Building emotional dictionary for sentiment analysis of online news

Article *in* World Wide Web · July 2014

DOI: 10.1007/s11280-013-0221-9

Yanghui Rao
City University of Hong Kong

Wenyin Liu
GuangDong University of Technology

Qing Li
City University of Hong Kong

Mingliang Chen
City University of Hong Kong

World Wide Web
DOI 10.1007/s11280-013-0221-9

Building emotional dictionary for sentiment analysis of online news

Yanghui Rao · Jingsheng Lei · Liu Wenyin · Qing Li · Mingliang Chen

Received: 28 September 2012 / Revised: 28 February 2013 / Accepted: 2 April 2013

Abstract Sentiment analysis of online documents such as news articles, blogs and microblogs has received increasing attention in recent years. In this article, we propose an efficient algorithm and three pruning strategies to automatically build a word-level emotional dictionary for social emotion detection. In the dictionary, each word is associated with the distribution on a series of human emotions. In addition, a method based on topic modeling is proposed to construct a topic-level dictionary, where each topic is correlated with social emotions. Experiment on the real-world data sets has validated the effectiveness and reliability of the methods. Compared with other lexicons, the dictionary generated using our approach is language-independent, fine-grained, and volume-unlimited. The generated dictionary has a wide range of applications, including predicting the emotional distribution of news articles, identifying social emotions on certain entities and news events.

Keywords Web 2.0 · Social emotion detection · Emotional dictionary · Topic modeling

Y. Rao · Q. Li
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong

Y. Rao
e-mail: raoroland@gmail.com

Q. Li
e-mail: itqli@cityu.edu.hk

J. Lei (✉)
College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China
e-mail: jshlei@shiep.edu.cn

L. Wenyin
College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China
e-mail: liuwenyin@gmail.com

M. Chen
Centre for Transport, Trade and Financial Studies, City University of Hong Kong, Kowloon, Hong Kong
e-mail: mlchen2@cityu.edu.hk

1 Introduction

In the traditional society, when we make a decision, opinions and emotions of others have always been important information for reference. Knowing the answer of "What others think and feel" is necessary for general people, marketers, public relations officials, politicians and managers. For example, people usually like to ask or collect others' opinions, before buying a book, a stock [8] or even a house. Marketers need to get and evaluate opinions of customers, before promoting a new product or consolidating their brand. Politicians, public relations officials and reporters have to track public emotions to current events, advertisers should track advertising campaigns, trend analyzers need to keep pace with technology and entertainment trends, and stock traders have to track emotions from stock market [32].

Recently, with the paradigm shift in the usage of the Web from information consumption to information production and sharing ("Web 2.0"), numerous social media services have emerged. Users can express their opinions and emotions conveniently through news portals, blogs and microblogs (tweeters), where they become both the consumers and producers. Facing the vast amount of data, the task of automatically detecting public emotions evoked by online documents is emerging [7] (such as the SemEval task 14). This task is treated as a classification problem according to the polarity (positive, neutral or negative) or multiple emotion categories such as joy, sadness, anger, fear, disgust and surprise. However, due to the limited information, annotating news headlines for emotions is a hard task. It is usually intractable to annotate headlines consistently even for human beings [13]. In this paper, we mainly focus on (1) annotating news bodies for emotions, and (2) building emotional dictionaries for online news as a result.

An algorithm and three pruning strategies are developed to generate a word-level emotional dictionary. We also build a topic-level lexicon by a method exploiting topic modeling. The constructed emotional dictionaries are evaluated on two data sets. The first one is a large-scale Chinese data set from the Sina society channel.[1] It has 40,897 news articles and 2,083,818 ratings distributed over 8 kinds of social emotions (i.e., touching, empathy, boredom, anger, amusement, sadness, surprise and warmness). The second one is an English data set used in the SemEval task 14. Experimental results show that the generated dictionary performs competitively in terms of social emotion classification. We also conduct qualitative investigation on samples of the emotional dictionary. The result shows that the dictionary can reflect not only explicit emotion words, but also implicit words that convey emotions potentially.

The main contributions are as follows:

- Algorithms of building the word-level and topic-level emotional dictionaries are proposed, which are totally automatic, and no human resource is needed.
- The approach is compared with the state-of-the-art algorithms by the means of social emotion classification. In addition, qualitative investigation is conducted to analyze the generated emotional dictionary.
- Compared with the existing emotional lexicons, the constructed emotional dictionary is language-independent, fine-grained, and can be updated constantly.

The rest of this article is organized as follows. Related work is given in Section 2. The problem formulation, the algorithm of building the emotional dictionary and potential applications of the dictionary are presented in Section 3. The experimental data sets,

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

evaluation metrics, results and discussions are illustrated in Section 4. Finally, we draw conclusions in Section 5.

2 Related work

2.1 Sentiment analysis of reviews

Most of the previous works focus on sentiment analysis of reviews [8, 22, 30, 32], which emerges from the year 2001 or so. Algorithms of both supervised and unsupervised learning have been applied for this task. Das and Chen [8] utilized classification algorithm to extract market emotions from stock message boards, which was further used for decision on whether to buy or sell a stock. However, the performance is heavily dependent on certain words. For instance, the sentence “It is not a bear market” means a bull market actually, because negation words such as “no”, “not” are much more important and serve to reverse meaning. Pang et al. [22] also applied three classification algorithms, Naïve Bayes, Maximum Entropy and Support Vector Machines, to classify movie reviews into positive and negative. They reported that those algorithms do not perform as well on sentiment classification as on text classification. In another early work [32], Turney applied an unsupervised learning algorithm to classify the emotional orientation of users’ reviews (i.e., reviews of movies, travel destinations, automobiles and banks). His approach firstly calculates the mutual information between each phrase and the word “excellent”, as well as the mutual information with the word “poor”. Then, the difference of the two mutual information scores is used to classify each review as “recommended” or not.

During this incipient stage of research on sentiment analysis of reviews, some of them focus on emotion categorization of the entire documents, which are based on the construction of discriminate-word dictionaries manually or semi-manually [8]. Other works focus on using linguistic heuristics or a set of preselected seed words, to classify the emotional orientation of words or phrases [31]. In addition to the above works on binary sentiment classification, there are also several works focusing on predicting the rating scores of reviews [11, 18]. As the rating scores are ordinal (e.g., 1–5 stars), the problem is tackled by regression. For sentiment analysis of reviews, as review holders are usually anonymous, useful information is often mixed with noisy data that makes sentiment analysis more difficult [20]. There are, for example, malicious users expressing offensive opinions, using their comments for the purpose of advertising, or even spreading rumors and fraudulent reviews. Considering this issue, Opinion Spam Detection [12] is essential to detect and filter out irrelevant information in reviews, which is an important subtask when performing sentiment analysis.

2.2 News emotion classification

Works of emotion classification of online news began from the SemEval tasks in 2007. The SemEval introduced the task of “Affective Text” [28] to explore the connection between news headlines and the evoked emotions of readers. The underlying assumption is that all words, even those neutral ones, can potentially convey affective meaning and provoke audience pleasant or painful experiences. For the task of news emotion classification, three systems were introduced: SWAT, UA and UPAR7.

The system SWAT [13, 28] adopted a supervised approach by developing a word-emotion mapping dictionary. The dictionary was then used to score each word of a headline to have an average score for the headline and to decide its emotion. UA [28] gathered statistics of the news headlines and emotions from three search engines and computed the Pointwise Mutual Information (PMI) scores to determine the emotion labels of headlines. UPAR7 [28] was a rule-based system which particularly relies on syntactic parser and lexicons. A similar piece of research work by Lin et al. [16, 17] also studied readers' emotional states evoked by Yahoo! Chinese news articles. In particular, a number of relevant features were extracted from the news articles and SVM was used to classify each article to appropriate emotion category. Due to the limited words in the headlines, those systems faced the problem of the small number of words available for the analysis.

Recently, Bao et al. [3, 4] proposed an emotion-topic model for social emotion classification by introducing an intermediate layer into Latent Dirichlet Allocation (LDA) [5]. The generation process of an affective document was modeled by the following steps. Firstly, for each document, a distribution over emotions is generated from a Multinomial distribution. Secondly, for each word in the document, a single emotion label is sampled according to above distribution. Thirdly, a latent topic is generated from a Dirichlet distribution conditioned on the emotion, and finally a term is generated from the latent topic, which is modeled by another Multinomial distribution over words. Their model is homologous to the author-topic model for authors and documents [23], which extends LDA to include authorship information by assuming that each author is associated with a Dirichlet distribution over topics and each topic is associated with a Multinomial distribution over words. Experimental results in [3, 4] show the emotion-topic model outperforms SVM and several other methods for social emotion detection task. However, those works focus on emotion classification, which is only one of the applications of emotional dictionaries.

2.3 Emotional dictionary construction

To construct an emotional dictionary, many pre-developed lexicons may be used [7, 15], e.g., Subjectivity Wordlist [2], WordNet-Affect [29] and SentiWordNet [1]. The Subjectivity Wordlist is built by a manually selected seed set of subjective words, a small raw corpus, and an online dictionary. It is a subjectivity lexicon distinguishing subjective versus objective words, which is useful in the pre-processing of reviews. The WordNet-Affect is a linguistic resource, in which the synsets representing emotional concepts are labeled. Synset is the synonym set that represents a sense or a concept in the WordNet lexicon. 2,874 synsets and 4,787 words are annotated in the WordNet-Affect. The SentiWordNet is a lexical resource developed for supporting emotion classification. It scores each synset in the WordNet along three emotional dimensions: positivity, negativity and neutrality. Those pre-developed emotional dictionaries can be applied to emotion classification directly. Chaumartin [7] utilized a linguistic and rule-based approach to tag news headlines for predefined emotions, which includes joy, sadness, anger, fear, disgust and surprise, and for polarity, i.e. positive or negative. The algorithm was based on existing emotional dictionaries, like WordNet-Affect and SentiWordNet. Kolya et al. [15] identified event and emotional expressions at word level from the sentences of TempEval-2010 corpus, in which the emotional expressions are also identified simply based on the sentiment lexicons, e.g., Subjectivity Wordlist, WordNet-Affect and SentiWordNet.

Emotion classification based on those existing lexicons has their limited utility. Firstly, the lexicons are mainly for public use in general domains, some resulting classifications of words can appear incorrect, and need to be adjusted to fit the personalized data set. Secondly, most of the lexicons are available only for bits of languages, such as English, and the volume of words annotated is restricted, which limits the applicability of those methods. Lastly, many of the existing lexicons label words on coarse-grained dimensions (positivity, negativity and neutrality), which are insufficient to individuate the whole spectrum of emotional concepts [29].

In this paper, we mainly focus on annotating news bodies for emotions, and building emotional dictionaries for online news. The emotion expressions are fine-grained (e.g., touching, empathy, boredom, anger and funny), rather than coarse-grained (positive, negative and neutral). The dictionary can be used to classify the emotional distributions of previous unseen news articles, and identify entities (e.g., product, brand and city) or news events that evoke different social emotions.

3 Emotional dictionary construction and applications

In this section, we firstly define our research problem. Then, we introduce the generating algorithm and pruning strategies of the word-level emotional dictionary, as well as the method of building the topic-level emotional dictionary. Finally, we discuss the potential applications of the dictionaries.

3.1 Problem formulation

The research problem of constructing the emotional dictionary is defined as follows.

Given N training news articles, a word-level and a topic-level emotional dictionaries are generated. The word-level dictionary is a $W\times E$ matrix, where W is the number of unique words, E is the number of emotions, and the (j, k) item in this matrix is the score (probability) of emotion e_k conditioned on word w_j. The topic-level dictionary is a $K\times E$ matrix, where K is the number of latent topics, and the (m, k) item in this matrix is the probability of emotion e_k conditioned on topic z_m.

For each document d_i (i=1, 2, …, N), the news content, the publication date (timestamp), and the distribution of ratings of emotions in the predefined list are available. Figure 1 presents an example from a popular news portal in China (i.e., Sina.com.cn). The predefined emotions in this website are touching, empathy, boredom, anger, amusement, sadness, surprise and warmness. Those emotions are voted by 3,064 online users for a particular news article. From all the contents of N news articles, a vocabulary is obtained as the source of the word-level emotional dictionary. The j-th word in the vocabulary is denoted by w_j (j=1, 2, …, W). A low-rank space generated by topic modeling is used as the

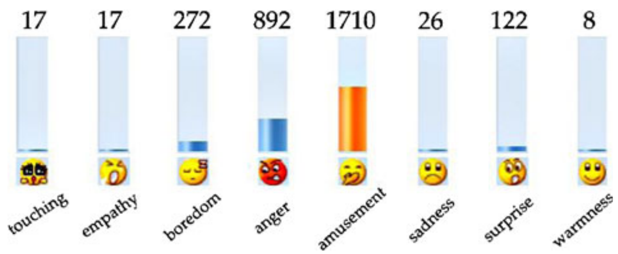

Figure 1 An example of social emotions and user ratings

source of the topic-level emotional dictionary, and the m-th topic is denoted by z_m (m=1, 2, …, K). The user-defined emotions is denoted by e=(e_1, e_2, …, e_E), the normalization form of ratings of d_i over e is denoted by r_i. r_i=(r_{i1}, r_{i2}, …, r_{iE}), and $|r_i|$=1.

3.2 Constructing word level emotional dictionary

In this section, we present the algorithm of generating the word-level emotional dictionary, as well as the pruning strategies of the constructed dictionary.

3.2.1 Generating algorithm

We start with introducing the method of generating emotional dictionary based on maximum likelihood estimation and Jensen's inequality. The notation of frequently-used variables is summarized in Table 1.

For each document d_i, the probability of r_i conditioned on d_i can be modeled as:

$$P(r_i|d_i) = \sum_{j=1}^{W} P\left(w_j|d_i\right)P\left(r_i\middle|w_j\right). \tag{1}$$

In the above, the probability of r_i conditioned on w_j is a multinomial distribution, and $P\left(r_i\middle|w_j\right) = \prod_{k=1}^{E} P\left(e_k\middle|w_j\right)^{r_{ik}}$. Then,

$$P(r_i|d_i) = \sum_{j=1}^{W} P\left(w_j|d_i\right) \prod_{k=1}^{E} P\left(e_k\middle|w_j\right)^{r_{ik}}. \tag{2}$$

In the above, words in document d_i are assumed to be independent of each other.

Since ε_i=$P(d_i)$, $\sigma_{ij} = P\left(w_j|d_i\right)$ and $\theta_{jk} = P\left(e_k\middle|w_j\right)$, the log-likelihood of $P(d_i, r_i)$ over all the N documents can be defined as:

$$\log l = \log\left(\prod_{i=1}^{N}\left(\varepsilon_i \sum_{j=1}^{W} \sigma_{ij} \prod_{k=1}^{E} \theta_{jk}^{r_{ik}}\right)\right) = \sum_{i=1}^{N} \log\left(\varepsilon_i \sum_{j=1}^{W} \sigma_{ij} \prod_{k=1}^{E} \theta_{jk}^{r_{ik}}\right). \tag{3}$$

Table 1 Notation of frequently-used variables

Symbol	Description
N	number of training documents
W	number of unique words
E	number of emotion labels
d_i	the i-th document in the training set
ε_i	the prior probability of i-th document
r_i	the normalized ratings of i-th document over E emotions
w_j	the j-th word in the dictionary
e_k	the k-th emotion in the dictionary
σ_{ij}	the probability of j-th word conditioned on i-th document
θ_{jk}	the probability of k-th emotion conditioned on j-th word
r_{ik}	the normalized rating of i-th document over k-th emotion

According to Jensen's inequality, we reconstruct the log-likelihood as follows:

$$\log l \geq \sum_{i=1}^{N}\sum_{j=1}^{W}\varepsilon_i\sigma_{ij}\sum_{k=1}^{E}r_{ik}\log\theta_{jk}. \tag{4}$$

Since $\sum_{k=1}^{E}\theta_{jk}=1$, we add a Lagrange multiplier to the log-likelihood equation as follows:

$$\widehat{l}=\sum_{i=1}^{N}\sum_{j=1}^{W}\varepsilon_i\sigma_{ij}\sum_{k=1}^{E}r_{ik}\log\theta_{jk}+\lambda\left(\sum_{k=1}^{E}\theta_{jk}-1\right). \tag{5}$$

Then, we maximize the likelihood by calculating the first-order partial derivative of θ_{jk},

$$\frac{\partial\widehat{l}}{\partial\theta_{jk}}=\sum_{i=1}^{N}\frac{\varepsilon_i\sigma_{ij}r_{ik}}{\theta_{jk}}+\lambda=\frac{\sum_{i=1}^{N}\varepsilon_i\sigma_{ij}r_{ik}}{\theta_{jk}}+\lambda=0. \tag{6}$$

Thus,

$$\theta_{jk}=-\frac{\sum_{i=1}^{N}\varepsilon_i\sigma_{ij}r_{ik}}{\lambda}. \tag{7}$$

Since $\sum_{k=1}^{E}\theta_{jk}=1$, we have

$$\lambda=-\sum_{k=1}^{E}\sum_{i=1}^{N}\varepsilon_i\sigma_{ij}r_{ik}. \tag{8}$$

Then, by substituting Eq. (8) into Eq. (7) we get:

$$\theta_{jk}=\frac{\sum_{i=1}^{N}\varepsilon_i\sigma_{ij}r_{ik}}{\sum_{k=1}^{E}\sum_{i=1}^{N}\varepsilon_i\sigma_{ij}r_{ik}}. \tag{9}$$

In the above, θ_{jk} is the probability of emotion e_k conditioned on word w_j, from which we could generate the word-level emotional dictionary. r_{ik} is the normalized rating of document d_i over emotion e_k. σ_{ij} is the probability of word w_j conditioned on document d_i, which can be calculated by relative term frequency. The relative term frequency is the number of occurrences of the term w_j in d_i divide by the total number of occurrences of all the terms in d_i. ε_i is the prior probability of document d_i, which can be calculated by maximum likelihood estimation as follows:

$$\varepsilon_i=\sum_{k=1}^{E}P(d_i,e_k)=\sum_{k=1}^{E}\left(r_{ik}\Big/\sum_{i=1}^{N}r_{ik}\right). \tag{10}$$

The computation of all ε_i can be conducted in the preprocessing, so the time complexity of estimating each θ_{jk} is $O(N)$, where N is the number of training documents. The algorithm can be applied to an online system by using time window. The size of a time window specifies the number of documents to be considered when doing training. For each time window, the time complexity is the same with that in the offline situation described above. Considering when the size of the training data set increases, the scale/dimension of the dictionary extends, refining operation is necessary for such dictionaries.

3.2.2 Refinement by pruning

Pruning is a process that removes a subset from the original word-level emotional dictionary according to some criteria. Three strategies of pruning are used to refine the dictionary, each of which uses a different criterion to eliminate a desired degree of words.

As stated earlier, the emotional dictionary is generated from the real-world news corpus, thus there contains topical common words. For example, consider a news corpus reporting by Google news. Words like "Google", "news", "report", and reporters' names in Google news are frequently occur and merged with other words in the emotional dictionary. Those words convey little social emotions and may disturb the effect of utilizing the dictionary. We prune those topical common words by a simple algorithm as follows:

ALGORITHM 1. The Pruning Algorithm.

Input: The emotional dictionary *OD* generated by Eq. (9), pruning strategy (i.e., maximum, average, minimum)
Output: The refined emotional dictionary *RD*.

```
zero word document frequency: df_w;
for each document d in N training articles do
    for all words w in d do
        if w not exists in d before then
            df_w +1;
        end if
    end for
end for
for each word w in OD do
    get θ, the probabilities of E emotions conditioned on w;
    calculate the value λ_w of θ according to the pruning strategy;
    if λ_w ≥ df_w/N then
        add w into RD;
    end if
end for
```

The algorithm consists of two processes. Firstly, it calculates the document frequency of each word in the emotional dictionary, which can be conducted in the preprocessing. Secondly, it compares the relative document frequency with threshold λ_w for each word w, to determine whether w is added into the refined dictionary. As a result, the time complexity is $O(W)$, where W is the number of unique words. The threshold λ_w is calculated by the pruning strategy and probabilities of E emotions conditioned on w, which are from the original emotional dictionary *OD*. Assume that a word "report" occurs in 5 documents, the total number of training articles N is 10, the specified pruning strategy is "minimum", and the probabilities of 8 emotions conditioned on "report" are 0.1, 0.1, 0.2, 0.1, 0.1, 0.1, 0.1, 0.2. Then, the relative document frequency of "report" is 0.5, and the threshold λ is 0.1. The word "report" is thus considered to be a topical common word and pruned from *OD*.

3.3 Building topic level emotional dictionary

Works of model-based Feature/Dimension Reduction actually began from the method of Latent Semantic Indexing (LSI). LSI is based on Singular Value Decomposition (SVD) of the feature-by-document matrix of a corpus, which projects the original matrix to lower rank space, so as to cope with synonymous and to reduce the feature dimension. Recently LSI models are replaced by Probabilistic Latent Semantic Indexing (PLSI), which is one of the topic models and has a solid statistical foundation [10].

The core of PLSI is a latent variable model which associates a latent topic variable $z=\{z_1, \ldots, z_K\}$ with each occurrence of a word $w=\{w_1, \ldots, w_W\}$. Normally, the number of topics K is much smaller than the number of unique words W, and PLSI variables are estimated by Expectation Maximization (EM) algorithm. Latent Dirichlet Allocation (LDA) is another topic model [6, 14]. The LDA parameters are estimated by the approximate inference algorithms, such as variational EM and Gibbs Sampling [5, 9]. Both PLSI and LDA estimate latent variables by hypothesizing that a document is constructed by words which are generated based on several topics, and they can be applied to construct the topic-level emotional dictionary. We use LDA model as a demonstration in this paper. In the LDA model, latent topics that generate each document d are sampled from a topic distribution, and such a distribution is denoted as $p(z|d,\theta)$. The topics generated from a random set of words w are sampled from each topic's word distribution, and such a distribution is denoted as $p(w|z,\beta)$. For example, consider using themes to explore a news corpus about technology. At a broad level (i.e., small number of topics) the themes might correspond to Internet and Telecom. We could increase the number of topics to zoom in on a broad theme, such as Telecom, to reveal various aspects of it, e.g., the 3rd Generation Telecommunication (3G), mobile communication and mobile phone. Those themes (i.e., latent topics) are connected with each document by $p(z|d,\theta)$, and with each word by $p(w|z,\beta)$. They are further connected with social emotions here so that we could reveal public opinion on certain aspects.

For each document d_i, the probability of r_i conditioned on d_i is modeled as:

$$P(r_i|d_i) = \sum_{m=1}^{K} P(z_m|d_i) P(r_i|z_m). \tag{11}$$

Let $\delta_{im} = P(z_m|d_i)$ and $\eta_{mk} = P(e_k|z_m)$, the log-likelihood of $P(d_i, r_i)$ over all the N documents is defined as:

$$\log l = \log\left(\prod_{i=1}^{N}\left(\varepsilon_i \sum_{m=1}^{K} \delta_{im} \prod_{k=1}^{E} \eta_{mk}^{r_{ik}}\right)\right) = \sum_{i=1}^{N} \log\left(\varepsilon_i \sum_{m=1}^{K} \delta_{im} \prod_{k=1}^{E} \eta_{mk}^{r_{ik}}\right). \tag{12}$$

Then, by following the same way of deduction in Section 3.2.1, we generate the topic-level emotional dictionary as follows:

$$\eta_{mk} = \frac{\sum_{i=1}^{N} \varepsilon_i P(z_m|d_i) r_{ik}}{\sum_{k=1}^{E}\sum_{i=1}^{N} \varepsilon_i P(z_m|d_i) r_{ik}}. \tag{13}$$

In the above, η_{mk} is the probability of emotion e_k conditioned on topic z_m, from which we could generate the topic-level emotional dictionary. $P(z_m|d_i)$ is the probability of topic z_m

conditioned on document d_i. Compared with the word-level emotional dictionary constructed by Eq. (9), the emotional dictionary exploiting LDA is topic level. Each topic in the dictionary is associated with words by the probability of word w_j conditioned on topic z_m, i.e., $P(w_j|z_m)$. According to the LDA model using Gibbs Sampling, the time complexity for each iteration is $O(VK)$, where V is the total number of words, K is the number of topics. $P(z_m|d_i)$ and $P(w_j|z_m)$ can be estimated by the following formulas:

$$p(z_m|d_i) = \frac{n_{z_m}^{d_i} + \alpha}{\sum_{m=1}^{K} \left(n_{z_m}^{d_i} + \alpha\right)}, \tag{14}$$

$$p\left(w_j|z_m\right) = \frac{n_{w_j}^{z_m} + \beta}{\sum_{j=1}^{W} \left(n_{w_j}^{z_m} + \beta\right)}. \tag{15}$$

In the above, $n_{z_m}^{d_i}$ is the number of words in d_i assigned to topic z_m, $n_{w_j}^{z_m}$ is the number of instances of word w_j assigned to topic z_m, α and β are hyper parameters. Typical values of the number of topics K lie in the range of 100 to 300 for the English datasets [24]. We will discuss the influence of K in detail in Section 4.3.2. Similar to the previous works [3, 4, 9], the hyper parameters α and β are set to symmetric Dirichlet priors with values of $50/K$ and 0.1, respectively.

3.4 Emotional dictionary applications

After generating the emotional dictionary, classifying and predicting the social emotions of given news articles are straightforward. For the word-level emotional dictionary generated by Eq. (9), it can be used to predict the emotions of given news articles as follows:

$$\widehat{P}(e|d) = \sum_{w \in W} p(w|d)p(e|w). \tag{16}$$

In the above, $\widehat{P}(e|d)$ is the predicted probability of users having emotions e on a new document d, $P(w|d)$ is the distribution of document d on word w, which can be calculated by relative term frequency, $P(e|w)$ is the probability of emotions e conditioned on word w, which can be looked up from the emotional dictionary generated by Eq. (9).

For the topic-level emotional dictionary generated by Eq. (13), it can be used to predict the emotions of given news articles as follows:

$$\widehat{P}(e|d) = \sum_{z \in Z} p(z|d)p(e|z). \tag{17}$$

In the above, $P(z|d)$ is the distribution of a new document d on topic z, which can be calculated by Eq. (14), $P(e|z)$ is the probability of emotions e conditioned on topic z, which can be looked up from the emotional dictionary generated by Eq. (13).

The constructed emotional dictionaries can also be used to identify entities or news events that evoke certain social emotions.

4 Experiments

To test the effect of the emotional dictionary on social sentiment analysis, experiments are conducted on a Chinese data set and an English data set. The good performance and multilingual data sets reflect the method's effectiveness, reliability, and language-independent of building the emotional dictionary.

4.1 Data sets

To test the adaptiveness, effectiveness and language-independent of our method of building the emotional dictionary, large-scale and multilingual data sets are needed. Two kinds of data sets are employed in the experiment.

Sina This is a large-scale Chinese data set scrawled from Sina society, which is one of the most popular news sites in China.[2] The attributes include the URL address of the news article, the news headline (title), the publish date (from August 2009 to April 2012), the news body (content), the user ratings over emotions of touching, empathy, boredom, anger, amusement, sadness, surprise and warmness. The data set contains 40,897 valid news articles with the total number of ratings over the 8 emotions larger than 0.

SemEval This is an English data set used in the 14th task of the 4th International Workshop on Semantic Evaluations (SemEval-2007).[3] The attributes include the news headline, the user scores over emotions of anger, disgust, fear, joy, sad and surprise, which are normalized from 0 to 100. The data set contains 1,246 valid news headlines with the total score of the 6 emotions larger than 0.

The information of the two data sets is summarized in Table 2. The number of articles of each emotion label represents the sum of documents having the most ratings over that emotion. For the Sina data set, the largest category is "anger". There are 18,287 news articles having the most user ratings over that emotion. The smallest category is "warmness", which only has 388 documents. The category distribution is quite imbalanced and poses challenges to the method of building the emotional dictionary.

4.2 Experiment design

For given news articles, classifying and predicting the emotions are efficient ways to validate the effectiveness of the generated emotional dictionary. The following algorithms are implemented for comparison:

1. Emotion-Term method (ET). The emotion-term method was formulated by improving the Naive Bayes classifier [3, 4]. Different from traditional Naïve Bayes, the method takes into account emotional ratings when calculating the probability of a category and the probability of a term given an emotion label.
2. Emotion-Topic Model (ETM). The emotion-topic model was built by introducing an additional layer (i.e. latent topic) into ET and LDA. The parameters of ETM are set according to the description in [3, 4].

[2] The data set is publicly available at http://www.hkws.org/public-sources/sinanews.zip.
[3] The data set is publicly available at http://www.cse.unt.edu/~rada/affectivetext/.

Table 2 Statistics of the data sets

Data set	Emotional label	# of articles	# of ratings
Sina	touching	5,672	254,488
	empathy	2,676	149,883
	boredom	2,603	151,599
	anger	18,287	830,927
	amusement	6,812	320,824
	sadness	2,888	212,316
	surprise	1,571	105,585
	warmness	388	58,196
SemEval	anger	87	12,042
	disgust	42	7,634
	fear	194	20,306
	joy	441	23,613
	sad	265	24,039
	surprise	217	21,495

3. SWAT system (SWAT). The SWAT system was one of the top-performing systems on the SemEval Affective Text task [25]. It scored the emotions of each word w as the average of emotions of every news headline, in which w appears [13, 28].
4. Word-Emotion method (WE). This is the proposed method based on building the word-level emotional dictionary. We firstly construct the dictionary according to Eq. (9). Then, the probabilities of social emotions are predicted by Eq. (16). The methods combined three pruning strategies in Algorithm 1 (i.e., maximum, average and minimum) are denoted by WE-max, WE-ave and WE-min, respectively.
5. Emotion LDA model (ELDA). This is the proposed algorithm of jointly modeling emotions and topics by LDA. We firstly build the topic-level dictionary according to Eq. (13). Then, the probabilities of social emotions are predicted by Eq. (17).

As mentioned above, the social emotions are detected by predicting the probabilities of them conditioned on each article. The larger the conditioned probability that an emotion has, the higher probability readers of the news article is likely to arouse the corresponding feeling. To evaluate the proposed algorithms, we compare the predicted probabilities with the actual distributions of emotions. The coarse-grained and fine-grained evaluation metrics are employed as indicators of the performance [21, 27]. For the coarse-grained evaluations, each emotion is mapped to a 0/1 classification, and we use the metric *Acc@k*. The fine-grained evaluations are based on Pearson's correlation coefficient, and the metric *AP* is calculated for all documents.

Acc@k, or the accuracy at top k is used in [3, 4]. Given a news article d, the top ranked predicted emotion e_p, and the truth emotion set $E_{topk@d}$ including the k top-ranked emotions, $Acc_d@k$ is calculated as

$$Acc_d@k = \begin{cases} 1 & if \quad e_p \in E_{topk@d} \\ 0 & else. \end{cases} \tag{18}$$

Then, the *Acc@k* for the entire collection D is

$$Acc@k = \sum_{d \in D} Acc_d@k/|D|, \tag{19}$$

where $|D|$ is the size of D, and the parameter k is usually set to 1. According to [3, 4], $Acc@1$ is considered to be the most important metric, which is the same as micro-averaged F_1 [19]. The F_1 measure equally weights precision and recall, and micro-averaging is one of the methods to compute a single aggregate measure when processing a collection with several two-class classifiers [19]. Micro-averaging pools per-document decisions across categories, and then computes an effectiveness measure on the pooled contingency table. Due to the very imbalanced distribution of articles in categories (ref. Table 2), it is unnecessary to compute the F_1 measure of each category or a macro-averaged F_1 [19] which takes the average of F_1 for all categories.

AP stands for the averaged Pearson's correlation coefficient. Pearson's correlation coefficient was also used as a measure in sentiment analysis of news [7, 13, 21, 27]. Given a news article d, the predicted and truth probability of users having emotions e on document d (i.e., X and Y), P_d is calculated as

$$P_d = \frac{\sum_{i=1}^{E} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(E-1)\sigma_X \sigma_Y}, \tag{20}$$

where X_i and Y_i are respectively the i-th element of X and Y, $\overline{X}$ and $\overline{Y}$ represent their means, and σ_X and σ_Y are the standard deviations. Then, the averaged value of all P_d [21] is used as *AP*, which has the range of [−1, 1], where 1 means a perfect positive correlation.

4.3 Results and analysis

4.3.1 Emotional dictionary samples

In this subsection, we present two kinds of emotional dictionary samples: the probability of emotions conditioned on a word and on a latent topic.[4] These samples are also compared with one of the manually developed lexicons [1], so as to show the differences between them.

Given the two data sets (i.e., Sina and SemEval), the proposed Word-Emotion method (WE) is used to generate the word-level dictionary, in which the probability of social emotions conditioned on each word is calculated by Eq. (9). Although the existing SWAT system (SWAT) can also generate such dictionaries, it was only proposed for news headlines [13, 28]. Table 3 lists the representative words of the word-level emotional dictionaries generated on Sina and SemEval, in which the probability of social emotions conditioned on each word is shown in parenthesis. The positive score (PosScore) and negative score (NegScore) of each word are also searched from the SentiWordNet lexicon [1], which is developed manually. If a certain word is not included in the SentiWordNet yet, then we note it as "NA".

The result shows that the constructed word-level dictionary by our method indentifies not only explicit emotional words, but also implicit words that convey social emotions potentially. However, explicit emotional words are mainly annotated by SentiWordNet, e.g., "touched", "rescue", "animal" and "fault". Moreover, as the SentiWordNet lexicon is coarse-grained, it is insufficient to individuate the whole spectrum of emotional concepts. The implicit emotional words identified by our dictionary reflect entities evoking certain social emotions. For instance, the word "Tao Xingzhi" (the name of a famous educationist in China) is assigned a large score with respect to the emotion of "warmness", as he has made

[4] The samples are publicly available at http://www.hkws.org/public-sources/lexicon.zip.

Table 3 The representative words (All the words in Sina are translated from Chinese to English)

Data set	Word	The word level dictionary	The SentiWordNet lexicon
Sina	touched	touching (0.57), empathy (0.08), sadness (0.08)	PosScore (0), NegScore (0. 5)
	rescue	touching (0.37), sadness (0.15), empathy (0.13)	PosScore (0.125), NegScore (0)
	pollutant	anger (0.58), sadness (0.18), empathy (0.07)	PosScore (0), NegScore (0)
	animal	surprise (0.22), amusement (0.17), anger (0.15)	PosScore (0.125), NegScore (0)
	Tao Xingzhi	warmness (0.90), empathy (0.09), touching (0.01)	NA
SemEval	fault	sad (0.50), fear (0.31), anger (0.11)	PosScore (0), NegScore (0.625)
	unveil	surprise (0.62), joy (0.38)	PosScore (0), NegScore (0)
	London	joy (0.96), surprise (0.04)	PosScore (0), NegScore (0)
	Madonna	joy (0.63), surprise (0.35), sad (0.01)	PosScore (0), NegScore (0)
	NASA	surprise (0.52), joy (0.48)	NA

great efforts to the national education. The news titled "Game on! London exhibition celebrates the history of video games" mainly triggers the large score of the emotion of "joy" over the word "London". The large scores of the emotions of "joy" and "surprise" over the word "Madonna" (the name of a pop singer in United States) are trigged by the news titled "Madonna's New Tot 'Happy at Home' in London" and "Madonna's new baby's daddy didn't realize adoption was 'for good'", respectively. The word "NASA" (i.e., the National Aeronautics and Space Administration) is assigned a large score with respect to the emotions of "surprise" and "joy", for the news titled "NASA revisiting life on Mars question" and "NASA spacecraft to measure Sun".

We also generate the topic-level dictionary by the proposed Emotion LDA model (ELDA). The probability of social emotions conditioned on each topic is calculated by Eq. (13), and the probability of words conditioned on each topic is estimated by Eq. (15). Since ELDA is mainly developed for news bodies, we list the representative topics of the topic-level dictionary on Sina in Table 4. There are two kinds of topics identified by the topic-level emotional dictionary. The first one is topics focus on certain emotions, such as topic 13, 212 and 219. Topic 13 evokes the emotion of "touching" mainly, and the representative words are "help", "touched" and "hope". Topic 212 and 219 are related to several similar emotions. The second one is topics with merged emotions, such as topic 126 and 375. For instance, many news documents on "dog breeding" evoke the emotions of "surprise" and "amusement", while others trigger the emotion of "anger".

Although the existing Emotion-Term method (ET) and Emotion-Topic Model (ETM) can respectively connect social emotions with words and with latent topics, they estimate different probabilities compared with the SentiWordNet, SWAT, WE and ELDA. According

Table 4 The representative topics (All the words are translated from Chinese to English)

Topic ID	Top 3 words in each topic	The topic level dictionary
13	help (0.004), touched (0.004), hope (0.003)	touching (0.47), sadness (0.11), empathy (0.11)
126	dog (0.015), animal (0.015), breeding (0.010)	surprise (0.21), amusement (0.18), anger (0.15)
212	death (0.012), rescue (0.009), die (0.009)	sadness (0.27), anger (0.26), empathy (0.14)
219	hospital (0.011), doctor (0.010), cure (0.010)	touching (0.20), empathy (0.18), sadness (0.16)
375	university (0.020), student (0.015), school (0.012)	touching (0.19), anger (0.18), amusement (0.15)

to [3, 4], the probability of words and topics conditioned on each social emotion, i.e., $P(w|e)$ and $P(z|e)$ are calculated by ET and ETM, respectively. As a result, ET and ETM can not generate emotional dictionaries such as the ones illustrated in Tables 3 and 4. The topic number of ELDA is set to 400 in Table 4. Those samples in Tables 3 and 4 are generated on 90 % of the data set. Next, we further evaluate the performance of algorithms with different parameters in terms of social emotion classification. Firstly, the influence of topic number for ELDA and ETM is studied. Then, all algorithms are compared and the effect of dictionary pruning is analyzed. The evaluation is conducted on the large-scale Sina data set, since all algorithms are mainly designed for news bodies except SWAT. Finally, we examine the performance of algorithms on short documents (i.e., SemEval data set).

4.3.2 Influence of topic number

The number of topics indicates how many latent aspects of documents can be derived, which is a parameter of algorithms exploiting LDA, i.e., the proposed Emotion LDA model (ELDA) and the existing Emotion-Topic Model (ETM) [3, 4]. To evaluate the influence of topic number, we use ELDA and ETM to classify the same testing set (90 %) based on the same training set (10 %) from Sina, and vary the number of topics from 20 to 400. The other parameters are set according to the description in [3, 4]. Figure 2 presents the performance of ELDA and ETM with different topic numbers. The results show that the topic number has a different impact on ELDA and ETM.

Firstly, the performance of ELDA increases largely as the increase of topic numbers in terms of both metrics. ELDA achieves the best result for 400 topics, where the *Acc@*1 and *AP* are 55.5 % and 0.55, respectively. It is consistent with the nature of ELDA, i.e., works as a model-based feature reduction method essentially. The larger the topic number is, the more aspects are revealed by ELDA. However, the performance of ETM increases slightly as the increase of topic numbers, and the increment is unstable. ETM achieves the best result for 300 topics, where the *Acc@*1 and *AP* are 55.5 % and 0.50, respectively. The reason is that ETM is essentially a feature-based classifier. The latent topics are mainly used to estimate and smooth the probability of words conditioned on social emotions in ETM.

Secondly, ELDA yields competitive results with ETM in terms of *Acc@*1, while for all topic numbers, ELDA outperforms ETM in terms of *AP*. To evaluate the differences of the performance statistically, statistical significance test is also performed on paired algorithms, in which the conventional significance level is 0.05. The average *Acc@*1 of ELDA and ETM are 53.3 % and 54.4 %, and the differences are not statistically significant with p-value= 0.12. The averge *AP* of them are 0.52 and 0.47, and the differences are statistically significant with p-value=0.0002. The ideal value of *AP* is 1, which means the predicted distribution of all emotions is completely-correct for all documents. This indicates that

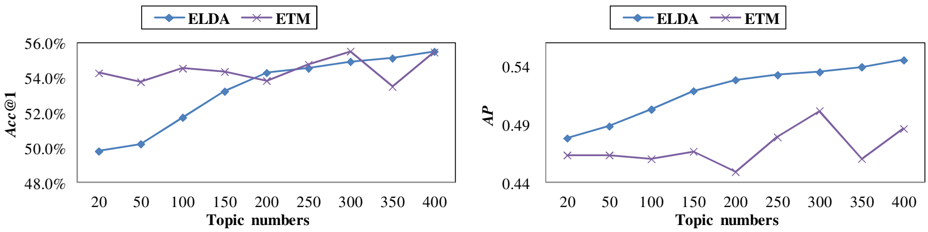

Figure 2 The *Acc@*1 and *AP* of ELDA and ETM with different topic numbers

ELDA is better at predicting a correct distribution of all emotions than ETM according to *AP*. Both of them show competitive performance at predicting the most popular emotion according to *Acc*@1.

4.3.3 Comparison with baselines

The proposed Word-Emotion method (WE), pruning strategies together with WE (i.e., WE-max, WE-ave and WE-min), and Emotion LDA model (ELDA) are compared with three baselines, by using *x* of the corpus for training and the remaining (1-*x*) for testing. To evaluate the scalability and generalizability of all algorithms, the value of proportion *x* varies from 10 % to 90 %, and one of the baselines designed for news headlines (i.e., SWAT) is also used on news bodies here. According to the experimental results in Section 4.3.2, the number of topics is set to 400 and 300 for ELDA and ETM, respectively.

The performance of all algorithms is plotted in Figure 3. As the increase of documents utilized for training, the quality of emotion classification increases also. The results of baselines ET and SWAT are worse than those of our algorithms for both metrics. The other baseline ETM performs competitively in terms of *Acc*@1, while not in terms of *AP*.

We present the average *Acc*@1 and *AP* of different algorithms in Table 5. Compared with the baselines ETM, SWAT and ET, the proposed algorithm of building the topic-level emotional dictionary (i.e., ELDA) improves 2.5 %, 14.8 %, 18.4 % for *Acc*@1, and 12.9 %, 17.2 %, 33.2 % for *AP*; the proposed algorithm of generating the word-level emotional dictionary together with pruning by minimum (i.e., WE-min) improves 4.6 %, 17.2 %, 20.8 % for *Acc*@1, and 13.0 %, 17.3 %, 33.4 % for *AP*. To evaluate the differences of the performance statistically, we also perform statistical significance test on paired algorithms. In terms of *Acc*@1, there are mainly four observations. Firstly, the proposed ELDA and WE-min outperform three baselines (i.e., ETM, SWAT and ET) significantly with *p*-values<0.05. Secondly, the differences between WE-ave and ETM, as well as WE-max and ETM are not statistically significant. Thirdly, the baseline ETM outperfoms WE significantly. Lastly, the proposed WE-ave, WE-max and WE all outperfom two baselines, i.e., SWAT and ET significantly. In terms of *AP*, the proposed five algorithms (i.e., ELDA, WE-min, WE-ave, WE-max and WE) all outperform three baselines (i.e., ETM, SWAT and ET) significantly with *p*-values<0.05.

4.3.4 Effect of dictionary pruning

Pruning strategies are conducted to refine the word-level emotional dictionary generated by WE (ref Algorithm 1). As shown in Table 5, WE-min outperfoms others and the differences

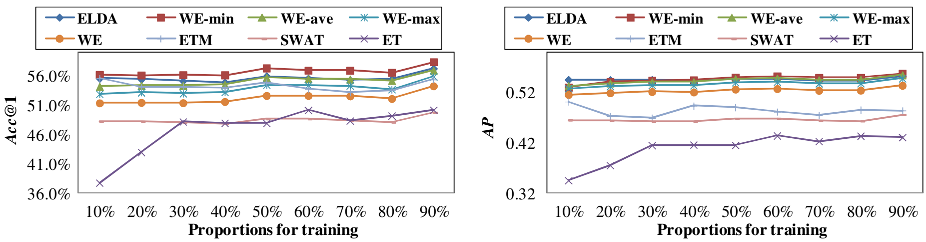

Figure 3 The *Acc*@1 and *AP* of algorithms with different proportions of corpus for training

Table 5 The average *Acc@*1 and *AP* of different algorithms

Algorithms	Average Acc@1	Average AP
ELDA	55.6 %	0.5457
WE-min	56.7 %	0.5465
WE-ave	55.1 %	0.5437
WE-max	53.9 %	0.5369
WE	52.2 %	0.5235
ETM	54.2 %	0.4836
SWAT	48.4 %	0.4658
ET	46.9 %	0.4097

are statistically significant; meanwhile, it prunes nearly 60 % words from the original dictionary. To evaluate the reliability of the strategy, we firstly refine the original dictionary by randomly pruning the same percentage of words as WE-min does, and denote the algorithm as WE-minrand. The average *Acc@*1 and *AP* of WE-minrand are 52.3 % and 0.5214, respectively. Table 6 illustrates the *p*-values of statistical significance test on paired algorithms. On one hand, for the algorithm without pruning (i.e., WE) and the algorithm of pruning randomly (i.e., WE-minrand), the *p*-value is much larger than the conventional significance level 0.05 in terms of *Acc@*1, and smaller than 0.05 in terms of *AP*, which indicates the dictionary after pruning randomly is not significantly different from or even worse than the dictionary without pruning. On the other hand, for the algorithm of pruning by minimum (i.e., WE-min) and WE, or WE-min and WE-minrand, all the *p*-values are much smaller than the conventional significance level 0.05, which indicates that the proposed WE-min achieves significant performance improvement statistically.

We also randomly prune 10 % to 90 % words from the original dictionary, which is denoted by WE-random, and compare the performance of WE-random with WE-min on the same training set (10 %) from Sina. The results are plotted in Figure 4. As WE-min is nonparametric, its result remains the same for all parameters. The proposed WE-min outperforms WE-random significantly with *p*-value<0.05 for both metric. On one hand, the influence of randomly pruning on *Acc@*1 is slightly with different percentage of pruned words. This may indicate that a small proportion of dictionary is sufficient to classification in terms of *Acc@*1, which only focuses on predicting the most popular emotion. However, the value of *Acc@*1 decreases largely when 90 % words are pruned. On the other hand, the percentage of pruned words has huge effect on *AP*. When more than 50 % words are pruned, the quality of predicting a correct distribution of all emotions reduces largely.

4.3.5 Evaluation on short documents

Despite that our focus is mianly on sentiment analysis of news bodies, it would be interesting to evaluate all algorithms on short texts. SemEval data set contains 1,246 valid news

Table 6 The *p*-values of statistical significance test on paired algorithms

Paired algorithms	Acc@1	AP
WE & WE-minrand	0.4554	0.0352
WE-min & WE	8.8E-12	4.2E-8
WE-min & WE-minrand	1.5E-8	2.0E-8

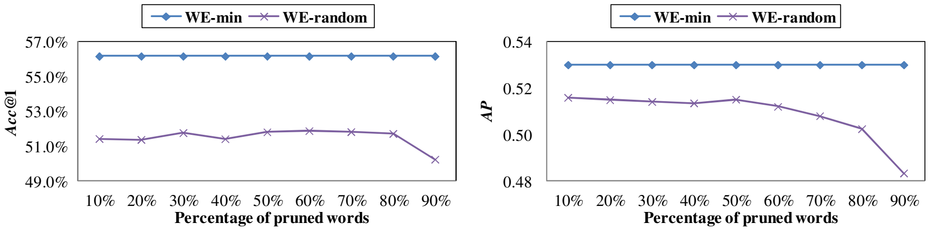

Figure 4 The *Acc*@1 and *AP* of WE-min and WE-random with different percentage of pruned words

headlines and their emotional scores annotated by users. The news documents in SemEval are of sentence level and short enough. SemEval is officially divided into a trail-set with 246 documents and a test-set with 1,000 documents. We train all algorithms on the trial-set, and evaluate them by classifying the same test-set, thus the statistical significance test is not further performed. Within the test-set, there are 28 documents' words not appeared in the trail-set. The emotional distributions of those documents can only predicted randomly for ET, SWAT and WE. More documents in the test-set will be randomly predicted if pruning strategies are used to refine the dictionary. For instance, there are 438 (43.8 %) testing documents' words not appeared in the dictionary generated by WE-min. Considering the limited information of the data set, we evaluate the proposed algorithm without pruning (i.e., WE) and Emotion LDA model (ELDA). The topic number of ELDA and the baseline ETM varies from 20 to 400 and the average result is used. The performance of different algorithms is shown in Table 7.

The results on short documents are different from those on news bodies. Firstly, the SWAT system outperforms the others. This is because SWAT generates a document by the Bernoulli model, in which a binary indicator is assigned for each term. For short documents (e.g., news headlines), the Bernoulli model works best [19]. Secondly, the algorithms exploiting LDA (i.e., ELDA and ETM) perform worse on short documents. It is caused by a particular challenge faced by LDA and other topic models. The existing experimental result shows that short documents lack enough content from which statistical conclusions can be drawn easily [26].

5 Conclusions

Emotion and opinion mining are useful and meaningful from political, economical, commercial, social and psychological perspectives. As the first step to meet the needs, we have presented algorithms of constructing a word-level and a topic-level emotional dictionaries in this paper. Different from previous methods, our algorithm of building the emotional

Table 7 The *Acc*@1 and *AP* of algorithms on SemEval

Algorithms	Acc@1	AP
SWAT	33.5 %	0.33
WE	33.0 %	0.31
ET	31.0 %	0.24
ELDA	20.5 %	0.27
ETM	25.6 %	0.23

dictionary is automatic, language-independent, volume-unlimited, and fine-grained. The main conclusions are as follows:

The pruning strategies are effective in refining the word-level emotional dictionary, and efficient in improving the performance of emotion classification. For the three pruning strategies, i.e., maximum, average and minimum, the last one achieves the biggest improvement on the performance, and the improvement is statistically significant under the statistical significance test. For the method of building the topic-level emotional dictionary, the dimension of the dictionary can be reduced largely, and the performance of applying the topic-level emotional dictionary is competitive when compared with others.

As the number of training documents increases, the quality of social emotion classification increases also. The performance of our algorithms is better than two baselines (i.e., ET and SWAT), and it is competitive with the other baseline ETM.

For annotating emotions of news headlines, it is unnecessary to prune the dictionary, due to the limited vocabulary in the short texts. Although the performance of our algorithm is much better than the previous method on news bodies, it is less noticeable on news headlines. Thus, further investigations on emotional annotation for both long and short texts constitute a major part of our future research.

Acknowledgments We are grateful to the anonymous reviewers for their constructive comments. The research work described in this article has been funded by a Strategic Research Grant from City University of Hong Kong (project no. 7002912), Natural Science Foundation of China (No. 61272437, No. 61073189), Innovation Program of Shanghai Municipal Education Commission (No. 13ZZ131) and Foundation Key Project of Shanghai Science and Technology Committee (No. 12JC1404500), Project of Shanghai Science and Technology Committee (12510500700)

References

1. Baccianella S., Esuli A., Sebastiani F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of The 7th Conference on Language Resources and Evaluation, pp. 2200–2204, (2010)
2. Banea, C., Mihalcea R., Wiebe J.: A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In: Proceedings of The 6th International Conference on Language Resources and Evaluation, (2008)
3. Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., Yu, Y.: Mining social emotions from affective text. IEEE Trans. Knowl. Data Eng. **24**, 1658–1670 (2011)
4. Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., Yu Y.: Joint emotion-topic modeling for social affective text mining. In: Proceedings of The 9th IEEE International Conference on Data Mining, pp. 699–704, (2009)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Cai, D., Mei, Q., Han, J., Zhai, C.: Modeling hidden topics on document manifold. In: Proceedings of The 17th ACM Conference on Information and Knowledge Management, pp. 911–920, (2008)
7. Chaumartin, F.R.: Upar7: A knowledge-based system for headline sentiment tagging. In: The 4th International Workshop on Semantic Evaluations, 422–425, Association for Computational Linguistics, (2007)
8. Das, S., Chen, M.: Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: Proceedings of The 8th Asia Pacific Finance Association Annual Conference, (2001)
9. Griffiths T. L., Steyvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America, 101, pp. 5228–5235, (2004)
10. Hofmann, T.: Probabilistic latent semantic indexing. In Proceedings of The 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57, (1999)

11. Ifrim, G., Weikum, G.: The bag-of-opinions method for review rating prediction from sparse text patterns. In: Proceedings of Coling, (2010)
12. Jindal, N., Liu, B.: Opinion spam and analysis. In: The International Conference on Web Search and Web Data Mining, pp. 219–230, (2008)
13. Katz, P., Singleton, M., Wicentowski, R.: Swat-mp: The semeval-2007 systems for task 5 and task 14, In: The 4th International Workshop on Semantic Evaluations, 308–313. Association for Computational Linguistics, (2007)
14. Koga, H., Taniguchi, T.: Developing a user recommendation engine on twitter using estimated latent topics. In: Proceedings of The 14th international conference on Human-computer interaction: design and development approaches - vol. Part I, pp. 461–470, (2011)
15. Kolya, A., Das, D., Ekbal, A., Bandyopadhyay, S.: Identifying event-sentiment association using lexical equivalence and co-reference approaches. In: Workshop on Relational Models of Semantics Collocated with ACL, pp. 19–27, (2011)
16. Lin, K.H.-Y., Yang, C., Chen, H.-H.: Emotion classification of online news articles from the reader's perspective. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 220–226, (2008)
17. Lin, K.H.-Y., Yang, C., Chen, H.-H.: What emotions do news articles trigger in their readers?. In: Proceedings of The 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 733–734, (2007)
18. Liu, J., Seneff, S.: Review sentiment scoring via a parse-and-paraphrase paradigm. In: Empirical methods in natural language processing, ACL, (2009)
19. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge University Press, pp. 156–281, (2008)
20. Moreo, A., Romero, M., Castro, J.L., Zurita, J.M.: Lexicon-based comments-oriented news sentiment analyzer system. Expert Syst. Appl. **39**, 9166–9180 (2012)
21. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Affect analysis model: novel rule-based approach to affect sensing from text. Nat. Lang. Eng. **17**(1), 95–135 (2011)
22. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Empirical methods in natural language processing, pp. 79-86, (2002)
23. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of The 20th Conference in Uncertainty in Artificial Intelligence, pp. 487–494, (2004)
24. Smet, W. D., Moens, M. -F.: An aspect based document representation for event clustering. In: Proceedings of The 19th meeting of Computational Linguistics, (2009)
25. Snow, R., Connor, B.O', Jurafsky, D., Ng, A.Y.: Cheap and fast-but is it good? Evaluation non-expert annotations for natural language tasks. In: Empirical Methods in Natural Language Processing, pp. 254–263, (2008)
26. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2330–2336, (2011)
27. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, Fortaleza, Brazil, pp. 1556–1560, (2008)
28. Strapparava, C., Mihalcea, R.: Semeval-2007 task 14: Affective text. In: Proceedings of The 4th International Workshop on Semantic Evaluations, pp. 70–74, (2007)
29. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: Proceedings of The 4th International Conference on Language Resources and Evaluation, pp. 1083–1086, (2004)
30. Tong, R.M.: An operational system for detecting and tracking opinions in on-line discussions. In: Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, pp. 1–6, (2001)
31. Turney, P.D., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus, Technical Report EGB-1094, National Research Council Canada, (2002)
32. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of The 40th annual meeting of the Association for Computational Linguistics, pp. 17–424, (2002)