A bilingual approach for conducting Chinese and English social media sentiment analysis

Article *in* Computer Networks · December 2014

DOI: 10.1016/j.comnet.2014.08.021

Computer Networks 75 (2014) 491–503

Contents lists available at ScienceDirect

Computer Networks

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

A bilingual approach for conducting Chinese and English social media sentiment analysis

Gongjun Yan [a],*, Wu He [b], Jiancheng Shen [c], Chuanyi Tang [d]

[a] Management and Information Sciences, University of Southern Indiana, Evansville, IN 47712, United States
[b] Information Technology & Decision Sciences Department, Old Dominion University, Norfolk, VA 23529, United States
[c] Finance Department, Old Dominion University, Norfolk, VA 23529, United States
[d] Marketing Department, Old Dominion University, Norfolk, VA 23529, United States

ARTICLE INFO

Article history:
Received 15 November 2013
Received in revised form 20 July 2014
Accepted 11 August 2014
Available online 5 November 2014

Keywords:
Social media
Social network
Sentiment analysis
Bilingual approach
Text mining
Objective opinion

ABSTRACT

Due to the advancement of technology and globalization, it has become much easier for people around the world to express their opinions through social media platforms. Harvesting opinions through sentiment analysis from people with different backgrounds and from different cultures via social media platforms can help modern organizations, including corporations and governments understand customers, make decisions, and develop strategies. However, multiple languages posted on many social media platforms make it difficult to perform a sentiment analysis with acceptable levels of accuracy and consistency. In this paper, we propose a bilingual approach to conducting sentiment analysis on both Chinese and English social media to obtain more objective and consistent opinions. Instead of processing English and Chinese comments separately, our approach treats review comments as a stream of text containing both Chinese and English words. That stream of text is then segmented by our segment model and trimmed by the stop word lists which include both Chinese and English words. The stem words are then processed into feature vectors and then applied with two exchangeable natural language models, SVM and N-Gram. Finally, we perform a case study, applying our proposed approach to analyzing movie reviews obtained from social media. Our experiment shows that our proposed approach has a high level of accuracy and is more effective than the existing learning-based approaches.

1. Introduction

In today's globalized environment, consumers from different countries become increasingly connected. People from different cultures can easily express their comments or opinions on the same events or products via social media tools. As a result, multi-language comments on the same topic have become more common than ever. For example, a popular movie can be reviewed and discussed by fans all over the world and these fans may use many different languages. The multi-language comments bring both opportunities and challenges to researchers and practitioners. On the one hand, multilingual social media provides international political and business practitioners insightful information about individuals in other countries. For example, researchers can predict the result of a political election campaign by tracking the sentiments toward each candidate on the foreign country's social media websites. Multinational enterprise can apply bilingual sentiment approaches to examine the foreign country's social media to collect the information regarding the foreign consumers' attitudes and preferences toward a certain product

* Corresponding author.
E-mail addresses: gyan@usi.edu (G. Yan), whe@odu.edu (W. He), jshen@odu.edu (J. Shen), ctang@odu.edu (C. Tang).

http://dx.doi.org/10.1016/j.comnet.2014.08.021
1389-1286/© 2014 Elsevier B.V. All rights reserved.

or brand. On the other hand, how to effectively utilize the valuable multi-language data contained in social media becomes challenging due to the lack of effective tools to analyze the multi-language information reliably and accurately.

In recent years the rapid development of social media has significantly influenced the way in which people communicate with one another and obtain information. Nowadays social media is ubiquitous and plays an important role in people's daily lives. For example, in the business field, more and more consumers rely on user comments posted on Facebook, Twitter or Amazon to valuate products and services prior to making a purchase [12,34,23,13]; in the finance field, investors' opinions and moods posted in social media have been utilized to predict future stock returns and earnings [5,2]; in health care services, many patients use social media to discuss medical services and their doctors in order to optimize treatments [11]; in education, many students and parents use social media platforms to exchange opinions and discuss the pros and cons of their interested colleges [7].

To better leverage social media, researchers have proposed and developed a number of methods to analyze social media data. Sentiment analysis is one of the most popular methods and has been widely used to analyze user-generated text content of social media [27,21]. However, the existing research on sentiment analysis largely focuses on a single language (i.e., English).

Little attention has been paid to bilingual sentiment analysis. Bilingual sentiment analysis is challenging because it deals with multi-culture comments and analyzing English and Chinese comments separately is inefficient and less persuasive. For example, there are a lot of Chinese and English comments on Huawei's fight to gain access to the U.S. market. Huawei is the largest Chinese telecommunication equipment maker. Most Chinese Internet users' comments are positive on this event since they support Huawei. In contrast, most English comments are negative on the same event. In this situation, it is hard for researchers to make a fair judgment if they run sentiment analysis separately as the results on Chinese and English comments are contradictory with each other. Our method combines both English and Chinese comments and treat them as unified comments data. It can indistinctively process these comments no matter which language is used in them and obtain an overall sentiment opinion based on the multi-language comments. Thus, the sentiment opinion is relatively more objective and consistent than the results of analyzing the comments in different languages separately.

More specifically, this study proposes a novel bilingual approach for social media sentiment analysis. Our bilingual approach contributes to the sentiment analysis literature in several important ways. First, our single model can process two different languages (in the case of our study, English and Chinese) simultaneously in a single application. In contrast, in other existing approaches the two languages have to be processed in two separate applications, which may end up with contradictory results. Second, although we focused on English and Chinese, the design of our bilingual module allows for the expansion to other languages. Finally, our bilingual model can be modified to incorporate popular text mining models (in the case of our study, N-Gram and SVM) into bilingual sentiment analysis. In addition to introducing a new method, we also validate our proposed method by conducting a case study in which we adopted the proposed bilingual approach to process movie-related review comments.

The remainder of the paper is organized as follows. Section 2 provides a review of sentiment analysis as well as the recent progress in sentiment analysis of multiple languages. Section 3 proposes a novel bilingual approach for social media sentiment analysis. Section 4 presents a case study with movie-related review comments we collected from both English and Chinese social media sites. Conclusions and future research are provided in Section 5.

2. Literature review

2.1. An overview of sentiment analysis

There is a growing interest in using sentiment analysis methods to mine user-generated data. Sentiment analysis has been used to determine the attitude of customers and online users on some specific topics, such as consumer products (e.g., books, movies) reviews, hotel service reviews, public relations statements, and financial blogs [27,21,22]. Sentiment analysis mainly relies on machine learning techniques, such as Support Vector Machine (SVM), Naive Bayes Classifier, and Maximum Entropy, to classify texts into positive or negative categories [21,28]. Naive Bayes Classifier is a probabilistic and supervised classifier whose algorithm is mainly implemented to calculate the probability of a data to be positive or negative [31].

Bollen et al. [2] used sentiment analysis to mine a large corpus of Twitter messages to determine the mood of the Twitter population on a given day. They found that the mood of the Twitter population was able to predict the movement of the Dow Jones Industrial Average (DJIA) on the following day with a claimed 87.6% accuracy. There exist Multilanguage comments in twitter. However, only English comments are analyzed. A method which can process multiple languages is needed to obtain comprehensive opinions. Stieglitz and Dang-Xuan [33] also used a sentiment analysis tool called SentiStrength [38] to analyze two data sets of more than 165,000 tweets and found that emotionally charged Twitter messages tend to be retweeted more often and more quickly compared to neutral ones. Klein et al. [18] have proposed a novel approach for extracting investor sentiment from a set of blog articles to predict future returns for investment managers and other stakeholders in the financial industry. Lee et al. [20] used sentiment analysis to extract the sentiment contained in each idea and comment collected from a website for building a recommendation system, which can help firms identify prospective ideas for their innovation among a large number of ideas.

In summary, sentiment analysis has often been used to show the positive and negative trends in the data sets and has been recognized as a feasible method to provide categorical insights into unstructured textual data [10]. However, the above studies mainly focused on the textual data written in English and it is not clear whether their findings can be generalized to other languages or the results are unbiased by language.

2.2. Sentiment analysis in multiple languages

Most of the previous sentiment studies focus on sentiment classification of texts written in a single language. However, effectively applying sentiment analysis in international research requires sentiment analysis techniques that can handle a variety of languages such as English, Arabic, Chinese, and French. To meet this need, a growing number of studies have proposed sentiment analysis methods that translate one language to another language [39], such as French and English text has been automatically collected and constructed as a training corpus [25] and Chinese and English translation has been studied [37].

Basic methods for cross-lingual sentiment classification include lexicon-based methods and corpus-based methods, which enable machine translation services to eliminate the gap between languages [39]. For the lexicon based methods, Kim et al. [16,17] proposed a method to automatically create a sentiment lexicon in a foreign language using a sentiment lexicon in a resource-rich language with only a bilingual dictionary. For the corpus-based methods, Bautin et al. [1] provided cross-language analysis across parallel corpora and performed sentiment analysis on the English translation of texts written in a foreign language. Schulz et al. [30] used an existing English corpus as a basis for the multilingual corpus, and developed a manually annotated multilingual corpus with fine-grained opinions and target annotations. To further eliminate the translation errors in machine translation services, a few existing studies used co-training approaches to improve multilingual sentiment classification accuracy [39,8]. These methods are mainly translation among different languages. We, instead, directly analyze sentiment in both English and Chinese without translation.

Sentiment analysis in Chinese has become a popular topic due to a large amount of social media data generated on Chinese websites. A few researchers, including Yao et al. [41], Tan and Zhang [36] and Zhang et al. [43], have proposed some sentiment analysis approaches in analyzing Chinese. However, generally the existing research of bilingual sentiment analysis in English and Chinese is still limited in literature and the classification algorithms, such as feature selection and feature dimension reduction methods, are still not mature and sufficient [10,40]. Chinese word segmentation is an open topic and is different from English word segmentation. In addition, the Chinese language also has a unique way of emotional expression. Thus, sentiment analysis methods that developed for English language may not be appropriate for handling Chinese language directly [43]. Therefore, we want to develop a method that can handle both English and Chinese to (1) compare the sentiment mining results and (2) obtain comprehensive understanding of opinions that cover a multicultural scenario.

Compared with the previous studies, our proposed approach (including the commonly used machine learning methods such as SVM and N-Gram and existing Chinese word segmentation techniques), which is discussed in further detail below, is able to address the important question on how to compatibly handle both English and Chinese sentiment analysis in a parallel way. Therefore, when the comments collected from social media include both English and Chinese, users do not need to manually pre-process text. The text comments related to the same topic regardless of the language used will be processed in a single model to achieve high utility and accuracy. In addition, our open design allows exchangeable text mining models. Different models can be tested and compared in our system. The most adaptive model can be tested and identified to further increase accuracy.

3. Proposing a bilingual approach for sentiment analysis

In this section, we propose a novel bilingual approach to conduct sentiment analysis. We choose to focus on English and Chinese because of their popularity—Chinese and English rank as two most commonly spoken languages. Because we chose English and Chinese, we needed to deal with the challenge of segmenting Chinese words. After segmenting the words, we then needed to trim down words and extract a few feature words that represent the main meaning of a paragraph or a document. Once the feature was selected, we needed to modify the existing models to achieve our bilingual analysis purpose.

3.1. Language segment

Unlike western languages (e.g., English), the Chinese language does not have delimiters (e.g., white-space). Words in Chinese are streams of characters, making it difficult to identify individual words. In fact there has been significant research on Chinese language segmentation [32,29,35,15,26,14]. We reviewed the extant research on Chinese word segmentation and introduce the fundamental ideas of Chinese word segmentation here.

The Chinese word segmentation in this paper is based on both a standard and customized dictionary. The basic processing steps of word segmentation are shown in Fig. 1. During the design process, two experienced data mining researchers who are acquainted with both English and Chinese conducted in-depth literature review, discussions and hands-on testing to fine tune the proposed approach (see Fig. 1). We first saved all the downloaded social media posts into text documents and saved the documents as samples. These samples were our input data files and were composed of streams of Chinese characters, English and numbers. Input data files included three types of characters: Chinese characters, English letters, and numbers. After these data samples enter the segmentation function through the segmentation interface, Chinese dictionaries—including both an up-to-date dictionary and customized dictionary—were loaded. The segmentation function searched for matches of words based on these dictionaries. The output of the search algorithm was a list of possible segmented Chinese words. Based on the similarity of predefined segmentation results, we computed the best

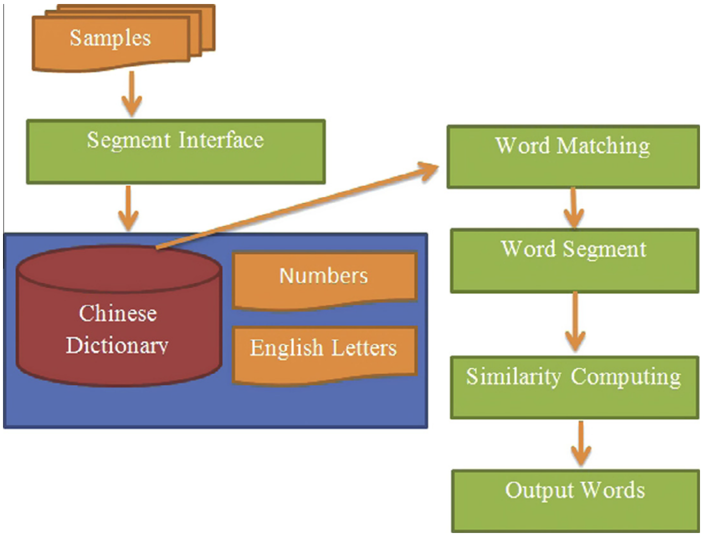


Fig. 1. A bilingual approach for Chinese and English word segmentation. English and Chinese can be simultaneously processed. Every box represents a system module which can be replaced by a different implementation and be extended to a more sophisticated one.

word segmentation. Finally, the word segmentation analyzer output a list of words which were the best segmented ones according to the selection rules.

Chinese word segmentation is an area open for exploration. The most significant problem that the Chinese word segmentation brings into sentiment analysis is its low accuracy. For example, "长春市长春花店" can be segmented into several ways, including "长春/市长/春花/店", or "长春市/长春/花店" or "长春市/长/春花/店". There are so many different ways to segment the sentence and the words segmented in one way greatly differ from the ones segmented in the other ways. The different ways a stream of Chinese characters may be segmented can greatly affect the accuracy of sentiment analysis. Therefore, it is important to choose the right Chinese segmentation methods/tools. There are actually quite a few well-performing Chinese segmentation algorithms and tools, which can be generalized into two methods: (1) dictionary based segmentation and (2) statistics and machine learning based segmentation.

For dictionary based segmentation, the input text is treated as a stream of characters. Common words in a lexicon dictionary are used to match the stream of characters. If there is a match, we can extract the words from the stream of characters. Normally there are some rules added (e.g., forward/backward searching for longest match and long-word first,) to improve its performance. The advantages of this type of algorithm are the speed (computing complexity $O(n)$) and the ease of implementation coupled with a decent overall performance (95% and above accuracy). However, this matching lexicon algorithm may lose accuracy when a word has different meanings, when a word is not in the dictionary, or when irony and sarcasm are applied. Examples of popular tools in this category are IKAnalyzer [15] and Paoding [26].

For the statistics and machine learning method, a certain number of words are manually marked as training data for a certain category. We use the training data for the specific category to optimize the best parameters of statistical models or text mining modules. Based on the optimized parameters, we can compute the probability of words falling into a category. Normally, the performance of this method, in terms of predicting words which are new or with multiple meanings is better than the dictionary based method. However, people have to manually mark a large number of words to achieve good performance. In addition, the computing time is normally longer than the method of matching lexicon words. An example of a popular tool in this category is ICTCLAS [14].

Our main focus in this paper is on bilingual sentiment analysis instead of Chinese word segmentation. Thus, we simply adopted the most well-known open source tool IKAnalyzer [15] to perform the Chinese word segmentation for us. IKAnalyzer is the most widely used open source Java tool for Chinese word segmentation in the market and has been widely used by researchers in processing the Chinese language [42]. The benefits of adopting this tool include: (1) saving time by avoiding having to manually mark a large number of words; (2) it supports multiple granularity matching; and (3) it is lean and fast.

3.2. Modified Chi-square feature selection

A large number of messages can generate thousands, millions, even billions of words. A large amount of words cause computational difficulty in analyzing and training a model and the "curse of dimensionality" [42], if all the words are counted in computing. An intuitive idea is to extract the features of each message and process similarity based on the extracted features instead of the whole messages. There are two steps to extract representative features: (1) elect feature sets and (2) extract feature values.

One of the challenges of electing feature sets is determining representative words without sacrificing the performance of the algorithm. Therefore, to improve the accuracy and reduce the computing time, we need to delete irrelevant or redundant features (words) in the messages to reduce the total number of features. In the literature, there are several methods to elect representative word sets, including Chi square [44], local/global document frequency, and information gain.

We use the Chi-square method and define Chi-square method of feature selection as the follows:

$$\chi^2(t,c_i) = \frac{N[P(t,c_i)P(\bar{t},\bar{c}_i) - P(\bar{t},c_i)P(t,\bar{c}_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)}$$

where N is the total number of documents. In this paper, we saved each social media comment in a single document. Chi-square assumes that t and c_i are independent of each other. According to the value of $\chi^2(t,c_i)$, t is independent to c_i if the value is very small, i.e. $\chi^2(t,c_i)$ is treated as measurement errors. Otherwise, t is not independent of c_i, i.e. t is the member of c_i.

In reality, the number of categories and the distribution of terms in each category can significantly impact the effectiveness of the methods. Define a function $\Im(\cdot,c_i)$ as the term t with the larger value of $\Im(t,c_i)$ is the term that more likely belongs to the category c_i Effective feature selection methods will:

1. select L_{F+} greatest term t-s where $\Im(t,c_i)$ is sorted in decreasing order,
2. select L_{F-} smallest term t-s where $\Im(t,c_i)$ is sorted in increasing order,
3. optimize L_{F+} and L_{F-} for better performance.

The Chi-square method may create significant problems. Only the presence of terms is considered while the frequency of the terms in a document is not considered. The frequency of terms often has a greater importance of the term. Therefore, we modified the Chi-square method by incorporating the frequency of terms. We present:

$$\Im(t,c_i) = a\frac{\chi^2(t,c_i)}{\Sigma_{j=1}^{N_c}\chi^2(t,c_j)} + b\frac{f(t,d_i)}{\Sigma_{j=1}^{N_d}f(t,d_j)}$$

where N_c is the total number of categories, $f(t, d_i)$ is a function that shows the frequency of term t in document d_i, N_d is the total number of document, and $a \in R_{\geqslant 0}$, $b \in R_{\geqslant 0}$ are coefficients with condition $a + b = 1$ $(R_{\geqslant 0} = \{\chi \in R : \chi \geqslant 0\})$.

3.3. Vector space model

The Vector Space Model (VSM) is a well-known model to convert text terms into a vector of identifiers, i.e. indexed terms.

Documents and queries are respectively represented by: $d_j = (t_{1,j}, t_{2,j}, t_{t,j})$ and $q = (t_{1,q}, t_{2,q}, t_{t,q})$. Each item corresponds to a separate term $t_{i,j}$. If a term $t_{i,j}$ occurs in the document d_j, $t_{i,j}$ is non-zero. All the terms will first filter through a pre-defined dictionary after stemlized, i.e. filtered out stem words. If a term is not in the pre-defined dictionary, the weight for the term will be 0, i.e. $t_{i,j} = 0$. There are several ways to give specified value to $t_{i,j}$. The term frequency–inverse document frequency (tf–idf) is one of the well-adopted methods [24]. For example, a message “please contact the seller by phone, phone number: 812-888-8888” can be processed as follows:

The algorithm that computes feature weights is very important. Only stem words will be kept and their weights will be calculated. The feature weight in Table 1 is obtained by counting the occurrence of the words. Only “phone” showed up twice. We need to make the feature weight a positive number less than one. Therefore, we can compute the norm of vector [0,1,1,2,1]. For example, 2.64 can be normalized as 0/2.64 or 0, 1/(2.64) or 0.38, 1/(2.64) or 0.38, 2/(2.64) or 0.75, 1/(2.64) or 0.38. Although the tf–idf has been well adopted in the literature, some research showed that the tf–idf may not be the best algorithm [19]. Sometimes, simply tf can perform better than tf–idf [9].

3.4. Classifier modules

As the primary purpose of our research is to harvest social opinion from multi-culture social media, our main effort is to develop a bilingual method which can simultaneously mine both Chinese and English and extract social opinions from social media. Classifier modules in text mining are not our main focus. However, we adopted two classifier modules to show the effectiveness of the proposed methods: N-Gram [3,4] and SVM [6]. There are two reasons: (1) we have developed software in our previous work and (2) the two modules are well-known and preform well. To better mine social opinions, we define our own cost functions in both models. Note that the classifiers in the proposed approach are exchangeable. In practice, other classifier models can be inserted or replaced in our framework.

3.4.1. Modified N-Gram

The basic idea of the N-Gram [3,4] method is to compute the probability that term t will show up after specific sequence of n terms $t_1, t_2, \ldots, t_n$, i.e. $P(t_i|t_1, \ldots, t_{i-1})$.

Since there are limited terms in documents and extremely expensive in computing time and capacity, N-Gram often assumes to have Markov property, i.e. $P(t_i|t_1,\ldots,t_{i-1}) \approx P(t_i|t_{i-(n-1)},\ldots,t_{i-1})$. That means the probability of observing the i-th term t_i in the context history of the continuous preceding $i-1$ terms can be roughly approximated by the probability of observing t_i in the shortened context history of the continuous preceding $n-1$ words (n-th order Markov property).

Since bigram will generate a lot 0 values (a lot of term combinations are not reasonable and logical), it is hard to compute a new term combination which is reasonable and logical. Therefore, we define a new probability to represent $h_\Theta(t_i) = aP_{\text{bigram}}(t_i|t_{i-1}) + bP_{\text{unigram}}(t_1)$ where $a \in R_{\geqslant 0}$, $b \in R_{\geqslant 0}$ are coefficients with the condition $a + b = 1$ $(R_{\geqslant 0} = \{\chi \in R : \chi \geqslant 0\})$. It is easy to prove that $0 \leqslant h_\Theta(x) \leqslant 1$.

Therefore, we can define the target function of the document as

$$cost_y(x) = -(ylogh_\Theta(x) + (1-y)\log(1-h_\Theta(x))) \tag{1}$$

Table 1
Example of vector space model.

ID	1	2	3	4	5	Norm of vector
Term	Please	Contact	Seller	Phone	Number	2.64
Feature weight	0	1	1	2	1	
Naturalized	0	0.38	0.38	0.75	0.38	

where y only has two possible values 0 or 1. $y = 1$ means that term is not in the category, $y = 0$ means the term is in the category.

The cost function is defined below:

$$\min_{\Theta} \frac{1}{N_t} \Sigma_{i=1}^{N_t} [y^{(i)} cost_1(\Theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\Theta^T x^{(i)})] + \frac{1}{2N_t} \Sigma_{j=1}^{N_c} \Theta^2 \quad (2)$$

where N_t is the total number of terms. The cost function is used to find the parameters like Θ to obtain the minimal values of cost function.

3.4.2. SVM

Support vector machine is a supervised learning tool. Therefore, messages in documents will be tagged with classes. We have Positive and Negative options on online comment data, marked as $\{x_i, y_i\}$, $i = 1,2,\ldots,n$, $y_i \in \{1,-1\}$, and x_i are every comment in document D. Suppose there is a perfect hyperplane H: $x + b = 0$ which can correctly partition all the text comments into positive and negative categories. There are two hyperplanes H_1 and H_2 which are parallel to H:

$$\omega \cdot x + b = 1$$

$$\omega \cdot x + b = -1$$

The positive comments that are closest to H are on H_1 and the negative comments that are closest to H are on H_2. Because we expect the largest margin distance, we need to find the smallest ω. It is easier to compute smallest $\frac{\omega^2}{2}$ with the condition $y_i(\omega \cdot x + b) - 1 \geqslant 0$. Therefore using Lagrange Multiplier, we can construct

$$L(\omega, b, a_i) = \frac{1}{2} \cdot //\omega//^2 - \Sigma_{i=1}^{n} a_i(y_i(\omega \cdot x_i + b) - 1) \quad (3)$$

where $a_i \geqslant 0$. We first write $\max_{a_i \geqslant 0} \min_{\omega,b} L(\omega, b, a_i)$ for (3). Therefore, we obtain

$$\frac{\partial L(\omega, b, a_i)}{\partial \omega} = \omega - \Sigma_{i=1}^{n} a_i y_i x_i = 0$$

$$\frac{\partial L(\omega, \mathrm{b}, a_i)}{\partial b} = -\Sigma_{i=1}^{n} a_i y_i = 0$$

We eventually can write $\max_{a_i \geqslant 0} \min_{\omega,b} L(\omega, b, a_i)$ equivalent to

$$\begin{cases} \max \left\{ \Sigma_{i=1}^{n} a_i - \frac{1}{2} \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} a_i a_j y_i y_j (x_i \cdot x_j) \right\} \\ \Sigma_{i=1}^{n} a_i y_i = 0 \\ a_i \geqslant 0 \end{cases} \quad (4)$$

The meaning of (4) is as following. We can compute all the values of a_i with the aid of computers. Therefore, we can compute the values of $\omega = \Sigma_{i=1}^{n} a_i \cdot x_i \cdot y_i$. With the condition, $a_i(y_i(\omega \cdot x_i + b) - 1) = 0$, we can compute the value of b. Therefore, we can find the best hyperplane, H_1 or H_2, to correctly classify the online comments.

3.4.3. Training and predicting

As shown in Fig. 2, the training process includes the following steps:

(1) collecting data from online social networks and saving them as samples;
(2) calling $\chi^2(t, c_i)$ and selecting features that represent data;
(3) generating a feature vector on the basis of extracted features;
(4) calling and preparing parameters for libsvm, which is a library with well-accepted SVM implementation;
(5) generating and saving training models; and
(6) testing the model by using testing data sets to obtain the precision, recall, and F function.

Based on training results, the prediction process is a much easier. The main prediction processes include: (1) generating feature vector from samples and (2) executing SVM models and computing the category indicators. We designed the system to run multiple SVM models on the same training model, as shown in Fig. 3. Multiple SVM computing can be executed by multiple processes and every process is independently calculated. This greatly speeds up the processing speed and reduces computing time.

4. A case study

In this section, we present a case study in which we adopted the proposed bilingual model to process review comments written in both Chinese and English. The objective of the case study is to demonstrate the applicability of our proposed model in practice and to compare the actual performance of the SVM/N-Gram model in handling both English and Chinese languages in a real-world situation. Because practitioners may have a preference to SVM or N-Gram based on their past experience, we provide a comparison of the results generated by SVM and N-Gram. The process we used to assess our proposed model is described below.

First, we manually collected review comments on a very popular movie from social media sites. Specifically, English comments were from Facebook and Twitter while Chinese comments were from Tianya forum and Weixin. We also manually examined the comments to collectively determine whether a comment should be qualified as positive

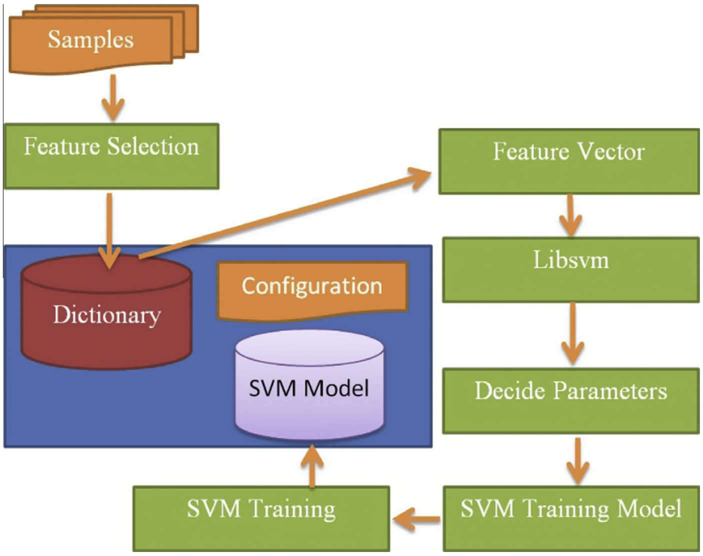


Fig. 2. Training process. *Note:* The arrows indicate the precedence of operations, i.e. execution work flow. Boxes refer to different software modules.

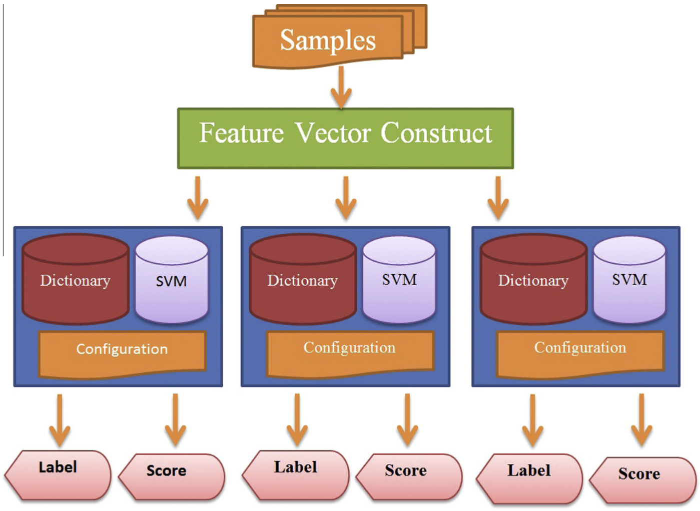
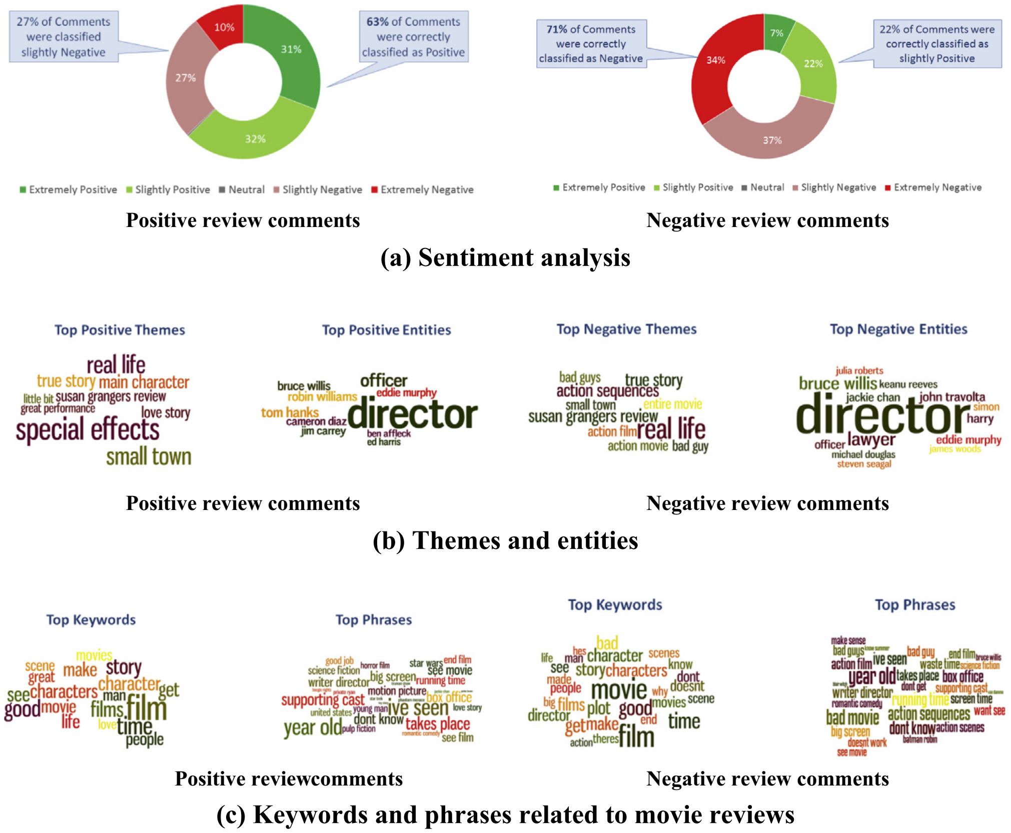


Fig. 3. Prediction by multi-models. *Note:* The arrows indicate the precedence of operations. Boxes refer to different software modules.

or negative. In the end, we collected 2000 English comments (1000 positive comments and 1000 negative comments) and 2000 Chinese comments (1000 positive comments and 1000 negative comments). For both Chinese and English comments, we partitioned 80% of comments as the training data set and 20% of comments as the testing data set. We used both the SVM and N-Gram as the classification model in our customized software. For English comments, we simply segmented words by using blank spaces. For Chinese, we adopted the proposed word segmentation method. The feature vector of comments was then constructed after word segmentation. Thereafter we trained SVM/N-Gram model to generate optimized parameters. Once the training was completed, we feed SVM/N-Gram the testing data to predict the category which the comment belongs to.

Next, we applied the N-Gram model on the English movie comments. We found that about 63% of positive comments were correctly classified and about 71% of negative comments were correctly classified, as shown in Fig. 4a. As the accuracy rate only gives us information regarding the performance of the classification, we do not have details about the content of each category of comments. However, in some circumstances, we may want to understand in more detail what people are feeling. Therefore, we analyzed the theme and entities of each category of comments. The results of theme and entity are shown in Fig. 4b. The more frequently a keyword appears in the comments, the larger the keyword appears in the theme and entity results. We noticed that “director” is the main entity for both positive and negative comments. But the positive comments have “special effects” as the top theme



Positive review comments

Negative review comments

(a) Sentiment analysis

Positive review comments

Negative review comments

(b) Themes and entities

Positive reviewcomments

Negative review comments

(c) Keywords and phrases related to movie reviews

Fig. 4. Results of sentiment analysis of movie reviews.

and the negative comments have “real life” as the top theme.

We were also interested in the language that people used in their comments. We analyzed the keywords and phrases of the movie review comments. We presented the keywords and phrases by using larger fonts if the keywords and phrases were more frequently used in the review comments. The results of keywords and phrases are shown in Fig. 4c. We noticed that the top 3 keywords were “film”, “time” and “good” for positive reviews and “movie”, “film” and “time” for negative reviews. These results, and the overlapping keywords, suggest that whether a film is positively or negatively reviewed is purely subjective. The audiences’ varying sentiments on the same movie are based on their own personal feelings and experiences. Some people like the film because they like the story, characters, special effects, or other aspects of the film and others dislike it for those same reasons.

Our next step was to use three metrics to measure the effectiveness of our proposed method. In text mining technologies, there are three important metrics that are used as the standards to compare the effectiveness of methodologies: precision, recall, and F-score. Thus, we compared the three metrics to demonstrate the effectiveness of our proposed methodologies through this case study.

For classification tasks, there are well accepted metrics: *P* for precision, *R* for recall and *F* for f-function. We define true positives as *tp*, true negatives as *tn*, false positives as *fp*, and false negatives as *fn*. Then precision is defined as:

$$Precision = \frac{tp}{tp + fp}$$

Recall is defined as:

$$Recall = \frac{tp}{tp + fn}$$

F-score is defined as F1 in this paper because recall and precision are equally weighted

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

And finally, accuracy is defined as:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

We performed a classification of the sentiment samples we obtained from the Internet. We extract part of the whole body of comments as the training data set and use the rest as the testing data. There are 64 negative and 65 positive samples in our testing set. The 95% Confidence Interval is 0.8372+/−0.0637.

Since repeated data could change the accuracy of the results in experiments, we intended to find out if this is the case in our bilingual sentiment analysis. We were particularly interested in how repeated samples impact the accuracy of results in SVM Model. We had 500 positive samples and 1000 negative samples in our sample set. We then duplicated the positive samples and merged them with the 1000 negative samples. The whole 2000 samples become our training sample set. Our test sample set includes 494 positive and 606 negative samples. The accuracy results are shown in Fig. 5.

Fig. 5 shows that the *F* and recall *R* values of the non-repeated samples are slightly higher than the repeated samples. Interestingly, the precision of the repeated samples is a little bit higher than the non-repeated samples. This is because the weights of the samples that repeat multiple times have been inflated. If the repeated samples were outliers, it would not significantly impact the results because the punishment parameters can control the outliers. On the other hand, the repeated samples actually increase the total number of samples, compared to the unique samples. The complexity of classification (including training and predicting) in the SVM model is $O(N^3)$. Therefore, duplicated samples will cause a significant increment of computing time.

We then wanted to evaluate the performance of the two data mining models, SVM and N-gram, by comparing the performance on test data. Out of the 2000 comments (1000 for positive and 1000 for negative), there are 400 comment messages (200 for positive and 200 for negative) functioning as testing data and the rest 1600 comments acting as training data. The overall performance of each of the two models is shown in Fig. 6. As seen in the figure, the SVM model generally out-performed N-Gram model because the SVM's values of *F*, *precision*, *recall*, and *accuracy* are all better than the ones of N-Gram.

Fig. 7 describes the results of the positive comments versus all comments. We noticed that the *accuracy* of

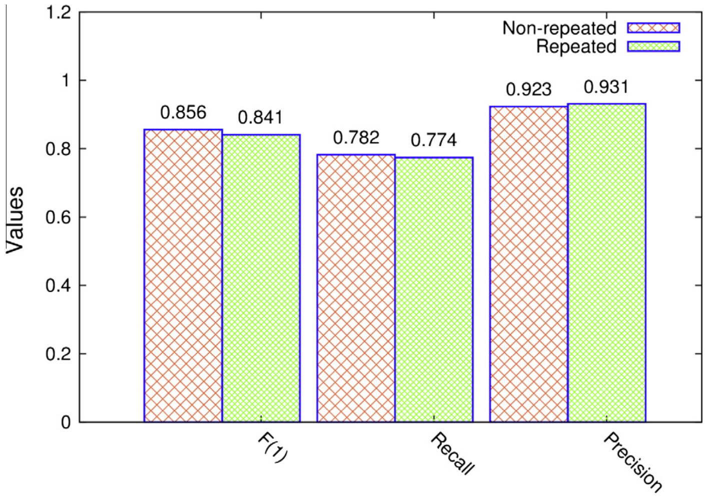


Fig. 5. Impact of repeated samples.

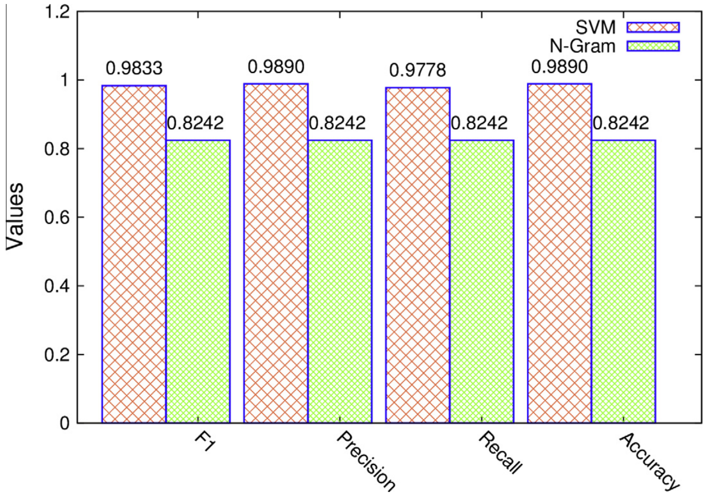


Fig. 6. Overall evaluation.

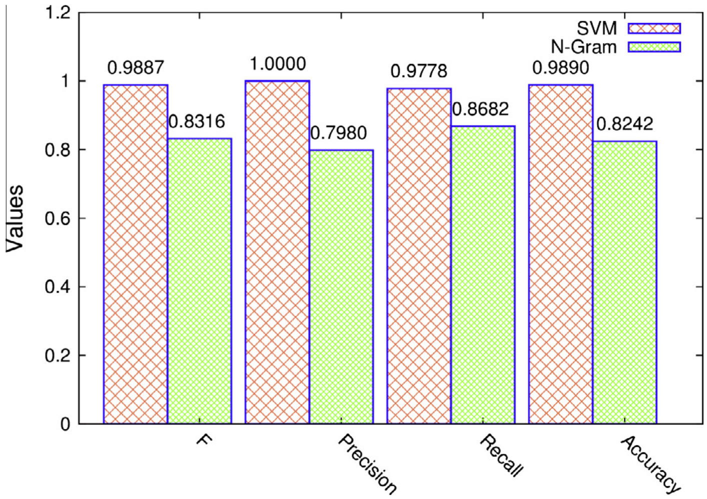


Fig. 7. Positive comments versus all.

SVM is about 20% (i.e. (0.9890 – 0.8242)/0.8242), which is higher than the value of N-Gram. This means that about 20% more positive comments were correctly classified by SVM than N-Gram. The *precision* values of SVM and N-Gram show that SVM recognized all positive comments (i.e. $fp = 0$, false positive is zero), but N-Gram only correctly recognized 79% of positive comments out of all the positive comments marked by the models. The *recall* values of the two models represent the percentage of positive comments that were marked as positive in all the processed comments. Again, the *recall* value of SVM shows there are only two positive messages that were marked negative messages. The values of *accuracy* show that SVM correctly marked positive message and negative messages at a success rate of 98.90%, while the success rate of N-Gram was only 82.42%.

Fig. 8 presents results about the negative comments versus all comments. *Recall* values represented the percentage of negative comments that were marked negative in all the processed comments. Our results show that SVM model out-performed the N-Gram model by more than 28% in terms of *recall*. In Fig. 8, the *recall* value of SVM model showed that false negative $fn = 0$. The *precision* values show that SVM recognized about 20% (i.e. (0.9785 – 0.8161)/0.8161) more negative comments than N-Gram model which correctly recognized 81% of negative comments out of all the comments. The values of *accuracy* showed that the SVM model correctly marked negative message at success rate 98.90% while N-Gram's success rate was 82.42%. Therefore, the *accuracy* of the SVM model out-performed N-Gram by about 20% (i.e. (0.9890 – 0.8242)/0.8242).

As Chinese language segmentation is very important, we conducted an experiment to investigate the performance of our Chinese language segmentation. The goal of this experiment was to show that some words (more than

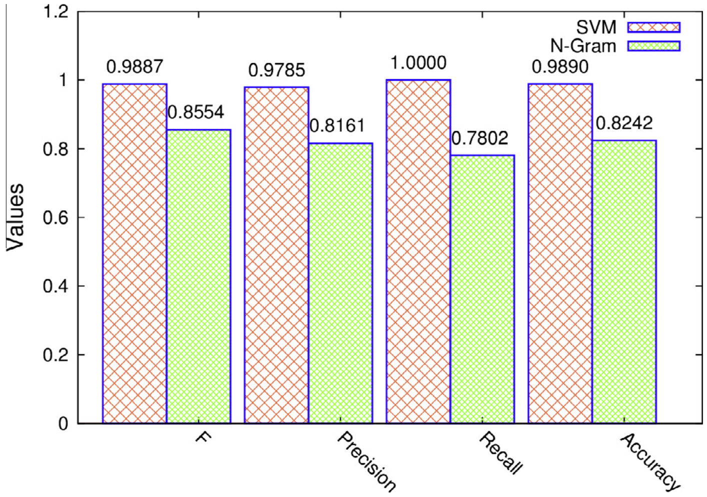


Fig. 8. Negative comments versus all.

Table 2
Comparison of Chinese language segmentation.

Item	Before	After
Precision	83.74	89.98
Recall	98.92	92.14
F1	90.38	91.04

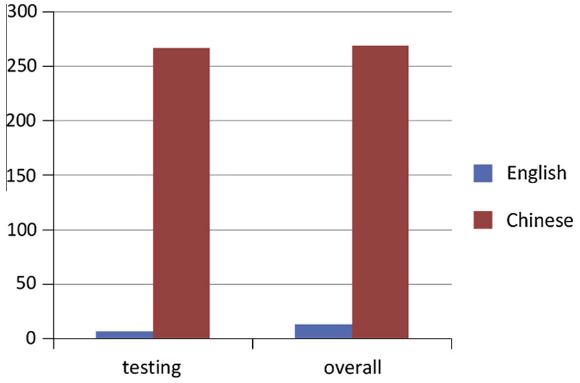


Fig. 9. Overall time consuming comparison.

two characters) can be correctly recognized and marked. In our development configuration file, we are able to edit special phrases which were treated as one word or term in the customized software. We compared the results before specifying the special phrases and after specifying the special phrases. The results are shown in Table 2. As expected, the results after specifying special phrases are much higher than those before specifying. The recall value for the before is 98.92%. The higher recall means the smaller *fn*. The recall value after specifying the special phrases dropped to 92.4%, which means higher *fn*. Meanwhile, the *precision* and *F*1 both increased, which means smaller *fp*.

As to the cost of the classification, we were interested in quantifying the computation time spent on the whole process. As our software has exactly the same code, the only difference is the feeding data set. Therefore, we compared how much time we needed to classify the Chinese reviews and English reviews. The result is shown in Fig. 9.

We expected that processing the Chinese review comments would take more time but we were surprised to see such a drastic difference in time. As shown in Fig. 9, processing the English review comments needed only about 13 s but processing the Chinese review comments needed about 269 s.

Because the values of the classification accuracy of the movie review comments is not very high, we switched to the SVM model and performed the classification again with the same set of review data. We also conducted the classification routines on the Chinese review comments. The classification accuracy of the SVM model is shown in Fig. 10, which is much higher than that of N-Gram. The accuracy increased from 63% to 85%, about a 34% increment. Another interesting fact is that the accuracy of the English comment classification is much better than that of Chinese comment classification. The reason lies in the fact that Chinese segmentation is not 100% accurate. We have about 90% accuracy in Chinese segmentation. In summary, the classification accuracy of Chinese comments is significantly lower than that of English comments.

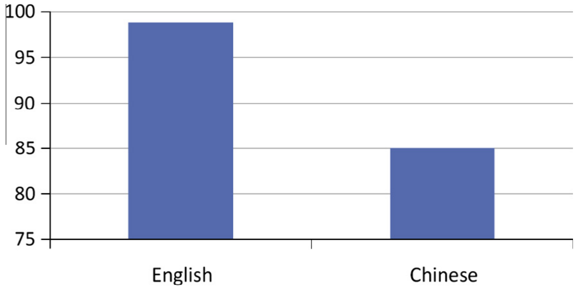


Fig. 10. Overall accuracy comparison.

5. Conclusion

There are many social media sites that use languages other than English, such as Chinese, French, and Arabic. As more and more non-English speaking people use social media, there is a growing need to improve existing sentiment analysis techniques and develop new approaches to handle social media textual content written in foreign languages. In this paper, we proposed a bilingual approach to conducting sentiment analysis on Chinese and English social media. A case study was conducted to explain how the proposed bilingual approach is used to process movie-related review comments we collected from both English and Chinese social media sites. The results show that our proposed approach is effective. Our results also show that the overall accuracy value of SVM is much higher than that of N-Gram in terms of classification performance. In addition, the results show that the accuracy of the English comment classification is a lot better than that of the Chinese comment classification. Our proposed approach is novel and different from the existing research in that the proposed model can process two different languages simultaneously and the module in the proposed model can also be replaced by other implementation methods.

6. Limitation and future research

Several potential limitations merit consideration. First, in this study we only made modifications to adopt two popular text mining models (N-Gram and SVM) for bilingual sentiment analysis. As there are many other methods or techniques to do sentiment analysis, it is possible to integrate those methods or techniques into our proposed model. Thus, one future research direction is to evaluate the different text mining approaches and identify the most effective one in analyzing multi-language social media. Second, although our case study provided some initial evidence for the validity and utility of our proposed approach, future research is needed to further validate our approach. For example, we tested our approach by using a small sample. Future research is needed to evaluate the performance of our proposed approach in handling a larger data set. Third, Chinese word segmentation is a complicated research area and involves complicated natural language processing. There is still a lack of effective visualization tools to generate beautiful clouds or figures for our Chinese sentiment analysis results as we did for English comments. In this study, we only provided the textual results generated from the sentiment analysis of movie review comments written in Chinese. Future research is needed to either find or develop a Chinese language-based visualization tool that can help us vividly display our Chinese results. Fourth, solutions need to be developed to address the limitations of natural language processing, such as the inability to identify irony and sarcasm and low accuracy in Chinese classification. Finally, cultural differences may cause expression differences which could affect the accuracy of the methods. More rigorous studies need to be conducted to address cultural differences in text mining. In addition, future research is needed to extend our approach to other languages so that a flexible sentiment analysis tool for multiple-language processing can be developed.

Despite the limitations, this paper provides a good starting point for building more powerful sentiment analysis tools for processing online content across multiple languages. Such a tool is becoming increasingly necessary as the number of non-English speaking Internet users continues to grow and because more companies are engaging in multilingual marketing via social media to increase global brand awareness.

References

[1] M. Bautin, L. Vijayarenu, S. Skiena, International sentiment analysis for news and blogs, in: ICWSM, April 2008.

[2] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.

[3] D. Bremner, E. Demaine, J. Erickson, J. Iacono, S. Langerman, P. Morin, G. Toussaint, Output-sensitive algorithms for computing nearest-neighbour decision boundaries, Discr. Comput. Geomet. 33 (4) (2005) 593–604.

[4] W.B. Cavnar, J.M. Trenkle, N-gram-based text categorization, in: Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, pp. 61–175.

[5] H. Chen, P. De, Y.J. Hu, B.H. Hwang, Wisdom of crowds: the value of stock opinions transmitted through social media, Rev. Finan. Stud. 27 (5) (2014) 1367–1403.

[6] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[7] N. Dabbagh, A. Kitsantas, Personal learning environments, social media, and self-regulated learning: a natural formula for connecting formal and informal learning, Internet Higher Educ. 15 (1) (2012) 3–8.

[8] D. Gao, F. Wei, W. Li, X. Liu, M. Zhou, Co-training based bilingual sentiment lexicon learning, in: AAAI (Late-Breaking Developments), June 2013.

[9] Z.H. Deng, S.W. Tang, D.Q. Yang, M.Z.L.Y. Li, K.Q. Xie, A comparative study on feature weight in text categorization, in: Advanced Web Technologies and Applications, Springer, Berlin Heidelberg, 2004, pp. 588–597.

[10] C. Haney, Sentiment analysis: providing categorical insight into unstructured textual data, Soc. Media, Soc., Surv. Res. (2014) 35–59.

[11] C. Hawn, Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care, Health Aff. 28 (2) (2009) 361–368.

[12] W. He, S. Zha, L. Li, Social media competitive analysis and text mining: a case study in the pizza industry, Int. J. Inform. Manage. 33 (3) (2013) 464–472.

[13] N.N. Ho-Dac, S.J. Carson, W.L. Moore, The effects of positive and negative online customer reviews: do brand strength and category maturity matter?, J Market. 77 (6) (2013) 37–53.

[14] ICTCLAS, January 2011. <www.ictclas.org/>.

[15] IKAnalyzer, December 2012 <http://code.google.com/p/ik-analyzer/>.

[16] J. Kim, H.Y. Jung, Y. Lee, J.H. Lee, Conveying subjectivity of a lexicon of one language into another using a bilingual dictionary and a link analysis algorithm, Int. J. Comput. Process. Lang. 22 (02n03) (2009) 205–218.

[17] J. Kim, H.Y. Jung, S.H. Nam, Y. Lee, J.H. Lee, Found in translation: conveying subjectivity of a lexicon of one language into another using a bilingual dictionary and a link analysis algorithm, in: Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy, Springer, Berlin Heidelberg, 2009, pp. 112–121.

[18] A. Klein, O. Altuntas, M. Riekert, V. Dinev, A Combined Approach for Extracting Financial Instrument-Specific Investor Sentiment from Weblogs, 2013.

[19] M. Lan, C.L. Tan, J. Su, Y. Lu, Supervised and traditional term weighting methods for automatic text categorization, Pattern Anal. Mach. Intell., IEEE Trans. 31 (4) (2009) 721–735.

[20] H. Lee, K. Choi, D. Yoo, Y. Suh, G. He, S. Lee, The More the Worse? Mining Valuable Ideas with Sentiment Analysis for Idea Recommendation, 2013.

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

[22] B. Liu, Sentiment analysis and subjectivity, Handbook Nat. Lang. Process. 2 (2010) 627–666.

[23] S. Ludwig, K. de Ruyter, M. Friedman, E.C. Brüggen, M. Wetzels, G. Pfann, More than words: the influence of affective content and linguistic style matches in online reviews on conversion rates, J. Market. 77 (1) (2013) 87–103.

[24] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, vol. 1, Cambridge University Press, Cambridge, 2008, p. 6.

[25] J.Y. Nie, M. Simard, P. Isabelle, R. Durand, Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 74–81.

[26] Paoding Analyzer, January 2010. <https://code.google.com/p/paoding/>.

[27] B. Pang, L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2004, p. 271.

[28] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, 2002, pp. 79–86.

[29] F. Peng, F. Feng, A. McCallum, Chinese segmentation and new word detection using conditional random fields, in: Proceedings of the 20th International Conference on Computational Linguistics, Association for Computational Linguistics, 2004, p. 562.

[30] J.M. Schulz, C. Womser-Hacker, T. Mandl, Multilingual corpus development for opinion mining, in: LREC, May 2010.

[31] P.K. Singh, M.S. Husain, Methodological study of opinion mining and sentiment analysis techniques, Int. J. Soft Comput. (IJSC) 5 (1) (2014) 11–21.

[32] R. Sproat, T. Emerson, The first international Chinese word segmentation bakeoff, Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, vol. 17, Association for Computational Linguistics, 2003, pp. 133–143.

[33] S. Stieglitz, L. Dang-Xuan, Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior, J. Manage. Inform. Syst. 29 (4) (2013) 217–248.

[34] J. Stewart, H. Strong, J. Parker, M.A. Bedau, Twitter keyword volume, current spending, and weekday spending norms predict consumer spending, in: Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, IEEE, 2012, pp. 747–753.

[35] X. Sun, H. Wang, W. Li, Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1, Association for Computational Linguistics, 2012, pp. 253–262.

[36] S. Tan, J. Zhang, An empirical study of sentiment analysis for Chinese documents, Expert Syst. Appl. 34 (4) (2008) 2622–2629.

[37] T. Tao, C. Zhai, Mining comparable bilingual text corpora for cross-language information integration, in: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005, pp. 691–696.

[38] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Am. Soc. Inform. Sci. Technol. 61 (12) (2010) 2544–2558.

[39] X. Wan, Bilingual co-training for sentiment classification of Chinese product reviews, Comput. Linguist. 37 (3) (2011) 587–616.

[40] H. Wang, P. Yin, L. Zheng, J.N. Liu, Sentiment classification of online reviews: using sentence-based language model, J. Exp. Theor. Artif. Intell. 26 (1) (2014) 13–31.

[41] J. Yao, G. Wu, J. Liu, Y. Zheng, Using bilingual lexicon to judge sentiment orientation of Chinese words, in: Computer and Information Technology, 2006. CIT'06. The Sixth IEEE International Conference on, IEEE, 2006, p. 38.

[42] A. Zimek, E. Schubert, H.P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min. 5 (5) (2012) 363–387.

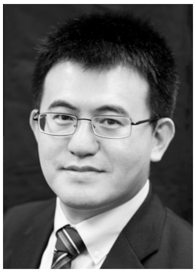
[43] C. Zhang, D. Zeng, J. Li, F.Y. Wang, W. Zuo, Sentiment analysis of Chinese documents: from sentence to document level, J. Am. Soc. Inform. Sci. Technol. 60 (12) (2009) 2474–2487.

[44] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, ACM SIGKDD Explor. Newslett. 6 (1) (2004) 80–89.

Gongjun Yan received his Ph.D. in Computer Science from Old Dominion University in 2010. He is currently an Assistant Professor in University of Southern Indiana. His main research areas include algorithms, security, privacy, routing, and healthcare in Vehicular Ad-Hoc Networks, Sensor Networks and Wireless Communication and Internet.

Wu He is currently an Assistant Professor of Information Technology at Old Dominion University, USA. He earned his PhD in Information Science and Learning Technologies from the University of Missouri. His research interests include data mining, knowledge management, information systems design, case-based reasoning, and information technology education.

Jiancheng Shen is currently a Ph.D. Candidate in Finance at Old Dominion University, USA. His research interests include neuro-finance, media and financial markets, high frequency financial data analysis and international economics.



Chuanyi Tang is currently an Assistant Professor of Marketing at Old Dominion University, Norfolk, VA USA. He received his Ph.D. in Retailing from the University of Arizona. His research interests include services marketing, consumer well-being, and consumer online communication.