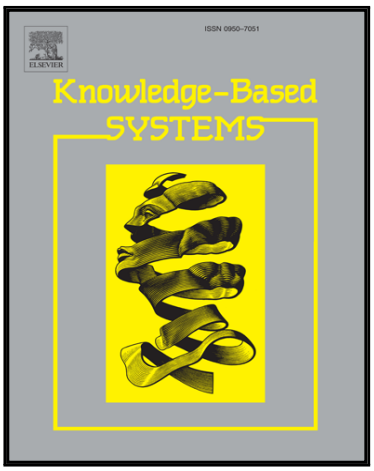


DWWP: Domain-specific New Words Detection and Word Propagation System for Sentiment analysis in the Tourism Domain

Wei Li , Kun Guo , Yong Shi , Luyao Zhu , Yuanchun Zheng

PII: S0950-7051(18)30055-8
DOI: 10.1016/j.knosys.2018.02.004
Reference: KNOSYS 4212

To appear in: *Knowledge-Based Systems*

Received date: 7 April 2017
Revised date: 23 January 2018
Accepted date: 3 February 2018

Please cite this article as: Wei Li , Kun Guo , Yong Shi , Luyao Zhu , Yuanchun Zheng , DWWP: Domain-specific New Words Detection and Word Propagation System for Sentiment analysis in the Tourism Domain, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.02.004

DWWP: Domain-specific New Words Detection and Word Propagation System for Sentiment analysis in the Tourism Domain

Wei Li[1,2,3], Kun Guo[1,2,3,*], Yong Shi[2,3,4,*], Luyao Zhu[1,2,3], Yuanchun Zheng[1,3]

[1] *School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190, China*

[2] *Fictitious Economy & Data Science Research Center, Chinese Academy of Sciences, Beijing, 100190, China*

[3]*Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences*

[4] *College of Information Science and Technology, University of Nebraska at Omaha, NE 68182, USA*

[*]*Correspondence: yshi@ucas.ac.cn (Shi); guokun@ucas.ac.cn (Guo)*

Abstract

Online travel has developed dramatically during the past three years in China. This results in a large amount of unstructured data like tourism reviews from which it is hard to extract useful knowledge. In this paper, a *DWWP* system consisting of domain-specific new words detection (*DW*) and word propagation (*WP*) is presented. *DW* deals with the negligence of user-invented new words and converted sentiment words by means of *AMI* (Assembled Mutual Information). Inspired by social networks, the new method WP incorporates manually calibrated sentiment scores, semantic and statistical similarity information, which improves the quality of sentiment lexicon in comparison with existing data-driven methods. Experimental results show that *DWWP* improves seventeen percentage points compared with graph propagation and four percentage points compared with label propagation in terms of accuracy on Dataset I and Dataset II, respectively.

Keywords: Online travel, DWWP, Chinese new words detection, Sentiment lexicon, Word propagation

1. Introduction

The rapid development of online travel websites, such as Ctrip[1], Qunar[2], and Tuniu[3], causes a significant increase in user-generated content (UGC) [1]. UGC refers to reviews and interactions among users on travel websites in this paper [2]. Despite the large scale of tourism reviews on travel websites, the application of online textual data is limited since the manual and comprehensive extraction of useful knowledge from the reviews is costly and time-consuming [3–5]. Unlike written languages, such as journalism, tourism reviews are informal and colloquial. For instance, “开心” (happy) is usually written as “开森” (happy) in the tourism reviews. Thus, effective sentiment analysis (SA) techniques must be introduced to process such textual data.

SA is the task of detecting whether a textual item (e.g., a product review and a blog post) expresses a *positive* or *negative* opinion in general or about a given entity (e.g., a product, person, or policy), which has been the research focus in natural language processing (*NLP*) [6]. Sentiment lexicon consists of a substantial amount of sentiment words and phrases with explicit sentiment scores. It has a significant influence on the performance of lexicon-based SA [7,8]. Dictionary-based and corpus-based methods mainly achieve the building of sentiment lexicon, where corpus-based methods are further divided into statistical-based and semantics-based methods according to the specific techniques [9]. Gonzalez-Rodriguez et al. [10] analyzed sentiment orientations of travel-related information on social media via AFINN-111 [11]. The final sentiment score of a review was calculated based on the relative amounts of four types of sentiment words. Xianghua et al. [12] employed topic modeling and HowNet lexicon to perform aspect-level SA for Chinese online social reviews. The method can either calculate sentiment orientations of words or identify topics but fails to cover user-invented words. More concretely, existing lexicon-based SA methods get bad performances on tourism reviews due to two issues. First, proper nouns,

[1] http://www.ctrip.com/
[2] https://www.qunar.com/
[3] http://www.tuniu.com/

converted words, user-invented words and multiword expressions (MWEs) in tourism reviews, are not included in traditional sentiment lexicons. For instance, “千古情” (a large costume show), a common proper noun in the tourism domain, is rarely seen in other domains. The manual detection of these informal words is time-consuming and costly, especially when faced with vast amounts of reviews. Second, some data-driven sentiment lexicon construction methods fail to provide rigorous sentiment scores or intensities for corresponding sentiment words. In addition, the advanced machine learning tool word2vec is rarely used to assist the construction of sentiment lexicon. Therefore, automatic and effective construction of a high-quality tourism-specific sentiment lexicon is of great value.

In contrast to English SA, Chinese SA was difficult due to the lack of segmentation symbols like blank spaces that separated consecutive Chinese characters into words. In this case, the aforementioned four types of words cannot be easily detected by Chinese word segmentation tools [11]. Thus, the direct use of English SA methods in Chinese SA tasks was not applicable. Besides, there were two limitations in the state-of-the-art sentiment lexicons. The first limitation was that few emoticons existed in the tourism reviews compared with microblogs and tweets, which made it hard to copy existing sentiment lexicon construction approaches in the tourism domain. Another limitation was that only statistical information was used to calculate the sentiment scores of words, causing poor performance on the sentiment scores.

To overcome these limitations, a domain-specific new words detection and word propagation system (*DWWP*) was proposed to build a robust high-quality sentiment lexicon. This system is not subject to types of languages and it improves the accuracy and robustness of SA in the tourism domain. Main contributions of this paper are shown as follows:

1. Considering the complexity of Chinese, *DW* is presented to detect user-invented new words, converted words, proper nouns and multiword expressions.

2. We present a sentiment lexicon construction method combining seed word scores, statistical information and semantic information derived from word2vec, which significantly improves the performance of the sentiment lexicon.

3. The combination of optimization function and iteration algorithm in *WP* contributes to increasing the efficiency and accuracy of word propagation.

4. The *DWWP* system outperforms traditional sentiment lexicons, state-of-the-art methods LP and GP significantly, and serves an important role in subsequent tasks like precision marketing and decision making.

The rest of the paper is organized as follows: In section 2, the related work is introduced. In section 3, the *DWWP* system is elaborated. Experimental results are discussed in section 4. The last section concludes the paper and gives the future insights.

2. Related Work

Research papers about domain new words detection and SA are introduced in this section. SA techniques mainly consist of two types, machine learning and lexicon-based approaches [9].

2.1 New words detection

Because Chinese lacks segmentation symbols (blank spaces), domain new words detection served an important role in SA [13]. Four types of new words detection methods existed [14]. The first type of methods incorporated new words detection into a word segmentation task [15]. In this case, the most probable word segments that were not included in existing dictionaries were regarded as new words. On the basis of complicated linguistic rules and knowledge [16], the second type of methods extracted new words, although it was hard to devise these rules. The third type of methods took new words detection as a classification task [17]. However, designing proper linguistic features and labeling enough training data were time-consuming and required extra human work. The fourth type of methods that belonged to statistics-based methods considered new words detection as an unsupervised learning task [18,19], which took use of

unlabeled data and did not need extra work, such as devising complex rules and designing proper linguistic features. Pointwise mutual information (*PMI*) [18], multiword expression distance (*MED*) [19] and enhanced mutual information (*EMI*) [20] were three important indicators in statistics-based methods. The proposed method *DW* is a statistics-based method, and it makes improvements based on *EMI* to satisfy specific conditions in Chinese and increases the quality of newly detected words.

2.2 Machine learning sentiment analysis approach

In the aspect of machine learning approaches, additional research focuses on supervised learning. Santos et al. [21] proposed a convolutional neural network named CharSCNN to extract relevant features from the character-level to the sentence-level to make SA from short texts. Khalil et al. [22] harnessed an ensemble binary classifier that consists of CNN and SVM for aspect category and sentiment extraction. Tai et al. [23] proposed a tree-LSTM to predict the semantic relevance of two sentences and achieved sentiment classification tasks. Besides basic machine learning approaches, some lexicon based machine learning approaches make a great progress in improving the performance of SA. Khan et al. [24] presented a hybrid approach named SWIMS that incorporated machine learning with a lexical based approach, where SentiWordNet was used to determine the feature weight and support vector machine was utilized to learn the feature weights. Moreover, Khan et al. [25] built a general purpose sentiment lexicon in a semi-supervised manner. They used Expected Likelihood Estimation Smoothed Odds Ratio to define word semantics, then the well-defined word semantics were incorporated with supervised machine learning based model selection approach. These hybrid approaches showed the superiority in SA and should be explored further.

In recent years, some insightful deep models have been published in the top-tier journals and major conferences [26–30]. S. Poria et al. [31] proposed the first deep learning approach to aspect extraction in SA and used the deep convolutional neural networks to detect sarcastic tweets [32].

After that, they presented a novel model extracted visual and textual features to feed a multiple kernel learning classifier and got state-of-the-art results on several challenging datasets [33].

2.3 Lexicon-based sentiment analysis approach

2.3.1 *Traditional sentiment lexicon*

Words that express opinions or sentiments are applied to textual SA [34]; Thus, it is of vital importance to build a sentiment lexicon that contains numerous sentiment words. SenticNet, SO-CAL, AFINN, Opinion lexicon, Subjectively lexicon, General Inquirer, WordNet, WordNet-affect, and SentiWordNet were the state-of-the-art lexical resources and were built in a supervised manner or by experts [9]. Traditional sentiment lexicons, manually built by experts in general [35], required huge human work and resources. Miller et al [36] proposed an English dictionary WordNet based on cognitive linguistics. The dictionary was widely used for English SA since its release. Strapparava et al. [37] proposed a WordNet-based sentiment dictionary named WordNet-Affect through the selection and annotation of the WordNet subset. In addition to WordNet, Richardson et al. [38] constructed another emotional dictionary named MindNet. Based on extensive area, MindNet was a broad-coverage natural language parser from fully automated construction. It provided the most credible prospect for supporting extraction of knowledge. Besides, the National Taiwan University Sentiment Dictionary (*NTUSD*) [39] proposed by Ku et al. was primarily used for analyzing Chinese microblogs and public mood of the Chinese Internet.

2.3.2 *Data-driven sentiment lexicon*

The traditional sentiment lexicons had some drawbacks. For instance, it failed to handle diverse SA tasks and had poor coverage in a new domain. Xu et al. [40] noted that the cross coverage rate of the existing sentiment lexicons HowNet and NTUSD was less than ten percentage points; thus, they proposed a word graph method using link analysis. It used a substantial amount of unlabeled data for training, which extended the existing sentiment lexicon

and calculated a sentiment score for each word. Afterwards, Tang et al. [41] proposed a large-scale Twitter-specific sentiment lexicon to solve the problem of remarkable differences between twitter vocabulary and WordNet sentiment words. Furthermore, Saif et al. [42] put forward a Twitter-specific sentiment lexicon in which they specified a fixed sentiment score for each word and then updated the scores by means of the co-occurrence number of sentiment words in different documents. Extensive experiments demonstrated that the method outperformed other SA methods of the same type. In addition, Feng et al. [43] proposed a data-driven method aimed at building a sentiment lexicon with extensive coverage. Sentiment scores of words were calculated by measuring the relative similarity degree towards positive and negative emoticons. The method cannot detect user-invented words, which weakened the value in use. To this end, Wu et al. [14] extended *Feng*'s work in Chinese microblog via detecting new words and introducing three types of sentiment knowledge. However, it should be noted that this method did not take the semantic knowledge into consideration. Cho et al. [44] took use of the merging, removing and switch operations to adjust sentiment lexicons to be suitable for various domains. Similarly, Bravo-Marquez et al. [45] expanded the opinion lexicon in a supervised manner considering word-level attributes that consist of morphological information and associations between words and sentiments in tweets, which were modeled by statistical approaches, pointwise mutual information semantic orientation (PMI-SO), and stochastic gradient descent semantic orientation (SGD-SO).

Existing lexicon-based SA methods get poor performances in specific domains. WordNet, HowNet, and some other traditional sentiment lexicons cannot handle complicated SA tasks. In contrast, data-driven approaches show its potential in increasing the textual SA performance and adaptability. Besides, few studies have been done on analyzing online tourism reviews, especially Chinese texts. Therefore, the SA system *DWWP* is presented to dramatically increase the classification accuracy and extend the scope of SA applications.

3. Proposed mechanism

The *DWWP* system, as shown in Fig. 1, consists of two parts: domain new words detection (*DW*) and word propagation (*WP*).

DW involves the detection of user-invented words, converted words, proper nouns and multi-word expressions, serving an important role in improving the quality of sentiment lexicons. It starts with the processing of tourism data. In this paper, tourism review data is taken as an example to elaborate *DW*. The *DW* part corresponds to the proposed new words detection mechanism in subsection 3.1. Then, a part of new words detected by *DW* are added to the seed word set to start *WP*.

Seed words with manually labeled scores in the seed word set, together with the semantic information, are used to calibrate numerous unlabeled words. Then, manual information, semantic and statistical information are integrated into an optimization function to optimize the sentiment scores of seed and unlabeled words. These optimized words update the primal seed word set. Next, the second part *WP* is iteratively performed until the convergence condition is satisfied. It is described in detail in subsection 3.2. In practice, the *DWWP* is available for auxiliary management, precision marketing, decision making and hot scenic spots. Furthermore, the tourism-domain SA are used as an application case to elaborate the entire DWWP system.

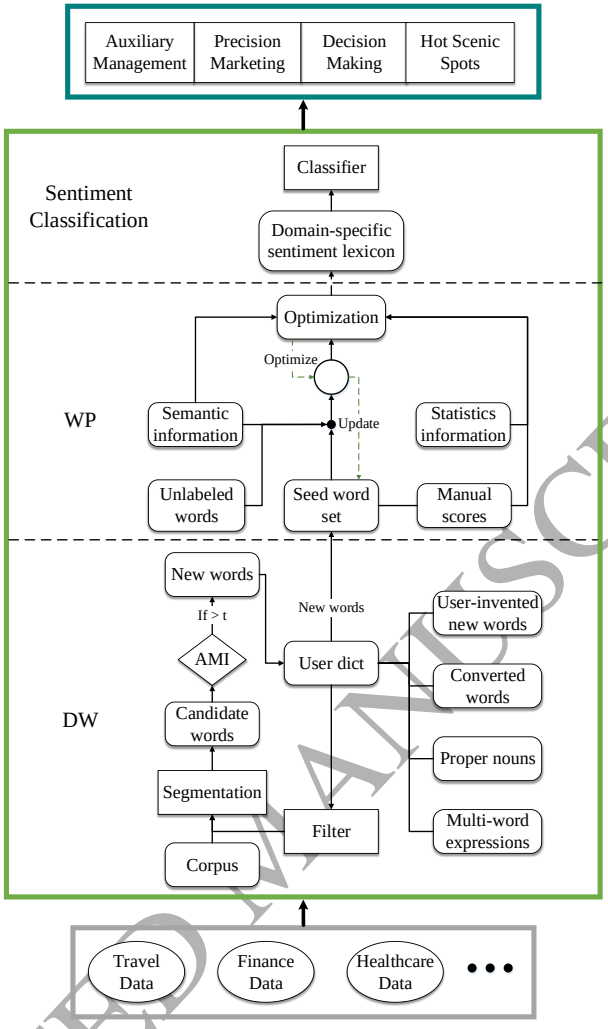


Figure 1. Domain words detection and word propagation system

3.2 Proposed new words detection mechanism

The majority of the research in the SA domain focuses on the English text. Although several attempts occurred in the Chinese SA, some specific problems remained. In English, two words are segmented by a blank space, whereas there is no token assistance used for word segmentation in Chinese. Unfortunately, the performances of *NLP* tasks deeply depend on the effect of Chinese word segmentation (e.g., word embedding, lexicon-based and machine learning SA approaches). What's more, many user-invented words cannot be recognized by segmentation tools, like *jieba* and *ANSJ*. For instance, "纯玩无购物" which stands for a trip that guarantees genuine goods at a

fair price, contains a slight positive sentiment when used in the tourism reviews. These words are referred to as converted words. Another key problem is that segmentation tools may not perform well in the practical cases. Numerous domain-specific words that convey sentiments cannot be detected due to a lack of words in *jieba* dictionary. For example, in “魅力湘西” (Charming Xiangxi), “魅力”(Charming) and “湘西”(Xiangxi) mean lure and a place name, respectively. “魅力湘西” has a special meaning in the tourism domain, which relates to a type of drama in Zhangjiajie, Hunan Province. Plenty of exclusive sentiment words exist in specific domains. “跟团游” (package tours), emerging with online travel, is not regarded as a word in the traditional dictionary.

Although it seems to be separable in appearance, some Chinese words in the tourism reviews cannot be segmented into characters and/or words further. Otherwise, the word would lose its original meaning. For instance, “我们是跟团游” (we are on a package tour) will be segmented into “我们/是/跟/团游”. The segmentation tool *jieba* cannot identify the new word “跟团游” (package tour); the new word is segmented into “跟” and “团游”, which is puzzling. To this end, *DW* is presented to detect new words. *EMI* [20] is selected as a basic indicator, and then it is extended into *AMI* (Assembled Mutual Information) to solve the practical issues. The definition of *EMI* is expressed as:

$$EMI(w) = \log \frac{n_w / N}{\prod_{i=1}^{T} [(n_{m_i} - n_w) / N]} \qquad \text{Formula 1}$$

where $\boldsymbol{w}$ means the candidate new word, composed of T single morphemes that are marked as $\boldsymbol{m_1}$, $\boldsymbol{m_2}$, $\boldsymbol{m_3}$,…, $\boldsymbol{m_T}$. It should be noted that the morpheme is the smallest component of a Chinese word and cannot be further segmented. To be specific, the morpheme $\boldsymbol{m_i}$ refers to the element from the initial segmentation result of *jieba*, which is a single character or a word that consists of multiple characters. $\boldsymbol{n_w}$ is the occurrence number of $\boldsymbol{w}$, whereas $\boldsymbol{n_{mi}}$ is the occurrence number of $\boldsymbol{m_i}$. $\boldsymbol{N}$ stands for the total number of documents [20]. If $\boldsymbol{m_1}$ $\boldsymbol{m_2}$ $\boldsymbol{m_3}$, … , $\boldsymbol{m_T}$ always occur in company with each

other and the occurrence number of a morpheme $\boldsymbol{m}_i$ is relatively small, $\boldsymbol{w}$ is more probable to be a new word.

However, *EMI* is not good enough when used in new words detection. Formula 1 is transformed into Formula 2 to give more insights about the indicator.

$$\begin{aligned} EMI(w) &= \log n_w - \log N - [\sum_T \log(n_{m_i} - n_w)] + T \log N \\ &= \log n_w - [\sum_T \log(n_{m_i} - n_w)] + (T-1)\log N \end{aligned} \quad \text{Formula 2}$$

First, as shown in Formula 2, *T* has an influence on *EMI*, whereas the influence is uncertain. The existence of *N* (usually significantly larger than n_w and n_{m_i}) makes *EMI* change rapidly as *T* changes. It is known that *EMI* works by setting a threshold to select new words. However, it is unreasonable to set a one-size-fits-all threshold for *T*-grams with a different value of *T*.

Second, *EMI* disregards the situation $n_{m_i} = n_w$, which usually occurs in proper nouns and user-invented words, such as “茶马古道” (ancient tea route) and “坑爹” (cheating). Under the circumatances, $\log(n_{m_i} - n_w)$ cannot be easily calculated without any adjustment.

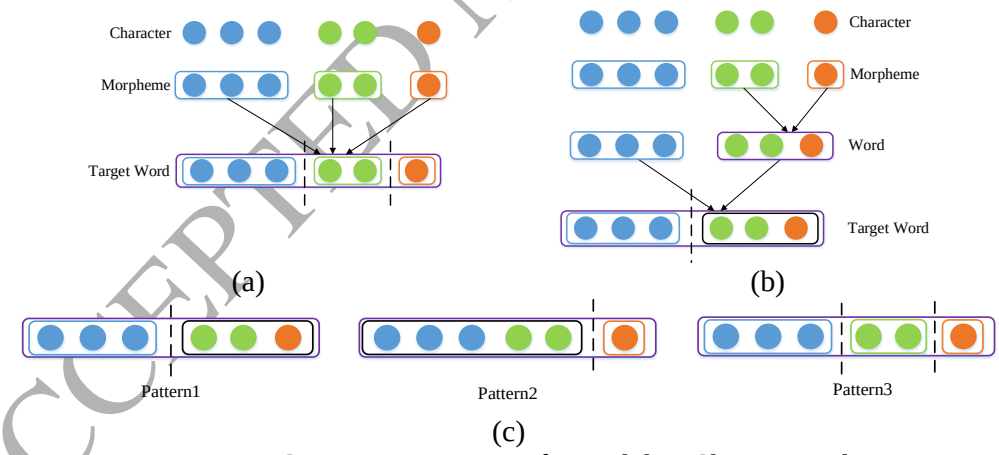


Figure 2. Constitution patterns of a candidate Chinese word

Due to the complexity of Chinese words’ structure, the morpheme of a word can be a single Chinese character or a word consists of more than one character. A word can be a morpheme or a combination of some other words. As shown in (b) in Fig. 2, a candidate word includes several words which can be the combination of morphemes or words. However, *EMI* only considers the

case in which a candidate word is directly composed of single morphemes, as shown in (a) in Fig. 2. In view of this, a qualified indicator should consider all possible constitution patterns of a candidate word. For example, in (c) of Fig. 2, a candidate word is treated as the combination of words and the combination of morphemes.

The basic indicator *EMI* is changed into *AMI* to overcome the aforementioned drawbacks. As shown in Formulas 3 and 4 (the identity transformation of Formula 3), the existence of *Tth* root reduces the influence of *T* and eliminates the effect of *N*. ***s_f*** is a smoothing factor that guarantees the robustness of the new indicator, solving the second issue where $\boldsymbol{n_w = n_{wi}}$. Formulas 3 and 4 consider different constitution patterns of a candidate word and summarize them with $\sum_j$, where *j* represents the *jth* constitution pattern of the word.

$$AMI(w) = \sum_j \left(\log \frac{n_w / N}{\sqrt[T]{\prod_{i=1}^{T} [(n_{w_i}^j - n_w + s\_f) / N]}} \right) \quad \text{Formula 3}$$

$$\begin{aligned} AMI(w) &= \sum_j \{\log n_w - \log N - \frac{1}{T}[\sum_{i=1}^{T} \log(n_{w_i}^j - n_w + s\_f) - T \log N]\} \\ &= \sum_j \{\log n_w - \frac{1}{T}[\sum_{i=1}^{T} \log(n_{w_i}^j - n_w + s\_f)]\} \end{aligned} \quad \text{Formula 4}$$

Due to the considering the constitution patterns of Chinese words, *AMI* works well in the recognition of multi-word expressions. A multi-word expression is a sequence of neighboring words "whose exact and unambiguous meaning or connotation cannot be derived from the meaning or connotation of its components" [46]. Note that the magnitude of MWEs was significantly greater than the magnitude given by linguistics [47]. According to an estimation, the number of MWEs in a speaker's lexicon was of the same order of magnitude as the number of single words [48]. F.B. et al. demonstrated that 41 percent of the entries were multi-word expressions [19]. In modern times, some specialized domain vocabularies, like terminologies and proper nouns, consist of MWEs. MWEs such as "全面二孩政策" (the universal two-child policy) and "长短期记忆模型" (long short-term memory) are common in Chinese.

The number of n-grams increases[4] sharply as n increases, requiring a large amount of memories. Therefore, the parameter n is set to three in the case that the number of candidate words increases rapidly. In addition, in consideration of the property of Chinese grammars, a stopword list and an adverb list are introduced to improve the performance of *DW*. Inspired by [14], an iterative method that prominently reduces the memory requirements is put forward compared with the method by Church et al. [18].

Considering all aspects, *DW* is given by pseudo code in Algorithm 1. Each iteration aims to detect suitable new words, and the new words are added to the customized user dictionary ***D***. To this end, the customized user dictionary ***D*** is modified in each iteration. Some words whose occurrence numbers are lower than the threshold t_f are discarded, whereas some newly discovered words are added. When the size of the customized user dictionary ***D*** remains constant, the algorithm stops.

Algorithm 1: New words detection algorithm used in the tourism domain.

1: ***Input:*** Tourism reviews corpus ***C***, Chinese segmentation tool ***jieba***, the threshold of word frequency t_f and threshold of word *AMI* score t_e. Stopword list_start ***S_s***; stopword list middle ***S_m***; stopword list end ***S_e***. Adverb list ***A***.
2: ***Output:*** Customized user dictionary ***D***.
3: Establish the empty dictionary ***D***, set hyper-parameters t_f and t_e, Initialize iteration number num==0.
4: ***while*** the size of customized dictionary ***D*** does not maintain constant:
5: Load the customized user dictionary ***D*** to ***jieba***.
6: Clean raw corpus, then segments cleaned corpus ***C*** using segmentation tool ***jieba***.
7: Go through corpus ***C***, then count words' occurrence numbers, save txt[5] file $\boldsymbol{F} = \{(w_i, n_i) \mid i=1,2,\ldots\}$ and filtered txt file $\boldsymbol{Ff} = \{(w_i, n_i) \mid n_i <= t_f\}$.

[4] We have tested zhiwiki, and almost 140 million words exist, whereas the number of candidate new words is nearly 300 million. https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2
[5] Text format is also essential as the size of corpus increases.

8: Filter the words in ***D*** according to ***Ff***.

9: Go through corpus ***C***, find out all the bi-grams and tri-grams according to stopword list ***S_s***, ***S_m*** and ***S_e*** , count their occurrence numbers, and filter them according to t_f, save txt file ***co_count*** = {$(w_i, n_i) \mid n_i <= t_f$ }.

10: Compute *AMI* scores of candidate new words in ***co_count*** according to Adverb list ***A***.

    If the first word of the candidate word in adverb list ***A***:

        Reward candidate word

    else:

        pass

11: Use t_e to filter candidate new words in ***co_count*** according to their *AMI* scores, then add remnant words in ***co_count*** to ***D***.

12: num+=1

13: ***end***

14: Sort customized user dictionary ***D*** according to their occurrence numbers.

15: Return customized user dictionary ***D***.

3.2 Proposed Sentiment Lexicon Construction Approach

Lexicon-based approaches have considerable merits in the SA domain. The traditional lexicons built for general SA tasks have poor performance in the tourism domain. Besides, word2vec has been combined with machine learning based SA approaches in the most recent studies since its release [49]. Motivated by these observations, word propagation (*WP*) that incorporates word2vec into the automatic construction of tourism-specific sentiment lexicon is proposed. It should be noted that *DW* promotes the training of word2vec used in *WP*.

3.2.1 *Seed Sentiment Words*

Each word is represented as a vector in vector space, and bilateral relationships exist between a pair of words. Thus, a lexicon is treated as a social network in terms of the structure, in which each word stands for a person. In this case, an analogy is drew between real-world social networks and the lexicon. The goal is to find as many sentiment words as possible. Similar to

social networks, a large number of words are identified from a few groups of seed words. Therefore, seed words consists of new words detected by *DW* in accordance with their frequencies and some seed words from pre-existing lexicons. Each negative seed word is manually calibrated with a sentiment score ranging from -1 to -3, and a positive seed word's sentiment score ranges from 1 to 3. Three authors of this paper manually assigned the sentiment scores in the light of their experience and pre-existing sentiment lexicons.

3.2.2 *Word Propagation*

Word2vec measures the semantic similarities among different words, and the similarity degree is calculated using the following formula

$$\cos(\theta) = \frac{vector(w_1) \cdot vector(w_2)}{|vector(w_1)||vector(w_2)|} \quad \text{Formula 5}$$

where $\cos(\theta)$ ranges from -1 to 1, indicating the similarity degree of two different words. Note that if the cosine value is high, then these two words convey more similar sentiment polarities. Everyone knows that we can find friends through friends in social networks. Similarly, for each positive seed word, some of its similar words are sorted according to the cosine values. Therefore, more positive sentiment words are obtained by positive seed words, which is similar to web crawlers. In this research, negative sentiment words are obtained in a parallel manner. Motivated by collaborative filtering algorithm [50], seed word scores are used to automatically calibrate unlabeled words. The formula is defined as

$$Score(word) = \sum_{i \in P} \frac{\cos\theta_i \times seed_i}{|\cos\theta_i|} \quad \text{Formula 6}$$

where $cos\theta_i$ is the cosine value between the *ith* word in *P* and the *word*; *P* represents the positive seed words set, and $seed_i$ stands for the manually calibrated score of the *ith word*. *Score (word)* is equal to the average score given by the positive seed words set.

3.2.3 *Statistical Similarity PMI*

A common phenomenon is that two sentiment words that have similar sentiment polarities frequently co-occur. For instance, “导游很负责，推荐！”(The guide is responsible, recommend!), where “负责”(responsible) and “推荐”(recommend) have similar sentiment polarities that frequently co-occur. The statistical indicator *PMI* (pointwise mutual information) is utilized to measure the statistical similarities between words. The calculation formula is defined as

$$\text{PMIsimilarity}(w_i, w_j) = \log \frac{n(w_i, w_j) / N}{(n(w_i) / N)(n(w_j) / N)} \quad \text{Formula 7}$$

where $n(w_i, w_j)$ is the co-occurrence number of w_i and w_j; $n(w_i)$ and $n(w_j)$ are the occurrence number of w_i and w_j, respectively, and N is the number of documents [20]. Inferred from Formula 7, the *PMIsimilarity* can be negative in some cases; therefore, the negative *PMIsimilarity* is set to zero to guarantee the validity of the indicator.

3.2.4 *Construct domain-specific sentiment lexicon*

The aforementioned methods are incorporated into a framework to build a sentiment lexicon with rigorous sentiment scores. The framework incorporates seed word scores, semantic similarity scores (given by Formula 6) and PMIsimilarity scores. An optimization function is used to integrate these three types of scores to build the sentiment lexicon and obtain exact sentiment scores. Some mathematical notations are introduced before describing the optimization function. The vector $\mathbf{x} \in \boldsymbol{R}^{\boldsymbol{D}\times\boldsymbol{1}}$ is denoted as the final optimized scores of sentiment words in the sentiment lexicon, of which $\boldsymbol{D}$ is the size. Denote $\boldsymbol{s} \in \boldsymbol{R}^{\boldsymbol{D}\times\boldsymbol{1}}$ as the sentiment scores obtained by the seed word scores and scores of other unlabeled words given by Formula 6. Denote $\boldsymbol{W} \in \boldsymbol{R}^{\boldsymbol{D}\times\boldsymbol{D}}$ as the semantic similarity matrix given by Formula 5. W_{ij} is the semantic similarity score between *word i* and *word j*. Denote $\boldsymbol{P} \in \boldsymbol{R}^{\boldsymbol{D}\times\boldsymbol{D}}$ as the statistical similarity matrix given by Formula 7. P_{ij} is the statistical similarity score between *word i* and *word j*. According to the notations, the optimization function is defined as

$$\begin{aligned}\arg\min_{x} &= \sum_{i=1}^{D}(x_i - s_i)^2 + \frac{1}{2}\alpha\sum_{i=1}^{D}\sum_{j\neq i}W_{ij}(x_i - x_j)^2 + \frac{1}{2}\beta\sum_{i=1}^{D}\sum_{j\neq i}P_{ij}(x_i - x_j)^2 + \lambda\sum_{i=1}^{D}x_i^2 \\ &= \|\mathbf{x}-\mathbf{s}\|_2^2 + \alpha\mathbf{x}^{\mathrm{T}}\mathbf{L}_{\mathrm{w}}\mathbf{x} + \beta\mathbf{x}^{\mathrm{T}}\mathbf{L}_{\mathrm{p}}\mathbf{x} + \lambda\|\mathbf{x}\|_2^2 \\ s.t.\quad & a <= x_i <= b,\quad i = 1,2,...,D\end{aligned} \qquad \text{Formula 8}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are non-negative weight parameters, and λ is the non-negative regularization parameter. L_w is the Laplacian matrix of W, which is defined as $\boldsymbol{L_w = D_w - W}$, where D_w is a diagonal matrix and $\boldsymbol{D_{wii} = \sum_{j=1}^{D} w_{ij}}$. L_p is the Laplacian matrix of P, which is defined as $\boldsymbol{L_p = D_p - P}$, where D_p is a diagonal matrix and $\boldsymbol{D_{pii} = \sum_{j=1}^{D} p_{ij}}$.

In Formula 8, by minimizing $\boldsymbol{\sum_{i=1}^{D}(x_i - s_i)^2}$, it is expected that the final sentiment scores are close to seed word scores or scores given by Formula 6. Minimizing $\boldsymbol{\sum_{i=1}^{D}\sum_{j\neq i} w_{ij}(x_i - x_j)^2}$ indicates that a pair of sentiment words with strong semantic similarity should be as close as possible in the matter of the final sentiment scores. Similarly, minimizing $\boldsymbol{\sum_{i=1}^{D}\sum_{j\neq i} p_{ij}(x_i - x_j)^2}$ indicates that a pair of sentiment words with strong statistical similarity, their final sentiment scores, should not substantially differ. $\boldsymbol{\|x\|_2^2}$ is the *L2-norm* regularization, which is inspired by [51] and can set some final sentiment scores near zero. The *L2-norm* is introduced to reduce the influence of non-sentiment words. Furthermore, the constrain means that final sentiment scores should be in an interval, where a and b are constant and represent the manually assigned infimum and supremum of sentiment scores, respectively. The optimization function is convex with a constrain; thus, the interior point method [52] is used to solve it. Gradient is useful for solving this problem, and is defined as

$$gradient = 2\mathbf{A}^{\mathrm{T}}(\mathbf{AX} - \mathbf{S}) + \alpha(\mathbf{L}_{\mathrm{w}} + \mathbf{L}_{\mathrm{w}}^{\mathrm{T}})\mathbf{X} + \beta(\mathbf{L}_{\mathrm{p}} + \mathbf{L}_{\mathrm{p}}^{\mathrm{T}})\mathbf{X} + 2\lambda\mathbf{X} \qquad \text{Formula 9}$$

where $\boldsymbol{A \in R^{D\times D}}$ is an identity matrix. The entire method of *WP* is summarized in Algorithm 2.

Algorithm 2. Sentiment Lexicon Construction Algorithm

1: **Input**: New words ***N***, tourism dataset ***D***, Chinese segmentation tool ***jieba***, positive seed words set $\boldsymbol{P}$ = {(seed_i, score_i) | i=1,2, …}.

2: **Output**: Final sentiment lexicon ***F***.

3: Initialize final sentiment lexicon ***F*** with ***P***.
4: Add new words ***N*** detected in algorithm 1 to Chinese segmentation tool ***jieba***, then use ***jieba*** segments tourism dataset ***D***.
5: Use segmented tourism dataset ***D*** to train word2vec model, save word2vec model.
6: Load word2vec model.
7: **while** the convergence condition is not satisfied **do**
8: iter = iter + 1.
9: for seed in ***P***:
10: Find top 20 new words according to cosine values computed by word2vec model.
11: end for
12: Sort new words according to cosine value, then select some new words with high cosine values.
13: Assign scores according to Formula 6 for each word selected in step 12, then save them in $\boldsymbol{S} = \{(w_i, score_i) \mid i=1,2, \ldots\}$.
14: Add positive seed words set to ***S***.
15: Find all the words pairs in ***S***, then compute their cosine similarity scores and PMI similarity scores. Save scores in matrix ***W*** and ***P***, respectively.
16: Solve optimization function by Interior Point algorithm, then obtain a solution ***X***.
17: Use ***X*** to extend ***P***.
18: **end while**
19: Use extended positive seed word set ***P*** to refresh final sentiment lexicon ***F***.
20: Return ***F*** as the output sentiment lexicon.

First, detected new words are added to *jieba dictionary* to improve the training of word2vec model. Then positive seed words are used to initialize the final lexicon ***F***. In each iteration, the wod2vec model is applied to propagate new sentiment words, then Formula 6 is used to automatically calibrate unlabeled words. Finally, interior point algorithm is taken to solve the optimization function, then positive seed word set is extended by the optimized new words. Different parts of the optimization function adjust sentiment scores to reflect the true sentiment degrees. Similarly, the Algorithm 2 is performed again to build a negative sentiment lexicon.

4. Experiment

The *DWWP* system is performed on tourism reviews from Ctrip.com and Qunar.com, two of the largest travel websites in China.

The reviews, amounting to nearly 20 MB, captured by web crawlers, among which 14948 reviews come from Ctrip.com dating from 29th April 2012 to 30th June 2016 and 55581 ones are from Qunar.com dating from 1st November 2012 to 1st March 2016.

4.1 DW Experiment Analysis

Table 1 shows various indices of Top 10 new words detected by DW sorted by frequencies.

Table 1. Typical words detected by DW (Top 10)

No.	Word	Meaning	F	AMI	IA	IP	SA	SP	PN	NN	C
1	跟团游	Package tours	1052	4.51	O	0	-	-	1	0	0
2	千古情	Eternal love, a large costume show	533	8.89	O	0	-	-	1	0	0
3	高原反应	Altitude stress	481	5.97	O	0	S	-1	1	0	0
4	散客拼团	FIT(Foreign Independent Tourist) group	362	4.75	O	0	-	-	1	0	0
5	棒棒哒	Awesome, amazing, good	284	5.20	S	1	-	-	0	1	1
6	靠谱	Reliable	267	5.68	S	1	-	-	0	1	1
7	点个赞	Like	259	4.14	S	1	-	-	0	1	1
8	高大上	High-end and classy	243	3.40	O	0	S	1	0	1	1
9	篝火晚会	Bonfire party	201	6.11	O	0	-	-	1	0	0
10	玻璃栈道	Glass paths built along the face of a cliff	196	7.34	O	0	-	-	1	0	0

As shown in Table 1, *Word* column is the Chinese new words detected by *DW*. *Meaning* column corresponds to the English meaning of the words. *Inherent Attribute* (IA) column is an option on judging whether a detected word is subjective or objective intrinsically. *Inherent Polarity* (IP), refers to the inherent sentiment polarities of detected words, with -1 representing negative emotions, 0 standing for objectivity and 1 denoting positive emotions. *Special Attribute* (SA) column is marked with capital letters if it is a converted word, while S or O depends on that the transferred meaning tends to be subjective or objective; And default value is "–" when there is no transferred meaning. *Special Polarity* (SP) refers to special sentiment polarities of converted words, with -1 denoting negative emotions, 0 standing for objectivity and 1 representing positive

emotions. *Proper nouns* (PN) column, *Networking Neologism* (NN) column, and *Colloquialism* (C) column are options on judging whether the detected words come from proper nouns, networking neologism or colloquialism respectively, with 1 representing yes and 0 standing for not. Obviously, the detected words include catering, tourism products or styles, accommodation, weather and climate, shopping and feelings. These words mainly refer to tourism topics and rarely exist in other domains.

Some statistical results are shown in accordance with the attributes presented above and are concluded in Table 2 in the light of Top 25 new words detected by *DW*. *IAO* represents the words classified into objective ones while *IAS* represents the words classified into subjective ones in terms of *Inherent Attribute*. Other abbreviations are the same as in Table 1.

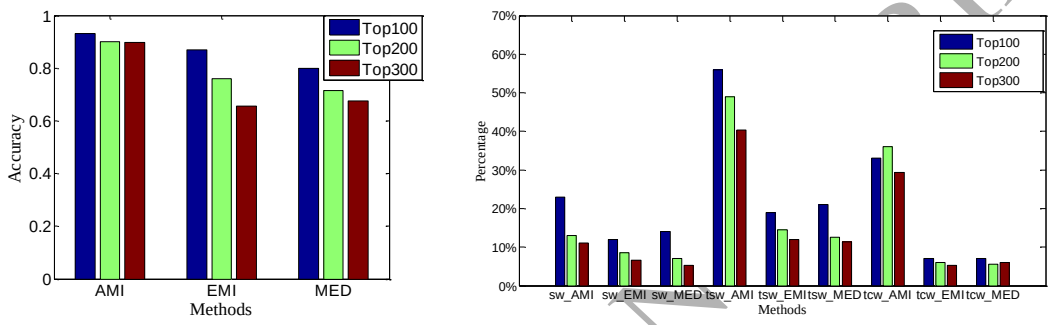
Table 2. Brief analysis of detected new words

Attribute	Count	Proportion
IAO	15	0.60
IAS	10	0.40
SP	5	0.20
PN	15	0.60
NN	9	0.36
C	9	0.36

It is shown in Table 2 that *DW* performs well in detecting subjective words, converted words, proper nouns, network neologism and colloquialism. For example, “纯玩无购物”, which means genuine tour without shopping, usually an objective word, has a special transferred meaning while used in tourism reviews. Such words usually convey the subjective emotions and should be included in proper nouns. As for IAO, the majority of them may have transferred meanings or belong to PN, NN or C. In addition, the results imply that *DW* contributes to the detection of words not included in general lexicons.

Fig. 3 shows the overall distribution characteristics of *EMI*, *MED* [19] and *AMI* performances on Chinese new words detection. Here are some illustrations on abbreviations in Fig. 3. All the indices stated below are based on Top100, Top200, Top300 new words given by

different methods. (a) of Fig. 3 shows the overall distribution of the accuracy of three methods. In (b) of Fig. 3, sw_xxx stands for the percentage of sentiment words; tsw_xxx denotes the percentage of sentiment words in tourism domain; tcw_xxx represents the percentage of converted words or connotations in the tourism domain, which can also be calculated by subtracting sw_xxx from tsw_xxx. And xxx corresponds to the names or abbreviations of different methods.



(a) Accuracy of detected new words (b) Proportion of sw, tsw and tcw

Figure 3. Performances analysis of three different methods (bar chart)[6]

As is shown, the local peaks of the bar chart always correspond to the statistical of *AMI*. Besides, there is no obvious distinction of performances between *EMI* and *MED*. So it is concluded that *AMI* performs apparently better than *EMI* and *MED* in terms of accuracy, tourism-specific sentiment words and converted words detection. What's more, the three methods' performances tend to get poorer with the growth of summarized new words numbers. That is, Top100 results perform better than Top200 while Top200 is better than Top300 in general.

4.2 WP Experiment Analysis

The first step of *WP* is to find seed sentiment words. The seed words consist of detected new words and words from pre-existing sentiment lexicons. After manually tuning, there are 209 positive and 257 negative sentiment words in which 99 positive and 79 negative words come

[6] The accuracy is calculated by manual inspection.

from detected new words. Seed sentiment words are composed of a lot of high-frequency sentiment words that guarantee the propagation process smoothness and some low-frequency words with wide coverage. Sentiment scores of positive seed words range from 1 to 3 while negative ones range from -3 to -1. The larger the absolute value of the sentiment score is, the higher the degree of the sentiment polarity becomes. Next, 1208 new words detected by *DW* are added to *jieba dictionary* to segment 70529 tourism reviews about tourism products. The values of α, β and λ are set to 0.5, 0.2, and 0.05, respectively.

Table 3: Typical examples of the tourism-specific sentiment lexicon

No.	Positive			Negative		
	#Word	#Meaning	#Score	#Word	#Meaning	#Score
1	超级棒	Very good	2.53447	不负责任	Irresponsible	-1.78432
2	挺靠谱	Very reliable	1.78023	冷嘲热讽	Cynicism	-1.64733
3	不二之选	Solid choice	2.67585	强买强卖	Hard sale	-2.39921
4	过瘾	Damn good	1.61217	受骗	Dupery	-1.7806
5	美如画	Picturesque	1.7876	商业化	Commercialized	-0.801148
6	嗨皮	Happy	1.77033	晚点	Delay	-0.56786

Then, 966 positive and 1080 negative sentiment words are obtained in the tourism-specific sentiment lexicon in which 170 positive and 190 negative sentiment words also occurred in the detected new words. Several typical examples with highest positive and negative sentiment scores are shown in Table 3. Some sentiment words like “超级棒”(Very good) and “不负责任”(Irresponsible), are prevalent MWEs and are important to SA tasks. In addition, some sentiment words such as “美如画”(picturesque) and “嗨皮”(happy), are popular user-invented sentiment words that play an important role in SA. What’s more, sentiment words like “晚点”(delay) that convey special sentiments in the tourism domain are not included in traditional sentiment lexicons. Such converted words improve the accuracy of SA in the tourism domain.

Then, the lexicon is applied to subsequent SA tasks. A worked example in Fig. 4 is utilized to show this process.

Figure 4 A worked example of lexicon-based SA process

It is a positive tourism review that contains some sentiment words with sentiment scores. The words in red color belong to positive sentiment lexicon, and the words in green color belong to negative sentiment lexicon. The summation of sentiment scores implies that the final sentiment polarity of this tourism review is classified as positive.

The lexicon-based sentiment classification method [53] which contains two main ideas is changed to evaluate the proposed lexicon. First, sentiment polarities of the words within the scope of the negation words are inverted. Second, sentiment scores of all the words in a review are summarized to represent whole sentiment score of the review. If a review's sentiment score is greater than zero, then it is classified into positive. On the contrary, if a review's sentiment score is less than zero, then it is classified as negative one. What's more, if the score is zero, which means the review cannot be classified by the proposed method.

30180 reviews are selected from Qunar.com and Ctrip.com and fifteen students are invited to label these reviews. The tagged dataset called Dataset II, containing 26186 positive and 3994 negative reviews, stands for the real-world dataset. Dataset I is composed of 4001 positive and 3994 negative reviews since it requires an equal proportion of positive and negative samples for a classification test[7]. Detailed statistics of reviews of Dataset I and Dataset II are placed as follows:

Table 4: Statistics of datasets

Name	#Positive	#Negative	#Total	#Type
Dataset I	4001	3994	7995	Selected
Dataset II	26186	3994	30180	Real World

[7] Dataset I is generated from dataset II.

The experiments are performed on several pre-existing sentiment lexicons, state-of-the-art sentiment lexicon construction methods and *DWWP*, in order to compare the performances of them.

The first sentiment lexicon is HowNet, which is the abbreviation of the HowNet Sentiment Analysis Word Library[8], with 836 positive words and 1254 negative ones. The second sentiment lexicon is NTUSD, which is the abbreviation of NTU Sentiment Dictionary[9], with 2811 positive words and 8277 negative ones. The third sentiment lexicon is SWOL, the abbreviation of Sentiment Word Ontology Library[10], with 11229 positive words and 10783 negative ones. The fourth lexicon SentiRUC[11], constructed by Information Systems Engineering Laboratory, Renmin University of China, is composed of 6847 positive words and 7432 negative words [54]. These four lexicons are all traditional sentiment lexicons for formal SA tasks. Table 5 shows the information of these sentiment lexicons.

Table 5: Statistics of sentiment lexicons

Lexicon	#Positive	#Negative	#Total
NTUSD	2811	8277	11088
HowNet	836	1254	2090
SWOL	11229	10783	22012
SentiRUC	6847	7432	14279
Our lexicon	966	1080	2046

In addition, LP and GP are the abbreviations of label propagation [55] and graph propagation [53] respectively, both of which are the state-of-the-art sentiment lexicon construction methods. Similarity information used in these two methods is *PMIsimilarity* introduced in Formula 7. Seed sentiment words used in these two methods are the same as *DWWP*. What's more, some simplified forms of *DWWP* are also shown in this paper. Basic_method1 is a similar method without optimization and iteration compared with *DWWP* while Basic_method2 is a similar

[8] http://www.keenage.com/html/c_bulletin_2007.htm.
[9] http://nlg18.csie.ntu.edu.tw:8080/opinion/pub1.html.
[10] http://ir.dlut.edu.cn/news/detail/215.
[11] https://pan.baidu.com/s/1jHAInlG

method without iteration in comparison with *DWWP*. The experiment performances of all the methods on Dataset I and Dataset II are all listed in Table 6.

Here are formulas of statistics in Table 6. *TP* is *True Positive*; *FP* is *False Positive*; *TN* is *True Negative*, and *FN* is *False Negative*. The concepts of these four indices originate from the confusion matrix.

$$Positive\ precision = \frac{TP}{TP+FP} \quad \text{Formula 10}$$

$$Positive\ recall = \frac{TP}{P} = \frac{TP}{TP+FN} \quad \text{Formula 11}$$

Positive recall is also called *Sensitivity* in the confusion matrix analysis.

$$Negative\ precision = \frac{TN}{TN+FN} \quad \text{Formula 12}$$

$$Negative\ recall = \frac{TN}{N} = \frac{TN}{TN+FP} \quad \text{Formula 13}$$

$$F1\_P = \frac{2*Positive\ precision*Positive\ recall}{(Positive\ precision+Positive\ recall)} \quad \text{Formula 14}$$

$$F1\_N = \frac{2*Negative\ precision*Negative\ recall}{(Negative\ precision+Negative\ recall)} \quad \text{Formula 15}$$

P is the size of the positive sample set; N is the size of the negative sample set. $F1\_P$ ($F1\_N$) is the comprehensive index measuring the accuracy and exhaustiveness of classification results.

Table 6: WP experiment analysis[12] on Dataset I and II

Dataset	Method	Positive precision	Positive recall	Negative precision	Negative recall	F1_P	F1_N	Accuracy
I	HowNet	0.5045	0.6758	0.5078	0.3350	0.5777	0.4037	0.5056
	NTUSD	0.6112	0.7136	0.6552	0.5453	0.6584	0.5952	0.6295
	SWOL	0.5608	0.8198	0.6640	0.3568	0.6660	0.4642	0.5885
	SentiRUC	0.6023	0.8858	0.7835	0.4141	0.7170	0.5418	0.6502
	LP	0.5270	**0.9543**	0.7560	0.1420	0.6790	0.2391	0.5485
	GP	0.6446	0.8608	0.7900	0.5245	0.7372	0.6304	0.6928
	Basic_method1	**0.8394**	0.8938	0.8862	**0.8287**	**0.8658**	**0.8565**	**0.8613**
	Basic_method2	0.8252	0.9098	**0.8993**	0.8070	0.8654	0.8506	0.8584
	DWWP	**0.8351**	**0.9100**	**0.9010**	**0.8200**	**0.8709**	**0.8586**	**0.8650**
II	HowNet	0.8727	0.6953	0.1436	0.3350	0.7739	0.2010	----

[12] The bold font in a same column of Table 6 means that the values are Top 2 best performances among all the methods.

NTUSD	0.9115	0.7140	0.2253	0.5453	0.8008	0.3189	----
SWOL	0.8964	0.8484	0.2642	0.3568	0.8717	0.3036	----
SentiRUC	0.9079	0.8807	0.3461	0.4141	0.8941	0.3771	----
LP	0.8801	**0.9609**	0.3566	0.1420	0.9188	0.2031	----
GP	0.9235	0.8757	0.3916	0.5245	0.8990	0.4484	----
Basic_method1	**0.9716**	0.8929	**0.5413**	**0.8287**	**0.9306**	**0.6549**	----
Basic_method2	0.9682	0.8951	0.5400	0.8070	0.9302	0.6470	----
DWWP	**0.9704**	**0.9003**	**0.5563**	**0.8200**	**0.9340**	**0.6629**	----

As shown in Table 6, *DWWP* outperforms all the sentiment lexicons or construction methods in most cases; Basic_method1 does better than Basic_method2 in terms of most indices except *Positive recall*, *Negative precision*. Note that LP works well in the *Positive recall*, which indicates that LP is good at finding *TP* reviews from the original positive samples than other lexicons or methods. This advantage is more significant when LP is applied to Dataset II, which contains almost nine times positive samples as many as negative ones. While the result also demonstrates that LP works worst in the *Negative recall,* that is, LP selects much fewer *TN* reviews out of the original negative samples than other lexicons or methods. Therefore, LP works worst in $F1\_N$, which indicates LP is not competent enough to process negative samples.

Table 6 also shows that data-driven methods, such as LP, GP and the proposed methods, get better performances than traditional sentiment lexicons on the tourism-specific SA tasks. This is because traditional sentiment lexicons do not contain user-invented words and converted words. Words such as “嗨皮”(happy), “强买强卖”(hard sale) and “晚点”(delay), very common in tourism reviews, are not included in traditional sentiment lexicons. Moreover, these lexicons cannot provide rigorous sentiment scores to differentiate the degrees of sentiment polarity.

Although Basic_method2 gets better performance than Basic_method1 in finding out positive samples (*Positive recall*) and maintaining the accuracy of classified negative samples (*Negative precision*), its other indices are rather worse than Basic_method1 and *DWWP*. Nevertheless, it is still better than LP, GP and traditional lexicons, which demonstrates that the proposed word

propagation algorithm based on semantic similarity is quite suitable for scoring sentimental words.

In addition, *DWWP* outperforms Basic_method2, which proves that scoring sentimental words with iterations is better than appointing scores to unlabeled words by solving optimization function all at once. Also, it is confirmed that solving the optimization function iteratively is necessary for sentiment lexicon construction since *DWWP* works better than Basic_method1 in most cases. This demonstrates that scoring sentimental words iteratively based on semantic similarity, PMIsimilarity, and L_2-norm regularization attains better performance. It can be explained as follows: manually selecting and scoring seed words cannot guarantee the representativeness and accuracy of seed sentiment words; there is no guarantee that the selected seed words are the closest (most similar) ones to unlabeled words either. In this case, if seed words are used to appoint sentiment scores to unlabeled words via semantic similarity, and extend seed word set iteratively, then the sentiment lexicon with rigorous scores is obtained in a more objective and accurate manner.

Table 7: The proposed methods vs. state-of-the-art methods

	manova1	anova1_PP	anova1_PR	anova1_NP	anova1_NR	anova1_F1P	anova1_F1N
DatasetI	**0.0043**	**8.92E-05**	0.2163	**0.0163**	**0.0017**	**6.22E-04**	**0.0033**
DatasetII	**0.0027**	**4.75E-04**	0.3159	**0.0025**	**0.0017**	0.0795	**5.63E-04**

Table 8: The proposed methods vs. data-driven methods

	manova1	anova1_PP	anova1_PR	anova1_NP	anova1_NR
DatasetI	**4.31E-06**	**1.12E-02**	0.9377	**0.0032**	**0.0425**
DatasetII	**6.18E-06**	**0.0245**	0.5363	**0.0013**	**0.0425**

Annova test is applied to estimate whether the performances of the proposed methods are better than all the state-of-the-art methods or data-driven methods LP and GP. The results in

Table 7 and Table 8 indicate that the null hypothesis should be denied in most cases[13] (the bold). In other words, the proposed methods are dramatically better than other methods.

4.3 Parameter Analysis

As is shown in the optimization function, different values of α, β and λ determine the relative importance of initial sentiment scores, semantic similarity scores, PMIsimilarity scores and L_2-norm regularization in *WP*. Therefore in this subsection, the influence of these parameters on sentiment lexicon performance is discussed. Since the parameter analysis result on Dataset II are similar to it is on Dataset I, detailed analysis is only performed on Dataset I.

In Fig. 5, as α increases from 0 to 1, so do *F1_P (F1_N)* score and *Accuracy* in the overall trend[14]. This indicates that semantic similarity is more important in *WP* relative to statistical similarity. As a rule of thumb, it is suggested to set α in the range of 0.8-1.0 as a qualified value for the proposed model.

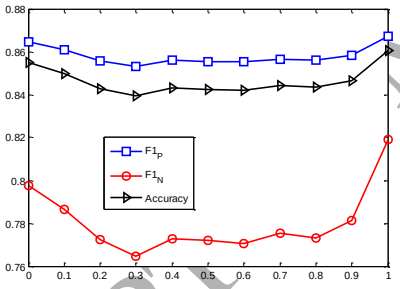


Figure 5. Parameter analysis of $\boldsymbol{\alpha}$

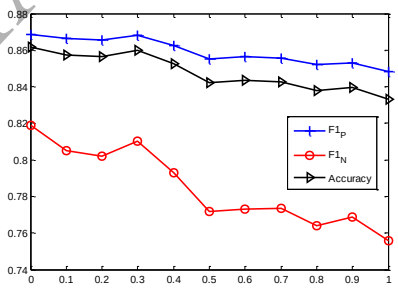


Figure 6. Parameter analysis of $\boldsymbol{\beta}$

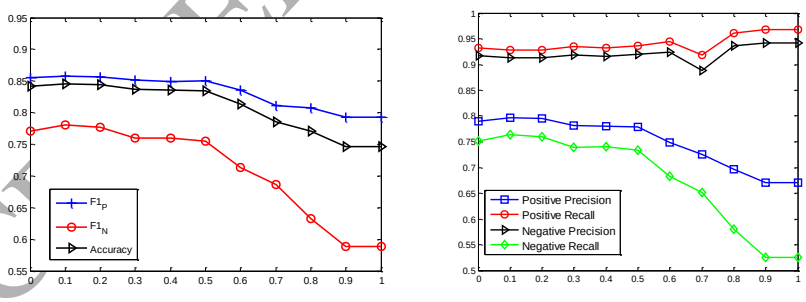
Similarly, in Fig. 6, as β changes from 0 to 1.0, the *F1_P (F1_N)* score and *Accuracy* increase a bit at first and reach the maximum when $\beta = 0.3$, then the indices decrease. This demonstrates statistical similarity is relatively less important in the proposed model, matching the analysis of α in Fig. 5. Moderate value of β ranges from 0.1 to 0.3. This can be explained in a more practical sense that statistical similarity, mainly considering the co-occurrence frequency of a word pair, is

[13] The P-value is set to 0.05

[14] Fig. 5 shows that negative samples are more sensitive to parameter α.

less accurate than semantic similarity in *WP*. That is because statistical similarity is suitable for processing synonym co-occurring in the same document. However, there are still quantities of words with similar sentiment scores that do not occur in one review due to different personal expression habits.

(a) in Fig. 7 shows that as the value of λ increases from 0 to 1, *F1_P (F1_N)* and *Accuracy* increase at first, especially when λ changes from 0 to 0.2, then decrease a little and almost maintain the similar values from 0.3 to 0.5. When λ changes from 0.5 to 1, all the indices decline sharply, especially *F1_P (F1_N)*. The rapid decrease of *F1* scores originates from the sharp decline of *Negative recall* as is shown in (b). *F1* scores are related to *Negative precision* and *Negative recall*. Although *Negative precision* increases as λ increases, the *Negative recall* decreases faster, causing the rapid decline of *F1* scores in the overall trend. Moreover, (b) indicates that *Positive precision* shows a variation pattern similar to *Negative recall* while *Positive recall* has a similar change tendency with *Negative precision*. The performance of the proposed model on negative samples is more sensitive to the change of λ relative to that on positive ones.



(a) (b)

Figure 7. The performance of *DWWP* with different values of $\boldsymbol{\lambda}$

There are some practical explanations about the results in Fig. 7. Increasing of λ from 0 to 0.3 indicates that L_2-norm regularization serves an important role in the optimization function, which leads to less outliers since L_2-norm regularization forces absolute value of sentiment scores to get

smaller in practice. This guarantees the smoothness of *WP* process. However, as λ increases from 0.5 to 1, L_2-norm regularization plays a more important role in the construction of sentiment lexicon, compelling sentiment scores to vary around 0. This implies that the dispersion of absolute value of sentiment scores becomes smaller. It should be noted that the relative amount of positive and negative sentiment words in a review determines the final sentiment polarity of a sample, whereas some important words with extreme values do not work anymore in this context. It is even more severe for classifying a negative sample because Chinese people usually put up some praises before the complaints. When λ increases from 0.5 to 1, L_2-norm regularization may force the word with greater sentiment degree to have a quite smaller absolute value of sentiment score. In this case, a negative sample may be classified as positive one since the number of positive words is more than negative ones and the sentiment scores of negative words do not match their real sentiment degrees. So the indices of negative samples in Fig. 7 are more sensitive to the change of λ. As λ increases from 0.5 to1, it is more difficult to find *TN* samples and more negative samples are classified incorrectly, so the *Negative recall* decreases; Meanwhile, because of the difficulty in finding *TN* samples, once a sample is classified as a negative one, it is less possible to be a wrong judgment, so the *Negative precision* increases. In turn, as λ increases from 0.5 to1, since more negative samples are classified into positive ones, *Positive precision* will decrease; and more difficulty in finding negative ones means it is relative more easy to find out positive samples, so *Positive recall* will increase. Furthermore, the figure indicates that moderate range of λ in *DWWP* is 0-0.3.

5. Conclusion

The proposed SA system *DWWP* mainly consisted of two parts, *DW* (domain new words detection) and *WP* (word propagation). The system was aimed at performing SA tasks through a high-quality domain-specific sentiment lexicon. *DW* made it possible to detect user-invented

words, proper nouns, converted words and MWEs in the tourism domain. The basic idea of *DW* lay in an indicator *AMI* mentioned in subsection 3.1. The indicator took use of statistical information of candidate new words and some necessary tricks according to Chinese grammars to select suitable new words. *WP* achieved the goal to build the high-quality tourism-specific sentiment lexicon, where the main idea was a data-driven iterative algorithm combining optimization function and machine learning tool word2vec. Seed words, semantic and statistical similarity help the algorithm get rigorous scores of sentiment words. Before that, *DW* contributed a lot to *WP*, especially in training the word2vec model.

Experimental results demonstrated that *DWWP* significantly outperformed traditional sentiment lexicons, two state-of-the-art methods LP and GP and two simplified forms of *DWWP*. More concretely, *DWWP* improved seventeen percentage points compared with GP and four percentage points compared with LP in accuracy on Dataset I and Dataset II, respectively.

The system can also be applied to hot scenic spots analysis and precision marketing. In future research, we will focus on improving the performance of the domain-specific sentiment lexicon, where a traditional sentiment lexicon combined with machine learning models like *RNN* and *LSTM* would be given more attention. What's more, The powerful meta-level features are available for improving the performance of sentiment polarity classification [56]. Besides, we plan to propose an effective sorting algorithm based on *DWWP*, picking up tourism reviews that contain more details and hot words. It will promote online travel and help tourists make better decisions.

Acknowledgement

This research is supported by the National Natural Science Foundation of China No. 91546201, No. 71331005 and No. 71501175, Shandong Independent Innovation and Achievement Transformation Special Fund of China (2014ZZCX03302), and the Open Project of Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences.

We would also like to thank the anonymous reviewers for their helpful comments.

Reference

[1] E. Marine-Roig, Online travel reviews: A massive paratextual analysis, in: Anal. Smart Tour. Des. Concepts Methods, 2017: pp. 179–202. doi:10.1007/978-3-319-44263-1_11.

[2] V. Baka, The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector, Tour. Manag. 53 (2016) 148–162. doi:10.1016/j.tourman.2015.09.004.

[3] K. Kim, O. Park, S. Yun, H. Yun, What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management, Technol. Forecast. Soc. Change. (2017). doi:http://dx.doi.org/10.1016/j.techfore.2017.01.001.

[4] M. Olmedilla, M.R. Martinez-Torres, S.L. Toral, Examining the power-law distribution among eWOM communities: a characterisation approach of the Long Tail, Technol. Anal. Strateg. Manag. 28 (2016) 601–613. doi:10.1080/09537325.2015.1122187.

[5] M. González-Rodríguez, Post-visit and pre-visit tourist destination image through eWOM sentiment analysis and perceived helpfulness, Int. J. (2016). doi:10.1108/IJCHM-02-2015-0057.

[6] P. Nakov, A. Ritter, S. Rosenthal, V. Stoyanov, F. Sebastiani, SemEval-2016 Task4: Sentiment Analysis in Twitter, Proc. 10th Int. Work. Semant. Eval. (2016) 1–18.

[7] A. Hogenboom, B. Heerschop, F. Frasincar, U. Kaymak, F. De Jong, Multi-lingual support for lexicon-based sentiment analysis guided by semantics, Decis. Support Syst. 62 (2014) 43–53. doi:10.1016/j.dss.2014.03.004.

[8] G. Wang, J. Sun, J. Ma, K. Xu, J. Gu, Sentiment classification: The contribution of ensemble learning, Decis. Support Syst. 57 (2014) 77–93. doi:10.1016/j.dss.2013.08.002.

[9] F.H. Khan, U. Qamar, S. Bashir, eSAP: A decision support framework for enhanced sentiment analysis and polarity classification, Inf. Sci. (Ny). 367–368 (2016) 862–873. doi:10.1016/j.ins.2016.07.028.

[10] M.R. Gonzalez-Rodriguez, M.R. Martinez-Torres, S.L. Toral, Monitoring travel-related information on social media through sentiment analysis, Proc. - 2014 IEEE/ACM 7th Int. Conf. Util. Cloud Comput. UCC 2014. (2014) 636–641. doi:10.1109/UCC.2014.102.

[11] F.Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, CEUR Workshop Proc. 718 (2011) 93–98.

[12] F. Xianghua, L. Guo, G. Yanyan, W. Zhiqiang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon, Knowledge-Based Syst. 37 (2013) 186–195. doi:10.1016/j.knosys.2012.08.003.

[13] M. Huang, B. Ye, Y. Wang, H. Chen, J. Cheng, X. Zhu, New Word Detection for Sentiment Analysis, Acl. (2014) 531–541.

[14] F. Wu, Y. Huang, Y. Song, S. Liu, Towards building a high-quality microblog-specific Chinese sentiment lexicon, Decis. Support Syst. 87 (2015) 39–49. doi:10.1016/j.dss.2016.04.007.

[15] X. Sun, H. Wang, W. Li, Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection, Acl. (2012) 253–262. http://dl.acm.org/citation.cfm?id=2390560.

[16] K.-J. Chen, W.-Y. Ma, Unknown word extraction for Chinese documents, Proc. 19th Int. Conf. Comput. Linguist. -. (2002) 1–7. doi:10.3115/1072228.1072277.

[17] H. Li, C.N. Huang, J. Gao, X. Fan, The Use of SVM for Chinese New Word Identification, Nat. Lang. Process. 2004. (2004) 723–732. http://www.springerlink.com/index/J676WDPWV9UY6JNL.pdf.

[18] K.W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Comput. Linguist. 16 (1990) 22–29. doi:10.3115/981623.981633.
[19] F.B.X. Zhu., Measuring the Non-compositionality of Multiword Expressions, in: Proc. 23rd Int. Conf. Comput. Linguist. (Coling 2010), 2010: pp. 116–124.
[20] W. Zhang, T. Yoshida, X. Tang, T.B. Ho, Improving effectiveness of mutual information for substantival multiword expression extraction, Expert Syst. Appl. 36 (2009) 10919–10930. doi:10.1016/j.eswa.2009.02.026.
[21] C.N. dos Santos, M. Gatti, Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts, Coling-2014. (2014) 69–78.
[22] T. Khalil, Samhaa R. El-Beltagy, NileTMRG: Deep Convolutional Neural Networks for Aspect Category and Sentiment Extraction in SemEval-2016 Task 5, Proc. 10th Int. Work. Semant. Eval. (SemEval 2016) - to Appear. (2016) 276–281.
[23] K.S. Tai, R. Socher, C.D. Manning, Improved semantic representations from tree-structured long short-term memory networks, Proc. ACL. (2015) 1556–1566. doi:10.1515/popets-2015-0023.
[24] F.H. Khan, U. Qamar, S. Bashir, SWIMS: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis, Knowledge-Based Syst. 100 (2015) 97–111. doi:10.1016/j.knosys.2016.02.011.
[25] F.H. Khan, U. Qamar, S. Bashir, Lexicon based semantic detection of sentiments using expected likelihood estimate smoothed odds ratio, Artif. Intell. Rev. (2016) 1–26. doi:10.1007/s10462-016-9496-4.
[26] E. Cambria, N. Howard, Computational Intelligence for Big Social Data Analysis, (2016) 8–9.
[27] et al. Cambria E, Schuller B, Xia Y, New avenues in knowledge bases for natural language processing, Knowledge-Based Syst. (2016) 1–4. doi:10.1016/j.knosys.2016.07.025.
[28] E.Cambria, Affective Computing and Sentiment Analysis, IEEE Intell. Syst. 31 (2016) 102–107. doi:doi: 10.1109/MIS.2016.31.
[29] S. Poria, E. Cambria, A. Gelbukh, Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis, Proc. 2015 Conf. Empir. Methods Nat. Lang. Process. (2015) 2539–2544. http://aclweb.org/anthology/D15-1303.
[30] I. Chaturvedi, E. Cambria, Bayesian Deep Convolution Belief Networks for Subjectivity Detection, (n.d.).
[31] S. Poria, E. Cambria, A. Gelbukh, Aspect extraction for opinion mining with a deep convolutional neural network, 108 (2016) 42–49. doi:10.1016/j.knosys.2016.06.009.
[32] S. Poria, E. Cambria, D. Hazarika, P. Vij, A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks, (2016) 1601–1612.
[33] S. Poria, I. Chaturvedi, E. Cambria, Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis, (2016).
[34] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (2014) 1093–1113. doi:10.1016/j.asej.2014.04.011.
[35] S.R. Das, M.Y. Chen, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, Manage. Sci. 53 (2007) 1375–1388. doi:10.1287/mnsc.1070.0704.
[36] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller, Introduction to wordnet: An on-line lexical database, Int. J. Lexicogr. 3 (1990) 235–244. doi:10.1093/ijl/3.4.235.
[37] C. Strapparava, A. Valitutti, WordNet-Affect: an affective extension of WordNet, Proc. 4th Int. Conf. Lang. Resour. Eval. (2004) 1083–1086. doi:10.1.1.122.4281.

[38] S.D. Richardson, W.B. Dolan, L. Vanderwende, MindNet: acquiring and structuring semantic information from text, in: COLING-ACL'98 Meet. Assoc. Comput. Linguist., 1998: pp. 1098–1102. doi:10.3115/980691.980749.
[39] L. Ku, Y. Liang, H. Chen, K. Lun-Wei, L. Yu-Ting, C. Hsin-Hsi, Opinion Extraction, Summarization and Tracking in News and Blog Corpora, Artif. Intell. pages (2006) 100–107. doi:citeulike-article-id:2913694.
[40] H. Xu, K. Zhao, L. Qiu, C. Hu, Expanding Chinese sentiment dictionaries from large scale unlabeled corpus, Proc. PACLIC 24. (2010) 301–310.
[41] D. Tang, F. Wei, B. Qin, M. Zhou, T. Liu, Building Large-Scale Twitter-Specific Sentiment Lexicon: a Representation Learning Approach, Proc. 25th Int. Conf. Comput. Linguist. (COLING 2014). (2014) 172–182.
[42] H. Saif, M. Fernandez, Contextual Semantics for Sentiment Analysis of Twitter Categories and Subject Descriptors, 52 (2016) 5–19.
[43] S. Feng, L. Wang, W. Xu, D. Wang, G. Yu, Unsupervised learning Chinese sentiment lexicon from massive microblog data, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2012: pp. 27–38. doi:10.1007/978-3-642-35527-1_3.
[44] H. Cho, S. Kim, J. Lee, J.S. Lee, Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews, Knowledge-Based Syst. 71 (2014) 61–71. doi:10.1016/j.knosys.2014.06.001.
[45] F. Bravo-Marquez, E. Frank, B. Pfahringer, Building a Twitter opinion lexicon from automatically-annotated tweets, Knowledge-Based Syst. 108 (2016) 65–78. doi:10.1016/j.knosys.2016.05.018.
[46] Y. Choueka, Looking for needles in a haystack or locating interesting collocational expressions in large textual databases, in: Proc. 2nd Int. Conf. Comput. Inf. Retr. (RIA '88), 1988: pp. 609–623. http://cat.inist.fr/?aModele=afficheN&cpsidt=7104181.
[47] R. Jackendoff, S. Stevenson, The Architecture of the Language Faculty, Comput. Linguist. 24 (1997) 652–655. doi:10.2307/417010.
[48] C. Fellbaum, WordNet: An Electronic Lexical Database, 1998. doi:10.1139/h11-025.
[49] T. Mikolov, G. Corrado, K. Chen, J. Dean, Efficient Estimation of Word Representations in Vector Space, Proc. Int. Conf. Learn. Represent. (ICLR 2013). (2013) 1–12. doi:10.1162/153244303322533223.
[50] S.M. Choi, S.K. Ko, Y.S. Han, A movie recommendation algorithm based on genre correlations, Expert Syst. Appl. 39 (2012) 8079–8085. doi:10.1016/j.eswa.2012.01.132.
[51] R. Tibshirani, Regression Selection and Shrinkage via the Lasso, J. R. Stat. Soc. B. 58 (1996) 267–288. doi:10.2307/2346178.
[52] A. Wächter, L.T. Biegler, On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming, Math. Program. 106 (2006) 25–57. doi:10.1007/BF01582568.
[53] L. Velikovich, S.B. Kerry, H. Ryan, The viability of web-derived polarity lexicons, Naacl. (2010) 777–785. doi:10.1056/NEJM199902113400601.
[54] X. Yang, Z. Zhang, Z. Zhang, Y. Mo, L. Li, L. Yu, P. Zhu, Automatic Construction and Global Optimization of A Multi-Sentiment Lexicon, 2016 (2016).
[55] D. Rao, D. Ravichandran, Semi-supervised polarity lexicon induction, EACL '09 Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguist. (2009) 675–682. doi:10.3115/1609067.1609142.
[56] F. Bravo-Marquez, M. Mendoza, B. Poblete, Meta-level sentiment models for big social data analysis, Knowledge-Based Syst. 69 (2014) 86–99. doi:10.1016/j.knosys.2014.05.016.