Knowledge-Based Systems 41 (2013) 89–97

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news

Liang-Chih Yu, Jheng-Long Wu, Pei-Chann Chang *, Hsuan-Shou Chu

Department of Information Management, Yuan Ze University, Chung-Li, Taiwan, ROC

ARTICLE INFO

Article history:
Received 7 August 2012
Received in revised form 30 December 2012
Accepted 2 January 2013
Available online 11 January 2013

Keywords:
Sentiment classification
Contextual entropy model
Word expansion
Natural language processing
Stock market news

ABSTRACT

Sentiment classification of stock market news involves identifying positive and negative news articles, and is an emerging technique for making stock trend predictions which can facilitate investor decision making. In this paper, we propose the presence and intensity of emotion words as features to classify the sentiment of stock market news articles. To identify such words and their intensity, a *contextual entropy model* is developed to expand a set of seed words generated from a small corpus of stock market news articles with sentiment annotation. The contextual entropy model measures the similarity between two words by comparing their contextual distributions using an entropy measure, allowing for the discovery of words similar to the seed words. Experimental results show that the proposed method can discover more useful emotion words and their corresponding intensity, thus improving classification performance. Performance was further improved by the incorporation of intensity into the classification, and the proposed method outperformed the previously-proposed pointwise mutual information (PMI)-based expansion methods.

1. Introduction

Stock trend prediction using technical indices such as moving averages (MAs) and relative strength indices (RSIs) has been extensively investigated [1–4]. Textual data, such as stock market news articles, may affect investor decisions, and is thus another important factor affecting share prices [5–8]. For example, positive news may encourage investors to buy shares, thus forcing the share price up, while negative news would have the opposite effect. However, the sheer volume of such news articles makes it difficult for investors to find such useful information from daily news sources. Sentiment classification of stock market news to identify positive and negative articles has thus emerged as a promising technique for stock trend prediction. This approach can also help investors identify sentimental tendencies in financial news and facilitate investment decision making. Accordingly, this study focuses on mining useful features to classify the sentiment of stock market news.

Stock market news articles are characterized by a specific set of emotion words which vary in intensity. Consider the sample passages presented in (1)–(4).

(1) Stocks **soar** on wings of hope. US stocks rallied Wednesday, with the Dow industrials and S&P 500 gaining the most for 2012, on hopes the world’s central bankers will move to bolster the global economy. (Star Tribune)

(2) US stocks **rise** most since December on stimulus hopes. The Dow Jones Industrial Average jumped 286.84 points, or 2.4%, to 12414.79, and the euro rose 1% to trade well above $1.25. (The Wall Street Journal)

(3) Fears of European **collapse** sink markets across globe. Fearing a financial rupture in Europe, investors around the world fled from risk Wednesday. (Kennebec Journal)

(4) Nokia’s share of mobile phone sales **fell** to 21% in the first quarter from 27% a year earlier, according to market data from IDC. Its share peaked at 40.4% at the end of 2007. (The Wall Street Journal)

Both examples (1) and (2) contain positive emotion words such as “soar” and “rise”, while examples (3) and (4) contain negative emotion words such as “collapse” and “fall”. Such emotion words are useful features for classifying the sentiment of stock market news [9,10]. Generally, news articles tend to be classified as having positive sentiment if they contain more positive emotion words, and vice versa. The intensity of such words is also useful for sentiment classification. For example, although “soar” and “rise” are positive, the intensity level of “soar” (e.g., 0.9) is greater than that of “rise” (e.g., 0.5).[1] Similarly, the intensity level of “collapse” is greater

* Corresponding author. Tel.: +886 3 463 8800x2305; fax: +886 3 435 2077.
E-mail address: iepchang@saturn.yzu.edu.tw (P.-C. Chang).

[1] The intensity levels in this example are assigned manually between 0 and 1 to help distinguish the sample words “soar” and “rise”.

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.01.001

than that of “fall”. Taking intensity into account is especially useful for classifying news articles containing similar number of positive and negative emotion words. Therefore, the proposed method uses both the presence of emotion words and their intensity as features for sentiment classification.

Several resources exist for obtaining emotion words and their intensity, including ANEW [11], WordNet-Affect [12], SentiWordnet [13], SentiFul [14], EmotiNet [15], and several Chinese sentiment lexicons [16–20]. These lexicons have been used for many applications such as hotspot detection and forecasting [21], detecting critical situations in online markets [22], and identifying depressive symptoms and negative life events from psychiatric texts [23,24]. However, lexicon-based approaches usually have limited word coverage, and thus may fail to recognize emotion words (especially domain specific words) not already defined in the lexicon due to various domain applications. For instance, the sample emotion words *rise*, *soar* and *collapse* in the sample passages above (1)–(4) commonly occur in the financial domain but are not included in ANEW [11]. Therefore, recent studies have investigated the use of corpus-based approaches [25–28] to automatically discover new emotion words with their corresponding intensity from large corpora based on a small set of seed words derived from existing emotion lexicons. Turney and Littman [25] proposed an automated system using pointwise mutual information (PMI) to infer the intensity of a word from its association from a set of paradigm words, and suggested various applications that could benefit from the use of sentiment analysis techniques. In calculating PMI scores, the co-occurrence frequencies between two words were retrieved from three different corpora using the AltaVista Advanced search engine with both NEAR and AND operators. Baroni and Vegnaduzzo [26] used PMI to acquire new subjective adjectives and their subjective scores by searching for adjectives that tend to co-occur with a small set of manually selected seed adjectives. Starting from an existing affect lexicon, Grefenstette et al. [27] further used PMI to identify new emotion words and their intensity from web search results to augment the lexicon. Abbasi et al. [28] evaluated lexicon features and PMI for sentiment analysis of online forums and blogs. The PMI used in this word expansion process was considered as a similarity measure to select words frequently co-occurring with the seed words as similar words for expansion.

This study proposes a *contextual entropy model* to expand a set of seed words by discovering more emotion words with their corresponding intensity from online stock market news articles. Similar to PMI, the contextual entropy model also adopts a similarity measure for emotion word selection, but the computation of word similarity is different. For PMI, the similarity between words is calculated based on the co-occurrence frequency. That is, two words which co-occur more frequently are more similar. However, this may also induce noisy words that frequently co-occur but are dissimilar from the seed words because frequent co-occurrence does not necessarily indicate semantic similarity. For the sample words shown in Fig. 1, the gray shadows represent words which frequently occur in the context of the words *boss*, *chief*, and *flower*. In this example, *stress*, *colleague*, and *client* frequently co-occurred with *boss* (or *chief*), thus they might be expanded by PMI when given *boss* (or *chief*) as the seed word, but not all of them are similar to *boss* (or *chief*).

Instead of directly using co-occurrence frequencies, the contextual entropy model measures the similarity between two words by comparing their contextual distributions using relative entropy [29]. This can be accomplished by transforming the co-occurrence frequencies of context words into probabilistic representations such that each word can be represented as a probabilistic distribution of its context words. Consider the example presented in Fig. 1: each row represents the vector representation of a word and its context words, and each entry in the vector represents the probability of a context word co-occurring with the target word. Based on this representation, the relative entropy can then be used to calculate the difference between the contextual distributions of two words to measure their similarity. That is, two words sharing more common contexts are more semantically similar because they have a similar contextual distribution, i.e., the difference (relative entropy) between their contextual distributions is small. Fig. 1 shows that *boss* and *chief* have similar contextual distributions, but are quite different from the contextual distribution of *flower*. Accordingly, with the input of *boss* (or *chief*), the words *chief* (or *boss*) will be expanded by the contextual entropy model, but the potential noisy word *flower* will not. This example shows the novelty of the proposed contextual entropy model—it considers both co-occurrence strength and contextual distribution to acquire more useful emotion words while better avoiding noisy words than PMI, which considers the co-occurrence strength alone. Once the words are expanded, the PMI scores and entropy scores are respectively considered to be the intensity levels of the words expanded by the two methods. Experiments were conducted to evaluate both methods to determine whether or not the expanded emotion words and their intensity could improve sentiment classification of stock market news, and to determine which expansion method could better generate emotion words and intensity for classification.

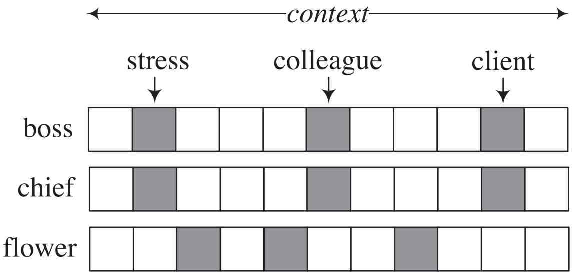

Fig. 1. Example of context distribution.

The rest of this paper is organized as follows. Section 2 describes the overall framework of emotion word expansion for sentiment classification. Section 3 describes the proposed contextual entropy model and previously proposed PMI-based method for expansion of emotion words and their intensity. Section 4 summarizes experimental results. Conclusions are finally drawn in Section 5.

2. System framework

The overall framework, as shown in Fig. 2, is divided into three parts: seed word selection, emotion word expansion and sentiment classification. A set of stock market news articles is first collected and manually annotated with positive or negative sentiments. These labeled news articles are then subjected to information gain (IG) [30,31], a technique commonly used for feature selection, to select a set of representative positive and negative emotion words as the seed words. Once the seed words are selected, they are passed to the next step for emotion word expansion. First, for each seed word, all content words (i.e., excluding stop words) with the same part-of-speech (POS) as the seed word in the unlabeled news articles are considered as candidates for expansion. Then, as shown in Fig. 2a, the proposed contextual entropy model is used to discover more emotion words similar to the seed words from among the candidates. The expanded words are associated with entropy scores as their intensity levels and then combined with the seed words to constitute a sentiment lexicon. Fig. 2b shows three validation methods provided for comparison with the proposed

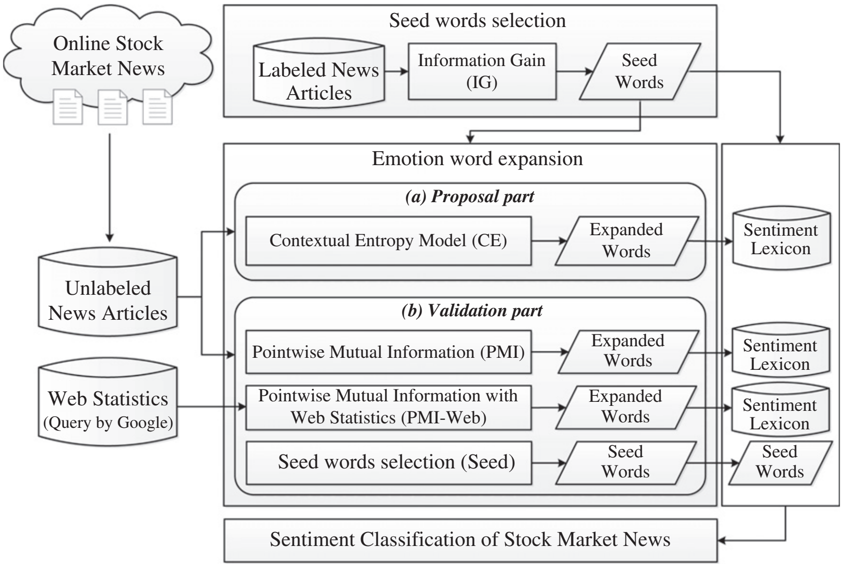

Fig. 2. Framework of emotion word expansion for the sentiment classification of stock market news. (a) Proposal part and (b) validation part.

method. For the PMI-based methods (i.e., PMI and PMI-Web) these validation methods use a different corpus for emotion word expansion. PMI uses a set of unlabeled online stock market news articles to retrieve the co-occurrence frequencies between seed words and candidate words, while PMI-Web retrieves them from web statistics by querying Google, i.e., taking the frequency returned by Google with any two words as a query. The expanded words are associated with PMI (or PMI-Web) scores as their intensity levels, and then combined with the seed words to constitute a sentiment lexicon. The method denoted as Seed represents the use of seed words alone for sentiment classification. Finally, the sentiment lexicon built from each method is used to classify each stock market news article as positive or negative for performance evaluation. In addition, comparing Seed to the expansion methods helps determine whether or not the expanded emotion words and their intensity could improve performance of sentiment classification, and comparing the PMI-based methods to the proposed method helps determine which expansion method best contributes to the classification task.

3. Emotion word expansion for sentiment classification

This section provides a detailed description of the seed word selection, emotion word expansion, and sentiment classification of stock market news.

3.1. Seed word selection

Given a corpus of annotated stock market news articles, the emotion seed words are selected using information gain [30,31]. Each distinct content word in the labeled corpus is considered as a candidate, and associated with an IG score representing its correlation to the positive and negative sentiment classes. These candidate words are then sorted in descending order of their IG scores. To ensure the quality of the seed words, the ranked list was presented to an annotator (a graduate student with a background in information management) to manually select representative positive and negative emotion words as the seed words. Once selection, the intensity levels of the positive and negative seed words are set to 1 and -1, respectively.

3.2. Expansion of emotion words and intensity

Once the seed words are generated, the next step is to acquire more emotion words and their intensity from an unlabeled corpus. The following subsections describe the two expansion methods: the proposed contextual entropy model and the validation method PMI used for comparison.

3.2.1. Proposed part—contextual entropy model (CE)

The proposed contextual entropy model considers both co-occurrence strength and contextual distribution between the candidate words and seed words to discover similar emotion words and their corresponding intensity from the unlabeled corpus. The contextual entropy model first uses a vector representation to represent the co-occurrence strength between the seed words and candidate words and their respective context words. The similarity between seed words and candidate words can then be calculated by comparing their vector representation of contextual distributions using relative entropy. Finally, a set of similar words and their corresponding intensities are selected for expansion. Below we describe the procedures for vector representation, similarity measurement, and expansion in the contextual entropy model.

3.2.1.1. Vector representation. The contextual entropy model uses a high-dimensional vector to record the co-occurrence strength between a word and its context words. In addition, a word has a left and right context depending on the context words preceding or following it in sentences. For instance, the left and right context of a word w_k in a sentence $W = d_1 \cdots d_{k-1}\ w_k\ d_{k+1} \cdots d_n$ are $\{d_1, \ldots, d_{k-1}\}$ and $\{d_{k+1}, \ldots, d_n\}$, respectively. Therefore, each seed word and candidate word can be represented as a pair of vectors of its left and right contexts. That is

$$\begin{aligned} context(w_k) &= \left(v^{left}_{w_k}, v^{right}_{w_k}\right) \\ &= \left(\left\langle m^{left}_{w_k d_1}, m^{left}_{w_k d_2}, \ldots, m^{left}_{w_k d_N}\right\rangle, \left\langle m^{right}_{w_k d_1}, m^{right}_{w_k d_2}, \ldots, m^{right}_{w_k d_N}\right\rangle\right), \end{aligned} \tag{1}$$

where $v^{left}_{w_k}$ and $v^{right}_{w_k}$ respectively denote the left and right context vectors of w_k; $m_{w_k d_i}$ denotes the weight of the ith dimension of a vector, representing the co-occurrence strength between w_k and its context word d_i, and N denotes the dimensionality of a vector, i.e., the number of distinct words appearing in the context of w_k in the unlabeled corpus. Fig. 3 shows the conceptual representation context information, where each row represents a word's left or right context vector.

Similar to the Hyperspace Analog to Language model [32–34], the weight $m_{w_k d_i}$ is calculated by considering both the location and distance of words in sentences. That is, a word in a context that is closer to the target word is given a greater weight. The proposed approach adopts an observation window of length ℓ over the corpus to calculate the weights. All words within the window are considered as co-occurring with each other. Thus, the weight between any two words of distance d within a window is calculated as $\ell - d + 1$. For example, given the sentence $W = \ldots d_1 d_2 d_3$ w_k $d_4 d_5 d_6 \ldots$, and a window size $\ell = 3$, the weight between d_4 and w_k is 3 – 1 + 1 = 3, i.e., $m_{w_k d_4} = 3$. Similarly, $m_{w_k d_3} = 3$, $m_{w_k d_2} = m_{w_k d_5} = 2$, and $m_{w_k d_1} = m_{w_k d_6} = 1$.

Once the weights are calculated, we use a weighting scheme analogous to TF-IDF [35–38] to re-weight the dimensions of a vector, defined as

$$m_{w_k d_i} = m_{w_k d_i} \times \log \frac{N(V)}{N(V_{d_i})}, \tag{2}$$

where $N(V)$ denotes the total number of vectors and $N(V_{d_i})$ denotes the number of vectors with d_i as the dimension. Based on the above re-weighting scheme, the dimension words appearing in more vectors will be de-emphasized because such words usually carry less information to discriminate among the vectors, and those appearing in fewer vectors will be emphasized because such words are usually more informative.

To further transform the vector representation of words into probabilistic distributions to calculate their similarity, the weights are also transformed into probabilistic representations, defined as

$$m_{w_k d_i} \equiv P(d_i|w_k) = \frac{m_{w_k d_i}}{\sum_i m_{w_k d_i}}, \tag{3}$$

where $P(d_i|w_k)$ denotes the probability that d_i appears in the context of w_k.

3.2.1.2. Similarity measure. The previous step describes how each seed word and candidate word is transformed into a probabilistic vector representation of their context distributions. The present step introduces the use of the *Kullback–Leibler (KL) distance* [39,40] to measure the difference between the probabilistic context distributions of a seed word and a candidate word. Let $c_i = \left(v_{c_i}^{left}, v_{c_i}^{right}\right)$ and $seed_j = \left(v_{seed_j}^{left}, v_{seed_j}^{right}\right)$ respectively be the vector representation of a candidate and a seed word. The KL divergence between their left (or right) contextual distributions, i.e., $D\left(v_{c_i} \| v_{seed_j}\right)$, is defined as

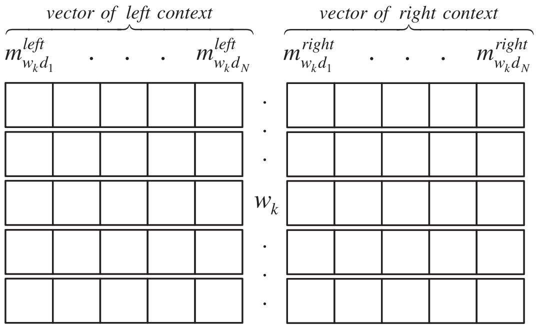
vector of left context
vector of right context
$m^{left}_{w_k d_1}$ . . . $m^{left}_{w_k d_N}$
$m^{right}_{w_k d_1}$ . . . $m^{right}_{w_k d_N}$
w_k

Fig. 3. Vector representation of contextual information.

$$D(v_{c_i} \| v_{seed_j}) = \sum_{k=1}^{N} P(d_k|c_i) \log \frac{P(d_k|c_i)}{P(d_k|seed_j)}, \tag{4}$$

where $P(d_k|c_i)$ and $P(d_k|seed_j)$ respectively denote the probabilistic weights of the kth dimension of the left (right) context vector of c_i and $seed_j$, and N denotes the dimensionality of a vector. Since the KL divergence presented in Eq. (4) is non-symmetric, i.e., $D(v_{c_i} \| v_{seed_j})$ is not necessarily equal to $D(v_{seed_j} \| v_{c_i})$, the following symmetric divergence measure is adopted by combining the two non-symmetric divergences such that

$$Div(v_{c_i}, v_{seed_j}) = Div(v_{seed_j}, v_{c_i}) = D(v_{c_i} \| v_{seed_j}) + D(v_{seed_j} \| v_{c_i}). \tag{5}$$

In this way, the distance between c_i and $seed_j$ can be calculated by the summation of the KL divergences of their left and right contextual distributions. That is

$$Dist(c_i, seed_j) = Div\left(v_{c_i}^{left}, v_{seed_j}^{left}\right) + Div\left(v_{c_i}^{right}, v_{seed_j}^{right}\right) \tag{6}$$

The two $Div(\cdot,\cdot)$ are added because both seed words and candidate words have their respective left and right context vectors, and this addition allows us to account for both their left and right contextual distributions. Once the distance between a seed word and a candidate word is calculated, the similarity between c_i and $seed_j$ can be defined as

$$CE(c_i, seed_j) = \frac{1}{1 + Dist(c_i, seed_j)} \tag{7}$$

where $CE(c_i, seed_j)$ denotes a similarity score output by the contextual entropy model, representing the contextual similarity between c_i and $seed_j$. In addition, the above equation indicates that a smaller distance between two words indicates a greater similarity between them.

3.2.1.3. Expansion procedure. Once the similarity between each candidate word and each seed word is calculated using the contextual entropy model, the next step is to determine the sentiment class (i.e., positive or negative) for each candidate. The seed words are first divided into positive and negative groups, respectively denoted as P_{seed} and N_{seed}. We then calculate the similarity between a candidate word c_i and each group. That is

$$CE(c_i, P_{seed}) = \frac{1}{|P_{seed}|} \sum_{seed_j \in P_{seed}} CE(c_i, seed_j), \tag{8}$$

$$CE(c_i, N_{seed}) = \frac{1}{|N_{seed}|} \sum_{seed_j \in N_{seed}} CE(c_i, seed_j), \tag{9}$$

where $CE(c_i, P_{seed})$ and $CE(c_i, N_{seed})$ respectively denote the similarity between c_i and the positive and negative group of seed words, calculated by separately averaging the similarity scores between c_i and all seed words in the respective positive and negative groups. By comparing $CE(c_i, P_{seed})$ and $CE(c_i, N_{seed})$, we get

$$r(c_i) = \frac{CE(c_i, P_{seed})}{CE(c_i, N_{seed})}. \tag{10}$$

The sentiment class of a candidate word is determined to be positive if its similarity to the positive group is greater than that to the negative group (i.e., $CE(c_i, P_{seed}) > CE(c_i, N_{seed})$ or $r > 1$), and its intensity is $CE(c_i, P_{seed})$. Otherwise, the sentiment class of the candidate word is negative, and its intensity is $CE(c_i, N_{seed})$. For the sake of simplicity, the intensity is normalized to 0–1 using a sigmoid function. That is

$$Intensity(c_i) = \begin{cases} \frac{1}{1+\exp^{-\gamma CE(c_i,P_{seed})+\theta}}, & \text{if the sentiment class of } c_i \text{ is positive,} \\ \frac{1}{1+\exp^{-\gamma CE(c_i,N_{seed})+\theta}}, & \text{if the sentiment class of } c_i \text{ is negative,} \end{cases} \tag{11}$$

where γ and θ are the parameters used to respectively control the slope and shift of the sigmoid function. Once the sentiment class and intensity of each candidate word are determined, both positive and negative candidates are ranked in the descending order according their intensity. A threshold is then used to select those with higher intensity to be expanded for the subsequent sentiment classification of stock market news because they are more similar to the seed words. The optimal threshold value for expanded word selection is determined empirically, and described in Section 4.2.

3.2.2. Validation part

3.2.2.1. Pointwise mutual information (PMI). Natural language processing commonly uses mutual information to measure the co-occurrence strength between two words [41–43], with a higher mutual information score indicating that the two words co-occur more frequently. Based on this notion, PMI adopts co-occurrence strength as the similarity measure [44] to select words frequently co-occurring with the seed words for expansion. Let $seed_j$ be a seed word, and c_i be a candidate word to be expanded in the unlabeled corpus, their PMI is defined as

$$PMI(c_i, seed_j) = \log_2 \frac{P(c_i, seed_j)}{P(c_i)P(seed_j)}, \tag{12}$$

where $P(c_i, seed_j) = C(c_i, seed_j)/N$ denotes the probability that c_i and $seed_j$ co-occur, which is calculated by the number of times c_i and $seed_j$ co-occur in the same sentence in the unlabeled corpus (i.e., $C(c_i, seed_j)$), divided by the total number of words in the unlabeled corpus (i.e., N). Similarly, $P(c_i) = C(c_i)/N$, where $C(c_i)$ is the number of times c_i occurs in the unlabeled corpus, and $P(seed_j) = C(seed_j)/N$, where $C(seed_j)$ is the number of times $seed_j$ occurs in the unlabeled corpus. Therefore, Eq. (12) can be re-written as

$$PMI(c_i, seed_j) = \log_2 \frac{C(c_i, seed_j) \cdot N}{C(c_i) \cdot C(seed_j)}. \tag{13}$$

Similar to the expansion procedure of the contextual entropy model (Section 3.2.1), once the PMI scores between each candidate word and each seed word are calculated, the next step is to determine the sentiment class of each candidate by calculating its co-occurrence strength relative to the positive (i.e., P_{seed}) and negative (i.e., N_{seed}) seed word groups. By replacing the contextual entropy scores presented in Eqs. (8) and (9) with the PMI scores, we get

$$PMI(c_i, P_{seed}) = \frac{1}{|P_{seed}|} \sum_{seed_j \in P_{seed}} PMI(c_i, seed_j), \tag{14}$$

$$PMI(c_i, N_{seed}) = \frac{1}{|N_{seed}|} \sum_{seed_j \in N_{seed}} PMI(c_i, seed_j), \tag{15}$$

where $PMI(c_i, P_{seed})$ and $PMI(c_i, N_{seed})$ respectively denote the PMI scores between c_i and the positive and negative group of seed words, calculated by separately averaging the PMI scores between c_i and all seed words in the respective positive and negative groups. By comparing $PMI(c_i, P_{seed})$ and $PMI(c_i, N_{seed})$, we get

$$r(c_i) = \frac{PMI(c_i, P_{seed})}{PMI(c_i, N_{seed})}. \tag{16}$$

The sentiment class of a candidate word is determined to be positive if its co-occurrence strength to the positive group is greater than that to the negative group (i.e., $PMI(c_i, P_{seed}) > PMI(c_i, N_{seed})$ or $r > 1$), and its intensity is $PMI(c_i, P_{seed})$;. Otherwise, the sentiment class of the candidate word is negative, and its intensity is $PMI(c_i, N_{seed})$. For the sake of simplicity, the intensity is normalized to 0–1 using a sigmoid function. That is

$$Intensity(c_i) = \begin{cases} \frac{1}{1+\exp^{-\gamma PMI(c_i,P_{seed})+\theta}}, & \text{if the sentiment classof } c_i \text{ is } positive, \\ \frac{1}{1+\exp^{-\gamma PMI(c_i,N_{seed})+\theta}}, & \text{if the sentiment class of } c_i \text{ is } negative, \end{cases} \tag{17}$$

where γ and θ are the parameters used to respectively control the slope and shift of the sigmoid function. Once the sentiment class and intensity of each candidate word are determined, both positive and negative candidates are ranked in the descending order according their intensity. A threshold is then used to select those with higher intensity to be expanded for the subsequent sentiment classification of stock market news because they co-occur more frequently with the seed words. The optimal threshold value for expanded word selection is determined empirically, and described in Section 4.2.

3.2.2.2. Pointwise mutual information with web statistics (PMI-Web). PMI-Web uses the same formula as PMI, but a different corpus for emotion word expansion. It adopts a method called *query-by-Google* [45] to calculate the co-occurrence strength between seed words and candidate words from web statistics instead of the unlabeled corpus used in PMI. In this way, the co-occurrence frequency of any two words can be obtained by taking the frequency returned by Google with the two words as a query.

3.2.2.3. Seed. This method uses the seed words alone for sentiment classification. The procedure of seed word generation is presented in Section 3.1.

3.3. Sentiment classification of stock news

Once the seed words have been expanded, both the seed words and expanded words are unified as an emotion lexicon for sentiment classification. That is, let $Pos = P_{seed} \cup P_{expand}$ be the unification of the sets of positive seed words and their expanded words, and $Neg = N_{seed} \cup N_{expand}$ be the unification of the sets of negative seed words and their expanded words. The resulting emotion lexicon is denoted as $Lex = Pos \cup Neg$. For classifying the sentiment of stock market news articles, two classification schemes are developed (i.e., binary and intensity), depending on whether or not intensity is used in classification. The binary classification scheme only compares the number of positive and negative emotion words contained in news articles without consideration of their intensity levels, whereas the intensity classification scheme compares the sum of the intensity levels of positive and negative emotion words in the articles. The binary classification scheme is defined as

$$l(D) = \begin{cases} positive & \text{if } \sum_{w_i \in D} \sum_{w_j \in Pos} I(w_i, w_j) \\ & \quad -\sum_{w_i \in D} \sum_{w_j \in Neg} I(w_i, w_j) > 0, \\ negative & \text{if } \sum_{w_i \in D} \sum_{w_j \in Pos} I(w_i, w_j) \\ & \quad -\sum_{w_i \in D} \sum_{w_j \in Neg} I(w_i, w_j) \leqslant 0, \end{cases} \tag{18}$$

where $l(D)$ denotes the sentiment label of a news article D, and $I(w_i, w_j)$ is an identity function used to identify whether or not two words are identical, defined as

$$I(w_i, w_j) = \begin{cases} 1 & \text{if } w_i = w_j \\ 0, & \text{if } w_i \neq w_j. \end{cases} \tag{19}$$

By matching between the emotion lexicon and news article D, $\sum_{w_i \in D} \sum_{w_j \in Pos} I(w_i, w_j)$ and $\sum_{w_i \in D} \sum_{w_j \in Neg} I(w_i, w_j)$ in Eq. (18) can be respectively calculated as the number of positive and negative emotion words in *D*, and *l*(*D*) is positive if the number of positive emotion words in *D* is greater than the number of negative ones.

In addition to the emotion words themselves, the intensity classification scheme further considers their intensity levels, defined as

$$l(D) = \begin{cases} positive & \text{if } \sum_{w_i\in D}\sum_{w_j\in Pos} I(w_i,w_j)\ Intensity(w_i) \\ & \quad -\sum_{w_i\in D}\sum_{w_j\in Neg} I(w_i,w_j)\ Intensity(w_i) > 0, \\ negative & \text{if } \sum_{w_i\in D}\sum_{w_j\in Pos} I(w_i,w_j)\ Intensity(w_i) \\ & \quad -\sum_{w_i\in D}\sum_{w_j\in Neg} I(w_i,w_j)\ Intensity(w_i) \leqslant 0, \end{cases} \tag{20}$$

where *Intensity*(w_i) denotes the intensity level of an emotion word in the lexicon. Based on this classification rule, a news article will be associated with a positive label if the sum of the intensity levels of positive emotion words in the article is greater than that of negative emotion words. This is useful when an article has a similar number of positive and negative emotion words.

4. Experimental results

This section presents the evaluation results for sentiment classification of stock market news articles. Section 4.1 describes the experimental setup, including experimental data, classifiers and feature sets, and the evaluation metric. Section 4.2 investigates the selection of optimal parameters for seed word selection and expansion. Section 4.3 presents the comparative results of using different feature sets for different classifiers.

4.1. Experiment setup

4.1.1. Experimental data

A total of 6888 stock market news articles were collected from YAHOO!News. Of these, 3262 articles were randomly selected for manual annotation, and then used for seed word generation. The remaining articles were used for 10-fold cross validation, and split into an expansion set, a development set and a testing set with at a ratio of 8:1:1 for each run. The expansion set was considered to be the unlabeled corpus for emotion word expansion, the development set was used to optimize the thresholds used in the expansion process, and these optimal settings were then used on the testing set to evaluate the classification performance.

4.1.2. Classifiers and feature sets

The classifiers involved in the experiments included *Binary* (Eq. (18)) and *Intensity* (Eq. (20)), and both used the following feature sets for sentiment classification:

- **Seed:** The seed words were generated using the information gain from the labeled news articles, and then manually selected by the annotator.
- **PMI:** The seed words plus the emotion words expanded by PMI from the unlabeled corpus.
- **PMI-Web:** The seed words plus the emotion words expanded by PMI from web statistics instead of the unlabeled corpus.
- **CE:** The seed words plus the emotion words expanded by the contextual entropy model from the unlabeled corpus.

4.1.3. Evaluation metric

The performance was measured by *accuracy*, i.e., the number of correctly classified news articles divided by the total number of test articles.

4.2. Evaluation of threshold selection

The thresholds included the size of the labeled corpus (α) used to generate seed words, the threshold (β) used to select expanded words, and the window size (ℓ) used to calculate the weights in the vector representation of the contextual entropy model. The best setting of these thresholds were tuned by maximizing the classification accuracy on the development set.

4.2.1. Seed word generation

Different sizes of the labeled corpus may affect the quantity and quality of the generated seed words. Fig. 4 shows the Binary and Intensity classification accuracy derived from using seed words alone (Seed) as features against different proportions of the labeled corpus used for seed word generation. The results show that, for both Binary and Intensity classification, increasing the size of the labeled corpus increased the classification performance because the corpus generated more useful seed words. However, this may also increase effort required by the human experts for annotation. The best performance for both Binary and Intensity was achieved at $\alpha = 1$ (using the whole labeled corpus), with respective accuracy rates of 67.90% and 70.07%.

4.2.2. Emotion word expansion

Based on the optimal setting $\alpha = 1$ for seed word generation, this experiment evaluated the parameters ℓ and β used for emotion word expansion. The window size ℓ was used in the contextual entropy model to determine the number of context words that can be observed in building vector representation. Too small a window may result in the loss of some important contextual information, while an overly large window may introduce noisy information. The threshold β was applied to the respective ranked lists of candidate emotion words produced by the contextual entropy model and PMI-based methods to control the number of emotion words expanded by each method (see Section 3.2). A small value of β (percentage) indicates a small number of words will be selected for expansion, but some useful emotion words ranked below β may be missed. Although a large value of β may include more useful emotion words, the ranked list may include noisy words at the bottom.

Figs. 5 and 6 show the respective Binary and Intensity classification accuracy of CE with different window sizes ℓ against different

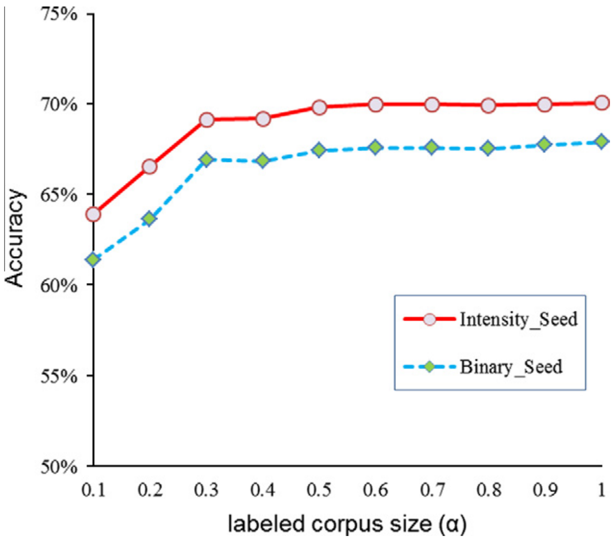

Fig. 4. Binary and intensity classification accuracy of Seed against different proportions of the labeled corpus (α) used for seed word generation.

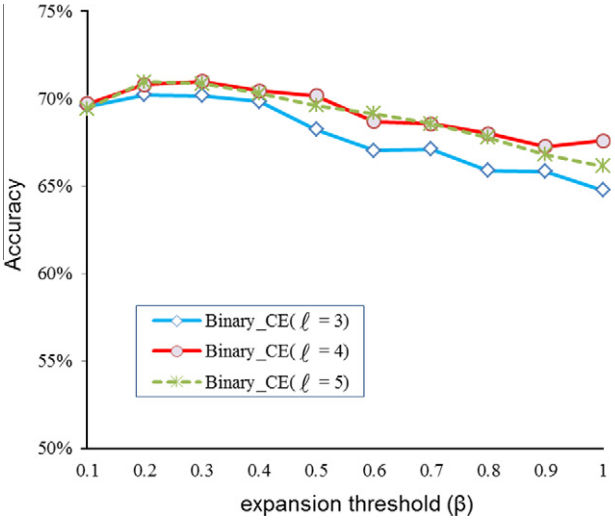

Fig. 5. Binary classification accuracy of CE with different window sizes ℓ against different threshold values of β.

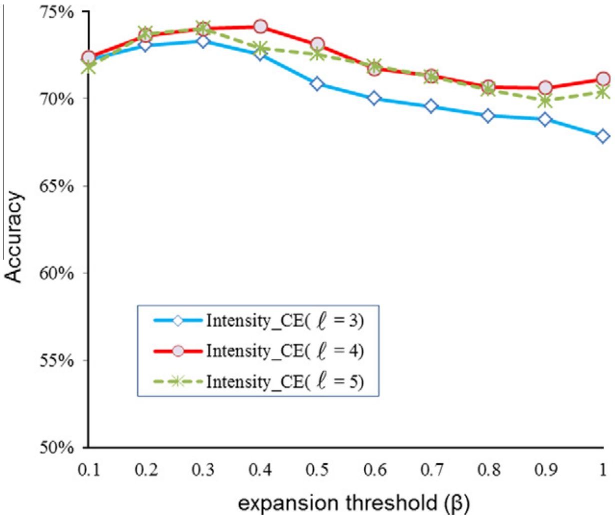

Fig. 6. Intensity classification accuracy of CE with different window sizes ℓ against different values of β.

threshold values of β for emotion word expansion. The results show that the optimal setting of the window size is $\ell = 4$ for both Binary and Intensity classification. Another observation is that increasing the value of β does not necessarily increase classification performance. For example, in the case of $\ell = 4$ in Fig. 5, the accuracy increased with β up to 0.3 (i.e., the top 30% emotion words of the ranked list), indicating that more useful emotion words at the top of the ranked list were expanded by the contextual entropy model, thus improving classification performance. Once β exceeded the best setting, the performance decreased gradually because of the expansion of more noisy words at the bottom of the ranked list, thus increasing classification ambiguity. Fig. 6 also shows a similar tendency.

To investigate the effect of β on the PMI-based methods, Figs. 7 and 8 show the respective Binary and Intensity classification results for both PMI and PMI-Web against different threshold values of β. The results show that PMI and PMI-Web produced similar levels of accuracy for both Binary and Intensity classification even though PMI-Web adopted a larger corpus (i.e., query by Google) for expansion. To further compare the proposed contextual entropy model and the PMI-based methods, the results of CE with the optimal window size setting ($\ell = 4$) are also presented in Figs. 7 and 8. The Binary classification results presented in Fig. 7 show that the accuracy levels of the two PMI-based methods increased with β up to 0.5, but were still lower than that of CE at the best setting $\beta = 0.3$. Additionally, the accuracy levels of the PMI-based methods dropped significantly when β exceeded the optimal setting, but the accuracy of CE did not drop significantly. The Intensity classification results presented in Fig. 8 also show similar phenomena. These findings indicate that CE can not only acquire more useful emotion words but also includes fewer noisy words than the PMI-based methods. An interesting distinction between the results of Binary (Fig. 7) and Intensity (Fig. 8) classification is that the performance difference between CE and the PMI-based methods for Intensity classification was greater than that for Binary classification, especially for β between 0.1 and 0.5. One possible reason is

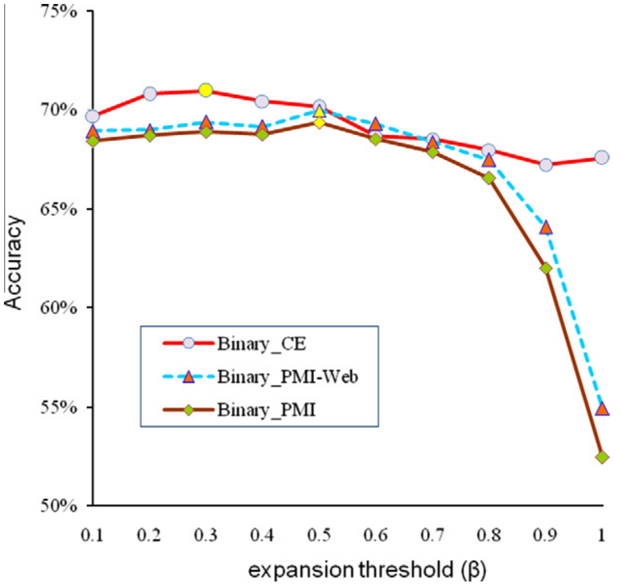

Fig. 7. Binary classification accuracy of PMI, PMI-Web and CE against different threshold values of β.

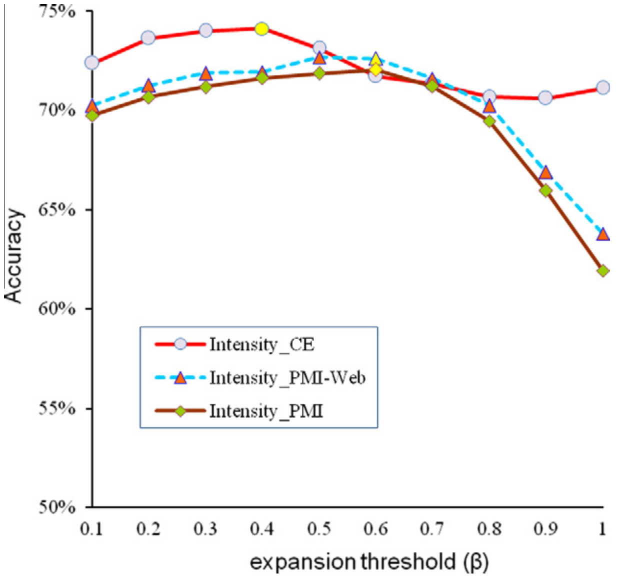

Fig. 8. Intensity classification accuracy of PMI, PMI-Web and CE against different threshold values of β.

Table 1
Comparative results of different methods for sentiment classification.

	Seed (%)	PMI (%)	PMI-Web (%)	CE (%)
Binary	67.90	68.40	68.95	73.19*
Intensity	70.07	71.43	72.15	76.67*

* CE vs PMI-Web significantly different ($p < 0.05$).

Table 2
Intensity classification accuracy against the difference of intensity levels between the positive and negative classes.

Difference	PMI-Web		CE	
	% Of test examples	Accuracy (%)	% Of test examples	Accuracy (%)
−0.5 to 0.5	8.77	51.26	8.55	52.58
±0.5 to 1.5	13.71	64.19	16.22	66.16
±1.5 to 2.5	13.82	68.86	12.74	68.83
±2.5 to 3.5	10.45	66.75	10.70	78.09
±3.5 to 5.5	19.03	74.64	16.88	80.07
±5.5 to 7.5	11.86	80.23	12.24	84.91
±7.5 to 9.5	10.45	82.32	10.01	88.15
>9.5	11.91	84.26	12.66	91.50
Sum/Avg.	100	72.15	100	76.67

that the intensity calculated by the contextual entropy model based on context distributions is more suitable than that calculated by the PMI-based methods based on co-occurrence frequency for determining the sentiment of stock market news articles.

4.3. Comparative results

Table 1 shows the evaluation results of using as features for sentiment classification the seed words alone (Seed), the seed words plus the emotion words expanded by the PMI-based methods (PMI and PMI-Web) and the contextual entropy model (CE). The paired, two-tailed *t*-test was used to determine whether the performance difference was statistically significant. The results show that both the PMI-based methods and CE outperform Seed, indicating that the use of expanded emotion words can improve classification performance. In addition, performance was further improved by taking the intensity of emotions words into account. As indicated, the accuracy levels of all methods for Intensity classification were higher than those for Binary classification. In comparing the expansion methods, CE outperformed both PMI-based methods, largely because the proposed method considers both co-occurrence strength and contextual distribution between the seed and candidate words, and can thus acquire more useful emotion words and exclude more noisy words than the PMI-based methods, which only consider the co-occurrence strength.

In classifying the sentiment of test examples, as shown in Eq. (20), some examples may have a similar sum of positive and negative class intensity levels, thus yielding a limited range of difference of intensity levels between the positive and negative classes. To explore the relationship between the difference of intensity levels and classification performance, we analyze the Intensity classification results of each test example for both PMI-Web and CE, as summarized in Table 2. The column labeled “Difference” represents groups of test examples divided according to their difference of intensity levels between the positive and negative classes. Columns labeled “% of test examples” and “Accuracy” respectively represent the corresponding percentages of the test examples and the accuracy levels for each method. The results show a small portion of test examples (8.77% for PMI-Web and 8.55% for CE) had a small difference (−0.5 to 0.5) of intensity levels between the positive and negative classes, and such examples tend to confuse the classifiers, thus yielding a lower degree of accuracy (51.26% for PMI-Web and 52.58% for CE). Once the difference of intensity levels increased, as indicated by ±0.5 to 1.5, the accuracy of both methods jumped above 64%, and continued to increase with the degree of difference because classifiers operated with increased confidence on examples with a larger degree of difference in intensity levels.

5. Conclusions

A contextual entropy model was proposed to expand a set of seed words by discovering similar emotion words and their corresponding intensities from online stock market news articles. This was accomplished by calculating the similarity between the seed words and candidate words from their contextual distributions using an entropy measure. Once the seed words have been expanded, both the seed words and expanded words are used to classify the sentiment of the news articles. Experimental results show that the use of the expanded emotion words improved classification performance, which was further improved by incorporating their corresponding intensities. Our proposed method considers both co-occurrence strength and contextual distribution, thus acquiring more useful emotion words and fewer noisy words and outperforming PMI which only considers the co-occurrence strength.

Future work will pursue three directions. First, both seed words and candidate words in this study exclusively represent individual words, and compound words or multi-words are not considered. Semantic composition of individual words can be investigated to extend the proposed method to automatically acquire compound/multi-words and their intensity. Second, co-reference techniques can be investigated to analyze whether an emotion word refers to a company or the whole market, thus enhancing classification performance when a particular company or market is specified. Additionally, more significant features such as compound words and co-reference information can be integrated to further improve classification performance. Third, both the sentiment and intensity of stock market news articles will be incorporated into a stock trend prediction model to further improve its performance.

References

[1] P.C. Chang, C.Y. Fan, A hybrid system integrating a wavelet and TSK fuzzy rules for stock price forecasting, IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews 38 (6) (2008) 802–815.

[2] W. Shen, X. Guo, C. Wu, D. Wu, Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm, Knowledge-Based Systems 24 (3) (2011) 378–385.

[3] P.C. Chang, T.W. Liao, J.J. Lin, C.Y. Fan, A dynamic threshold decision system for stock trading signal detection, Applied Soft Computing 11 (5) (2011) 3998–4010.

[4] S. Asadi, E. Hadavandi, F. Mehmanpazir, M.M. Nakhostin, Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction, Knowledge-Based Systems 35 (2012) 245–258.

[5] G.P.C. Fung, J.X. Yu, W. Lam, News sensitive stock trend prediction, in: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD-02), 2002, pp. 481–493.

[6] L. Zhou, Y. Zhang, C. Xing, Y. Sun, X. Zhu, Sentiment classification for stock news, in: Proceedings of the IEEE 5th International Conference on Pervasive Computing and Applications (ICPCA-10), 2010, pp. 99–104.

[7] S.W.K. Chan, J. Franklin, A text-based decision support system for financial sequence prediction, Decision Support Systems 52 (1) (2011) 189–198.

[8] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems 52 (4) (2011) 680–691.

[9] X. Liang, Neural network method to predict stock price movement based on stock information entropy, Lecture Notes in Computer Science 3973 (2006) 442–451.

[10] A. Devitt, K. Ahmad, Sentiment polarity identification in financial news: a cohesion-based approach, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07), 2007, pp. 984–991.

[11] M.M. Bradley, P.J. Lang, Affective Norms for English Words (ANEWs): Instruction manual and affective ratings, Technical Report C-1, Center for Research in Psychophysiology, University of Florida, 1999.

[12] A. Valitutti, C. Strapparava, O. Stock, Developing affective lexical resources, PsychNology 2 (1) (2004) 61–83.

[13] A. Esuli, F. Sebastiani, SentiWordNet: A publicly available lexical resource for opinion mining, in: Proceedings of the 5th Conference on, Language Resources and Evaluation (LREC-06), 2006, pp. 417–422.

[14] A. Neviarouskaya, H. Prendinger, M. Ishizuka, SentiFul: a lexicon for sentiment analysis, IEEE Transactions on Affective Computing 2 (1) (2011) 22–36.

[15] A. Balahur, J.M. Hermida, A. Montoyo, Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model, IEEE Transactions on Affective Computing 3 (1) (2012) 88–101.

[16] L.W. Ku, H.H. Chen, Mining opinions from the Web: beyond relevance retrieval, Journal of the American Society for Information Science and Technology 58 (12) (2007) 1838–1850.

[17] Y. Hu, X. Chen, D. Yang, Lyric-based song emotion detection with affective lexicon and fuzzy clustering method, in: Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR-09), 2009, pp. 123–128.

[18] C. Quan, F. Ren, Sentence emotion analysis and recognition based on emotion words using Ren-CECps, International Journal of Advanced Intelligence 2 (1) (2010) 105–117.

[19] G. Xu, X. Meng, H. Wang, Build Chinese emotion lexicons using a graph-based algorithm and multiple resources, in: Proceedings of the 23rd International Conference on Computational Linguistics (Coling-10), 2010, pp. 1209–1217.

[20] Y.W. Liu, S.B. Xiao, T. Wang, S.C. Shi, Building Chinese sentiment lexicon based on HowNet, Advanced Materials Research 187 (2011) 405–410.

[21] N. Li, D.D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, Decision Support Systems 48 (2) (2010) 354–368.

[22] C. Kaiser, S. Schlick, F. Bodendorf, Warning system for online market research – identifying critical situations in online opinion formation, Knowledge-Based Systems 24 (6) (2011) 824–836.

[23] C.H. Wu, L.C. Yu, F.L. Jang, Using semantic dependencies to mine depressive symptoms from consultation records, IEEE Intelligent Systems 20 (6) (2005) 50–58.

[24] L.C. Yu, C.L. Chan, C.C. Lin, I.C. Lin, Mining association language patterns using a distributional semantic model for negative life event classification, Journal of Biomedical Informatics 44 (4) (2011) 509–518.

[25] P.D. Turney, M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems 21 (4) (2003) 315–346.

[26] M. Baroni, S. Vegnaduzzo, Identifying subjective adjectives through Web-based mutual information, in: Proceedings of the Conference for the Processing of Natural, Language and Speech, 2004, pp. 17–24.

[27] G. Grefenstette, Y. Qu, D.A. Evans, J.G. Shanahan, Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes, Computing Attitude and Affect in Text: Theory and Applications 20 (2006) 93–107.

[28] A. Abbasi, H. Chen, S. Thoms, T. Fu, Affect analysis of web forums and blogs using correlation ensembles, IEEE Transactions on Knowledge and Data Engineering 20 (9) (2008) 1168–1180.

[29] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Mass, 1999.

[30] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the 14th International Conference on, Machine Learning (ICML-97), 1997, pp. 412–420.

[31] H. Uğuz, A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm, Knowledge-Based Systems 24 (7) (2011) 1024–1032.

[32] C. Burgess, K. Livesay, K. Lund, Explorations in context space. Words, sentences, discourse, Discourse Processes 25 (2–3) (1998) 211–257.

[33] L.C. Yu, C.H. Wu, J.F. Yeh, F.L. Jang, HAL-based evolutionary inference for pattern induction from psychiatry Web resources, IEEE Transactions on Evolutionary Computation 12 (2) (2008) 160–170.

[34] T. Xu, Q. Peng, Y. Cheng, Identifying the semantic orientation of terms using S-HAL for sentiment analysis, Knowledge-Based Systems 35 (2012) 279–289.

[35] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (5) (1988) 513–523.

[36] P.I. Chen, S.J. Lin, Word AdHoc network: using google core distance to extract the most relevant information, Knowledge-Based Systems 24 (3) (2011) 393–405.

[37] Y. Pan, H.X. Luo, Y. Tang, C.Q. Huang, Learning to rank with document ranks and scores, Knowledge-Based Systems 24 (4) (2011) 478–483.

[38] Q. Cao, W. Duan, Q. Gan, Exploring determinants of voting for the "helpfulness" of online user reviews: a text mining approach, Decision Support Systems 50 (2) (2011) 511–521.

[39] S. Kullback, Information Theory and Statistics, John-Wiley & Sons, New York, 1959.

[40] M.H.A. Hijazi, F. Coenen, Y. Zheng, Data mining techniques for the screening of age-related macular degeneration, Knowledge-Based Systems 29 (2012) 83–92.

[41] K.W. Church, P. Hanks, Word association norms, mutual information, and lexicography, Computational Linguistics 16 (1) (1990) 22–29.

[42] S. Deng, Z. He, X. Xu, G-ANMI: a mutual information based genetic clustering algorithm for categorical data, Knowledge-Based Systems 23 (2) (2010) 144–149.

[43] S. Cang, H. Yu, Mutual information based input feature selection for classification problems, Decision Support Systems 54 (1) (2012) 691–698.

[44] I. Novalija, D. Mladenić, L. Bradeško, OntoPlus: text-driven ontology extension using ontology content, structure and co-occurrence information, Knowledge-Based Systems 24 (8) (2011) 1261–1276.

[45] L.C. Yu, C.H. Wu, R.Y. Chang, C.H. Liu, E.H. Hovy, Annotation and verification of sense pools in OntoNotes, Information Processing & Management 46 (4) (2010) 436–447.