Guo et al. (2008)	Section 2
MSTM	multi-label supervised topic model	Rao et al. (2014b)	Section 2.4
SLTM	sentiment latent topic model	Rao et al. (2014b)	Section 2.4

(*continued on next page*)

Appendix B. Equations of $AP_{document}$ and $AP_{emotion}$

The average Pearson's correlation coefficient (AP) takes emotional distributions into consideration. For each document, $AP_{document}$ measures correlations between predicted probabilities and actual user ratings over all the emotion labels. We use X and Y represent predicted probabilities and actual votes over an document respectively, then $AP_{document}$ can be estimated as:

$$AP_{document} = \frac{\sum_{i=1}^{|E|} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{(|E| - 1)\sigma_X \sigma_Y} \tag{12}$$

For each emotion label, $AP_{emotion}$ measures correlations between predicted probabilities and actual votes over all documents. We use A and B to denote predicted probabilities and actual votes over an emotion label in all documents respectively, then $AP_{emotion}$ can be estimated as:

$$AP_{emotion} = \frac{\sum_{j=1}^{|D|} \left(A_j - \bar{A}\right)\left(B_j - \bar{B}\right)}{(|D| - 1)\sigma_A \sigma_B} \tag{13}$$

In the above equations, $\bar{X}$ represents the mean of vector X, $|E|$ represents the number of all emotion labels, $|D|$ represents the number of all documents, and σ represents standard deviation.

References

Adankon, M. M., & Cheriet, M. (2009). Support vector machine. *Computer Science, 1*(4), 1–28.

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., & Yu, Y. (2009). Joint emotion–topic modeling for social affective text mining. In *Proceedings of the Ninth ieee international conference on data mining* (pp. 699–704).

Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., & Yu, Y. (2011). Mining social emotions from affective text. *IEEE Transactions on Knowledge and Data Engineering, 24*(99). 1–1.

Blei, D. M., & Mcauliffe, J. D. (2010). Supervised topic models. *Advances in Neural Information Processing Systems, 3*, 327–332.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bollegala, D., Weir, D., & Carroll, J. (2011). Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Meeting of the association for computational linguistics: Human language technologies* (pp. 132–141).

Bosco, C., Patti, V., & Bolioli, A. (2013). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems, 28*(2), 55–63.

Cambria, E., Schuller, B., Liu, B., & Wang, H. (2013a). Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems, 28*(2), 12–14.

Cambria, E., Schuller, B., Liu, B., Wang, H., & Havasi, C. (2013b). Statistical approaches to concept-level sentiment analysis. *IEEE Intelligent Systems, 28*(3), 6–9.

Cambria, E., & White, B. (2014). Jumping nlp curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine, 9*(2), 48–57.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge & Data Engineering, 26*(12), 2928–2941.

Collobert, R., Weston, J., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research, 12*(1), 2493–2537.

Das, S., & Chen, M. (2001). *Yahoo! for amazon: extracting market sentiment from stock message boards.*

Eguchi, K., & Lavrenko, V. (2006). Sentiment retrieval using generative models. In *Conference on empirical methods in natural language processing* (pp. 345–354).

Gangemi, A., Presutti, V., & Reforgiato Recupero, D. (2014). Frame-based detection of opinion holders and topics: a model and a tool. *IEEE Computational Intelligence Magazine, 9*(1), 20–30.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(1) Suppl 1, 5228–5235.

Guo, J., Xu, S., Bao, S., & Yu, Y. (2008). Tapping on the potential of q&a community by recommending answer providers. In *Proceedings of the Acm conference on information and knowledge management, cikm 2008, napa valley, california, usa, october* (pp. 921–930).

He, X., Xu, H., Li, J., He, L., & Yu, L. (2017). Fastbtm: Reducing the sampling time for biterm topic model. *Knowledge Based Systems, 132*, 11–20.

He, Y., Lin, C., & Alani, H. (2011). Automatically extracting polarity-bearing topics for cross-domain sentiment classification.. In *The meeting of the association for computational linguistics: Human language technologies, proceedings of the conference, 19–24 june, 2011, portland, oregon, usa* (pp. 123–131).

Jin, O., Liu, N. N., Zhao, K., Yu, Y., & Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the Acm conference on information and knowledge management, cikm 2011, glasgow, united kingdom, october* (pp. 775–784).

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on web search and data mining* (pp. 219–230).

Katz, P., Singleton, M., & Wicentowski, R. (2007). Swat-mp: The semeval-2007 systems for task 5 and task 14. In *International workshop on semantic evaluations* (pp. 308–313).

Kazemzadeh, A., Lee, S., & Narayanan, S. (2013). Fuzzy logic models for the meaning of emotion words. *IEEE Computational Intelligence Magazine, 8*(2), 34–49.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Empirical Methods in Natural Language Processing*, 1746–1751.

Ku, L., Liang, Y., & Chen, H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. *In AAAI-CAAW*, 100–107.

Lau, R. Y. K., Xia, Y., & Ye, Y. (2014). A probabilistic generative model for mining cybercriminal networks from online social media. *IEEE Computational Intelligence Magazine, 9*(1), 31–43.

Li, X., Rao, Y., Xie, H., Lau, Y. K. R., Yin, J., & Wang, F. L. (2017). Bootstrapping social emotion classification with semantically rich hybrid neural networks. *IEEE Transactions on Affective Computing, PP*(99). 1–1.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the Acm conference on information and knowledge management* (pp. 375–384).

Lin, H. Y., & Chen, H. H. (2008). Ranking reader emotions using pairwise loss minimization and emotional distribution regression. In *Conference on empirical methods in natural language processing, emnlp 2008, proceedings of the conference, 25–27 october 2008, honolulu, hawaii, usa, a meeting of sigdat, a special interest group of the acl* (pp. 136–144).

Moreo, A., Romero, M., Castro, J. L., & Zurita, J. M. (2012). Lexicon-based comments-oriented news sentiment analyzer system. *Expert Systems with Applications An International Journal, 39*(10), 9166–9180.

Pan, S. J., Ni, X., Sun, J. T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the international conference on world wide web, www 2010, raleigh, north carolina, usa, april* (pp. 751–760).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the Acl-02 conference on empirical methods in natural language processing* (pp. 79–86).

Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the international conference on world wide web, www 2008, beijing, china, april* (pp. 91–100).

Poddar, L., Hsu, W., & Lee, M. L. (2017). Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews. In *Proceedings of the Conference on empirical methods in natural language processing* (pp. 472–481).

Quan, C., & Ren, F. (2010). An exploration of features for recognizing word emotion. In *Proceedings of the international conference on computational linguistics* (pp. 922–930).

Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing microblogs with topic models. In *Proceedings of the international conference on weblogs and social media, icwsm 2010, washington, dc, usa, may* (pp. 130–137).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 248–256).

Rao, Y. (2016). Contextual sentiment topic model for adaptive social emotion classification. *IEEE Intelligent Systems, 31*(1), 41–47.

Rao, Y., Lei, J., Liu, W., Li, Q., & Chen, M. (2014a). Building emotional dictionary for sentiment analysis of online news. *World Wide Web, 17*(4), 723–742.

Rao, Y., Li, Q., Mao, X., & Wenyin, L. (2014b). Sentiment topic models for social emotion mining. *Information Sciences, 266*(5), 90–100.

Rao, Y., Li, Q., Wenyin, L., Wu, Q., & Quan, X. (2014c). Affective topic model for social emotion detection. *Neural Networks, 58*(5), 29–37.

Rao, Y., Xie, H., Li, J., Jin, F., Wang, F. L., & Li, Q. (2016). Social emotion classification of short text via topic-level maximum entropy model. *Information & Management, 53*(8), 978–986.

Sahami, M., & Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the international conference on world wide web, www 2006, edinburgh, scotland, uk, may* (pp. 377–386).

Salton, G., & Yu, C. T. (1974). *On the construction of effective vocabularies for information retrieval*. Birkhîauser.

Santos, C. N. D., & Gattit, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the international conference on computational linguistics*.

Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2008). Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence, 22*(6), 558–601.

Shin, B., Lee, T., & Choi, J. D. (2016). Lexicon integrated cnn models with attention for sentiment analysis, (pp. 149–158).

Stoyanov, V., & Cardie, C. (2008). Annotating topics of opinions.. In *Proceedings of the international conference on language resources and evaluation, lrec 2008, 26 may - 1 june 2008, marrakech, morocco* (pp. 3213–3217).

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Annual Meeting of the Association for Computational Linguistics* (pp. 417–424).

Wang, X., & Mccallum, A. (2006). Topics over time:a non-markov continuous-time model of topical trends. In *Proceedings of the Twelfth acm sigkdd international conference on knowledge discovery and data mining, philadelphia, pa, usa, august* (pp. 424–433).

Zhan, X., Wang, Y., Rao, Y., Xie, H., Li, Q., Wang, F. L., & Wong, T. L. (2017). A network framework for noisy label aggregation in social media. In *Meeting of the association for computational linguistics* (pp. 484–490).