Information Discovery and Delivery

Identifying domain relevant user generated content through noise reduction: a test in a Chinese stock discussion forum

Xiangbin Yan, Yumei Li, Weiguo Fan,

Identifying domain relevant user generated content through noise reduction: a test in a Chinese stock discussion forum

Xiangbin Yan
University of Science and Technology, Beijing, China
Yumei Li
Harbin Institute of Technology, Harbin, China, and
Weiguo Fan
Department of Accounting and Information Systems, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA

Abstract

Purpose – Getting high-quality data by removing the noisy data from the user-generated content (UGC) is the first step toward data mining and effective decision-making based on ubiquitous and unstructured social media data. This paper aims to design a framework for revoking noisy data from UGC.

Design/methodology/approach – In this paper, the authors consider a classification-based framework to remove the noise from the unstructured UGC in social media community. They treat the noise as the concerned topic non-relevant messages and apply a text classification-based approach to remove the noise. They introduce a domain lexicon to help identify the concerned topic from noise and compare the performance of several classification algorithms combined with different feature selection methods.

Findings – Experimental results based on a Chinese stock forum show that 84.9 per cent of all the noise data from the UGC could be removed with little valuable information loss. The support vector machines classifier combined with information gain feature extraction model is the best choice for this system. With longer messages getting better classification performance, it has been found that the length of messages affects the system performance.

Originality/value – The proposed method could be used for preprocessing in text mining and new knowledge discovery from the big data.

Keywords Social media, User-generated content, Feature selection, Text classification, Domain lexicon, Noise reduction

Paper type Research paper

1. Introduction

With the emergence of information technology and applications, the total amount of global data has showed unprecedented explosive growth. At the same time, the data become increasingly more complex. The low value density is one of the characteristics of big data (Feng *et al.*, 2013), which means that there is much noise among the data we are really interested in. For example, in a video recording of uninterrupted monitoring having duration of 1 h, there may be only a few seconds containing potentially useful information. Getting rid of the noise information is an important step to go through the huge volume of data to extract useful information.

With the advent of Web 2.0, the internet has evolved to support multimedia-rich content delivery, end user personal content generation and community-based social interactions (Zhang *et al.*, 2011c). Online reviews, Web discussions and blog articles have become important channels for users to publish and share information, which lead to the explosive growth of user-generated content (UGC). This provides us with a great opportunity to explore the mass point of views on the Web (Liu *et al.*, 2007), and assess the value of these UGC. Web 2.0 sites may accumulate opinions from participants including any types of internet user, such as customers and investors of a company. These discussions could provide insights on many perspectives, such as consumers' view on products, stock values and public opinions on social events. This crowd-sourced information could provide us solutions that might answer our questions in a better way, and may provide a new perspective to the whole world. For example, Liu proves that online reviews offer significant explanatory power for both aggregate and weekly box office revenue (Liu, 2006). Previous research has shown that consumer behavior is increasingly influenced by peer opinions (Smith, 2009), which means that consumers' reviews are valuable to both companies and consumers. The communities powered by UGC have become prevalent and useful tools for knowledge acquisition, exchange and collaborative

The current issue and full text archive of this journal is available on Emerald Insight at: **www.emeraldinsight.com/2398-6247.htm**

Information Discovery and Delivery
45/4 (2017) 181–193
© Emerald Publishing Limited [ISSN 2398-6247]
[DOI 10.1108/IDD-04-2017-0043]

Received 20 April 2017
Revised 24 June 2017
Accepted 30 June 2017

decision-making (De Valck *et al.*, 2009). It is very important to get the valuable information quickly and accurately.

The UGC is always unstructured and often covers a variety of topics in different knowledge domains. The communities are always filled with news, pictures, rumors, advertisements, grumbles and something meaningless, as people can freely publish any information they want, some of which are even irreconcilable. Most of the time, we are only interested in a small portion of all the data. Most of the data will become noise, as we do not need it. Removing noisy data is an important step to perform any meaningful analysis in social media analytics process (Fan and Gordon, 2014). With the rapid expansion of digital content in social media community, the phenomenon of information overload has become extremely severe, which makes it even more difficult to get valuable information through the noise data. This makes an interesting yet challenging test bed for data mining applications.

However, it is often a very difficult and daunting task to identify valuable information from huge quantity of UGC, which cannot be handled manually. First, the amount of text content is overwhelming, so that it requires automation. Second, we need to handle the complexity of Chinese language processing. Chinese language processing is still a great challenge because of the special character encoding and linguistic characteristics (Zheng *et al.*, 2006). To the best of our knowledge, the detailed and rigorous studies for identifying and removing noise from massive UGC are scanty, especially in a Chinese language context. Getting the valuable information from the UGC through noise reduction is the central focus of this study.

Generally, the messages we need in a community could be attributed to one or several topics. We divide the messages into two major categories: the topic *relevant* messages that we are interested in and the topic *non-relevant* ones that we are not interested in. The topic non-relevant messages are the noise for us to get the valuable information. We propose a framework for removing non-relevant information, and construct the automatic Relevant Message Identification System (RMIS) based on text mining methods. Then we test the method on a Chinese stock Web forum. We compare experimental results from different perspectives, including feature selection methods, classification algorithms and their best combinations. Experimental results show that support vector machines (SVM) with the Information Gain (IG) feature selection obtain the best performance with the F-measure$_{\mathrm{REL}}$ of 88.2 per cent. We also study how the message length affects the result, which is one of the characteristics of the text itself. To the best of our knowledge, this is one of the first few studies to address data quality issues in online social media data. Our proposed approach of noisy data reduction can also be used as a reference for dealing with noisy data in other domains.

The rest of this paper is organized as follows. Section 2 lists some related works. In Section 3, we present the research design for topic relevant message identification by removing non-relevant noise data from the Web forum. A noise data automatic identification system is proposed and the details of the system are presented. Section 4 shows the experiment procedure. In Section 5, we discuss some of the experimental results followed by conclusions and future work in Section 6.

2. Related work

2.1 UGC and its impacts

With the advent of Web 2.0, UGC allows users to express their creativity and publish their comments on anything imaginable, in the form of photographs, videos, podcasts, articles, blogs, etc. A growing number of academic and commercial applications are tapping into the rich UGC. Moe *et al.* (Moe and Schweidel, 2012) study individuals' decision to contribute a product rating on an e-commerce platform based on UGC, and develop a model to examine whether to contribute as well as what to contribute. Bagozzi and Dholakia prove that online community participants are more likely to adopt suggestions and decisions of people who are similar to them (Bagozzi and Dholakia, 2002). Some notable examples are highlighted below. Abrahams *et al.* (2012) predict the existence of safety and performance defects across multiple brands of automobiles with UGC from online forums. Godes and Mayzlin (2004) show that a measure of the dispersion of conversations across communities has explanatory power in a dynamic model of TV ratings. Clemons *et al.* (2006) find that the variance of ratings and the strength of the most positive quartile of reviews play a significant role in determining which new products grow fastest in the marketplace. Liu *et al.* (2007) mine sentiment information from blogs and investigate ways to use such information for predicting product sales performance. Liu (2006, 2011) uses Yahoo movie site reviews to explain the earnings of the film's box office. Liu *et al.* (2007) study the online product reviews and how they affect consumers from a marketing strategy perspective. Recently, more and more scholars have focused on stock-related UGC, and investigated their impact on the stock market (Antweiler and Frank, 2004; Sabherwal *et al.*, 2008).

2.2 Removing noisy UGC data

The poor data quality has become an obstacle to taking advantage of the massive data in databases, data warehouses and information systems in every domain (Ouzzani *et al.*, 2013). Researchers have explored methods for removing various noise. Bertaglia and Nunes take advance of text normalization to remove the noise, the non-standard words like spelling errors, abbreviations, mixed case words, acronyms, internet slang, hashtags and emoticons, which are often in documents, especially the informal UGC (Bertaglia and Nunes, 2016). B clean boilerplate, remove non-UGC and delete duplicates from the consumer reviews to get the not relevant for the investigation of consumer perceptions (Egger and Schoder, 2017). These are not enough for getting through the big data to get the real information we want, which need to understand the messages and get the treasures from the messes. C even choose the messages manually (Schmunk *et al.*, 2014).

There has been attempts to remove noise information to get useful information from the massive text messages based on the understanding of the documents. The earliest attempts in noise reduction were done in the spam filtering area, where a combination of rule-based and content-based filtering strategies was proposed and used to filter spam emails

(Cunningham *et al.*, 2003). Within the rule-based method, a set of rules is applied to a message and a score is accumulated based on the rules. If the score for an email exceeds a threshold, the email is identified as spam. The features are extracted from the header of the email to help the classification in the content-based filtering strategy (Wang and Pan, 2005). There are a few attempts focusing on excluding noise and error messages from UGC data. Most existing researches just consider it as a pre-step of other data mining tasks, such as in research (Li and Wu, 2010; Devi and Bhaskarn, 2012) non-relevant postings are manually removed. Because there are more noise and error messages in small stocks bulletin board, Sabherwal *et al.* only include stocks that are on the TheLion.com's list of ten most actively discussed stocks in their experiment (Sabherwal *et al.*, 2008). Zhang *et al.* (2011a) make use of the similarity measurement method in removing domain non-relevant messages. First, they obtain words with great frequency from the sample set as the features set, then compute the similarity between the features set and the testing messages. If the similarity between a message and the features set were less than a given threshold, the message will be identified as noise. Das and Chen (2007) treat spam messages as neutral messages when classifying the messages into three classes and do no further processing with the spam messages. Rinser aligned information of the same thing across different language versions on Wikipedia to improve the quality of the information (Rinser *et al.*, 2013). Hu *et al.* (2016) proposed a semantics-based text classification technology to filter the spam, which select the related feature terms from the semantic meanings of the text content. Most of works on. As can be seen from the above discussions, more research is needed to help cope with the ever-increasing noisy data in UGCs for data or text mining applications.

2.3 Text classification

Text classification has recently received a lot of attention from both the academic and business community. It can automatically classify a given text document into known classes using advanced machine learning and AI techniques (Fan *et al.*, 2006). Text classification is typically formulated as a supervised learning task, in which a classifier learns how to distinguish between class labels in a given training set, using features automatically extracted from a collection of documents (Shi *et al.*, 2010). The procedure of text classification includes three steps: preprocessing text data, building classification model and performance evaluation (Guansong and Shengyi, 2012).

The preprocessing of text data is to represent the text to be effectively recognized by the text classifier, and it includes the text representation and feature selection. Currently, the Vector Space Model (VSM) is the main method adopted for text representation (Salton *et al.*, 1975). There are also several methods for text representation based on the semantic level, including Latent Semantic Indexing (Zhang *et al.*, 2011b), Locality Preserving Indexing, multi-words (Zhang *et al.*, 2011b). However, their effectiveness needs to be further tested. There are several feature selection methods, which are widely used, such as IG, Mutual Information (MI), Chi-square statistical methods (CHI) and expected cross entropy (Yang and Pedersen, 1997).

The classification algorithms are core parts in text classification. Some popular text classification algorithms include probability-based classification algorithms, e.g. Naive Bayesian (Lewis and Ringuette, 1994) and multivariate regression models (Schütze *et al.*, 1995), machine learning-based algorithms, e.g. Support Vector Machine (Platt, 1998), Neural Networks (Wiener *et al.*, 1995), K-Nearest Neighbor algorithm (Aha, 1992) and Decision Tree (Quinlan, 1993). Researchers show the SVM got better result than others (Haddoud *et al.*, 2016). The text classification models based on neural networks have become increasingly popular (Conneau *et al.*, 2016). Although these models get very good performance, the high training and testing time limit their use on very large data sets (Joulin *et al.*, 2017). Researcher try to combine classifiers, which show better performance than individuals (Jain and Mandowara, 2016).

The most frequently used Classification assessment methods is F-measure, which is based on the Recall and Precision, and considers the size of the class effect on the classification results (Guansong and Shengyi, 2012). Micro-averaging F-measure Value and the Macro-averaging F-measure Value are used to assess the overall classification performance based on multiple categories corpus (Guansong and Shengyi, 2012). The Break-Even Point is used to assess the classification result (Sebastiani, 2002).

Text classification is one of the most useful techniques for text mining and witnesses a large number of real-world applications. Text classification techniques have been used to solve many practical problems, such as topic classification (Joachims, 1997), spam detection (Sahami *et al.*, 1998), opinion identification (Pang and Lee, 2004), gender and age classification (Schler *et al.*, 2006), filtering spam messages in mobile phone (Liu and Yang, 2012), vehicle defect discovery using social media data (Abrahams *et al.*, 2012) and stock abnormal returns prediction using news data (Luss and d'Aspremont, 2012). We will employ text classification in our noise reduction efforts for UGCs in this paper.

3. Research design

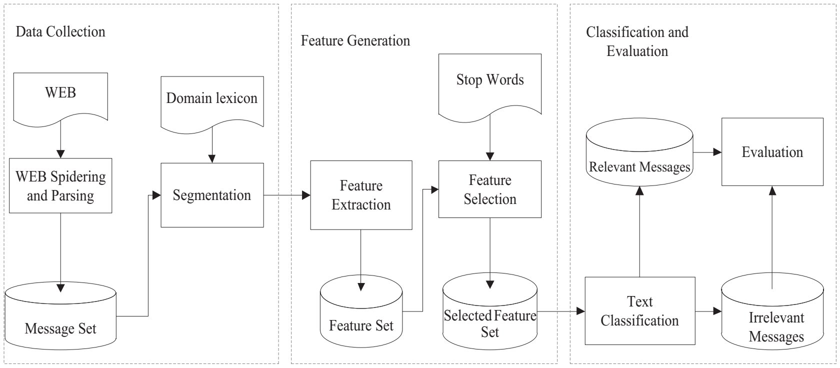
3.1 System overview

The RMIS we propose for identifying relevant information consists of three components as shown in Figure 1. The first one is the data collection component. In this part, we get the published messages from the Web using crawlers and store them in a relational database. With the aid of the domain lexicon, we performed Chinese word segmentation and POS Tagging with the software *ICTCLAS* (Institute of Computing Technology Chinese Lexical Analysis System) from Chinese Academy of Sciences. The second component of the system is feature generation. Each message is converted into a vector of words (VSM) which are the attributes of the message. Some stop words are excluded from the feature set. Not every word is included in the last classification steps. We engage the feature selection to remove some useless words. Finally, we identify messages that are topic relevant by text classifiers in the third component of our system. We next provide more details about each of the components.

3.2 Data collection

This component consists of the following steps: message collecting, message parsing and Chinese word segmentation.

Figure 1 System architecture



First, we download messages from the Web and store them in the database. Then we make use of the ICTCLAS to do Chinese segmentation with the aid of a domain lexicon dictionary.

3.2.1 Messages acquisition

Several tools can be used for getting messages from the Web. There are many crawlers for collecting data from the Web and many of them are free. The APIs are provided by many websites, which are convenient for us to crawl the data on the web. However, the most flexible method is writing a crawler ourselves.

3.2.2 Chinese word segmentation and POS tagging

Unlike English text in which sentences are sequences of words delimited by spaces, in Chinese text, sentences are represented as strings of Chinese characters or *hanzi* without similar natural demarcation. Therefore, identifying the sequence of words in a sentence and marking boundaries in appropriate places are the first step in a Chinese language processing task (Xue, 2003; Pengyu *et al.*, 2014). We use the automatic Chinese lexical analyzer ICTCLAS to undertake Chinese Word Segmentation and part-of-speech tagging. ICTCLAS is one of the best and free Chinese Lexical Analysis System developed by Institute of Computing Technology (Liu *et al.*, 2004). But it has the limited ability in identifying new words (Cao *et al.*, 2006). However, UGC is casually expressed and contains many newly created words. To improve the accuracy of Chinese word segmentation and POS tagging, a dictionary of commonly used Web words is adopted to identify the new words. The dictionary contains commonly used Web words extracted from *Sogou* Labs (www.sogou.com) and contains 130,000 words, which frequently appear on the Web.

As mentioned in the previous section, the RMIS identify the topic relevant messages from any other messages. Messages contain some special words in different topics from different domains. For example, the automobile reviews and stock reviews will contain their own domain-specific words, which is rarely used in other domains. The tools like ICTCLAS could not identify most of the domain-specific words. A domain lexicon of the concerned topic is helpful to get the exact words in the topic relevant messages. With the domain lexicon, the RMIS could identify the topic-related words and get the exact feature from the messages. The lexicon varies as the concerned topic changes.

3.3 Feature generation

3.3.1 Feature representation

To accomplish automatic text classification, we have to convert the set of documents into an acceptable representation that the learning machine can handle. The most common, simple and successful document representation so far is the VSM (Salton *et al.*, 1975). Each document is represented as a point in a vector space with one dimension for a term in the vocabulary. There are several possible representations for each word in the documents, such as Boolean representation and the term frequency. These methods tend to give equal importance to all terms.

Considering the idea that terms occurring in fewer documents are better ones, we adopt the *tf-idf* as the representation for each word in a document (Basili *et al.*, 1999).

3.3.2 Stop words removal

Some words get a very high frequency in the documents. However, not every word has the distinguishing potential between classes. One of the common characteristic of these frequent words is that they carry little information about the document. We usually refer to this set of words as stop words (Zou *et al.*, 2006). Silva verified that the application of stop words reduces the dimension of feature space and has a positive effect on improving the accuracy of a text classifier (Silva and Ribeiro, 2003). Since there is not a widely recognized Chinese stop words list, we treat a word as a stop word if the word appears in a large number of documents with a very high document frequency (DF) as Joachims did in (Joachims, 1999). The topic-special words may also get a high DF in the documents, so we will keep the word out of the stop words list if the word appears in the domain lexicon. We also filter out those words that appear in very few examples (documents) since they will unlikely represent a class category. With the frequency threshold we remove the ultra-low frequency words from the word set (Wang, 2005). Punctuations and numbers are also deleted from the messages.

3.3.3 Feature selection

There are still tens of thousands of words left after removing stop words and extremely low-frequency words. Not every word is useful in identifying the documents. Combining all words into the training process will lead to huge time cost and the decline of classification performance. Feature selection could help to improve the prediction performance of the classifiers by choosing the most useful predictors (Guyon and Elisseeff, 2003). A variety of feature selection methods can be used to improve the accuracy of text classification. Existing experiments show that IG (Yuan *et al.*, 2013), MI and CHI (Wu and kang, 2006) are the most effective methods (Yang and Pedersen, 1997).

For ease of explanation, we use Table I to show the term distribution information among training documents. A is the number of documents that contain term t and belong to the class C_i. B is the number of documents that contain term t but not belong to the class C_i. C is the number of documents that do not contain term t but belong to the class C_i. A is the number of documents that neither contain term t nor belong to the class C_i.

IG evaluates the worth of an attribute by measuring the information gain with respect to the class for feature selection (Crăciun *et al.*, 2006; Fan *et al.*, 2005).

$$IG(t) = \frac{1}{N}\left(A \log \frac{A}{A+B} + B \log \frac{B}{A+B} + C \log \frac{C}{C+D} + D \log \frac{D}{C+D}\right) \tag{1}$$

The CHI measures the correlation between class C and features t. Higher CHI means greater correlation between the feature and the class, thus the feature carrying more information can differentiate the classes. Yang and Pedersen (1997) have proved that IG and CHI are among the best feature selection methods, and cost almost the least time.

$$\chi^2(t, C_i) = \frac{N \times (AD - CB)^2}{(A+B) \times (B+D) \times (A+C) \times (C+D)} \tag{2}$$

$$\chi^2(t) = \max_{i=1}^{M} \{\chi^2(t, C_i)\} \tag{3}$$

MI is widely used in the statistical language model. A greater MI represents a higher degree of co-occurrence between feature t and category C_i, and the feature provides more information for C_i. To improve the representation of the selected vocabulary, we introduce the improved MI (Wu and kang, 2006).

$$MI(t, C_i) = \log_e\left(1 + \frac{A \times N}{(A+C) \times (A+B)} * TF(w, C_i)\right) \tag{4}$$

$$MI_{avg}(t) = \sum_{i=1}^{M} p(C_i) MI(t, C_i) \tag{5}$$

Table I The distribution of term t

	Class C_i	Not class C_i
Contains term contains term t	A	B
Does not contain term t	C	D

$$\sigma(w) = \sqrt{\sum_{i=1}^{m=1} (MI(w, C_i) - MI_{avg}(w))^2} \tag{6}$$

3.4 Classifiers and performance metrics

A wide range of supervised learning algorithms has been proposed for text classification. In this paper, we choose four classifiers, which have shown good performance in previous studies, to process the message information, including the Naïve Bayesian (Naïve Bayes), decision tree classifier (J48), SVM and k-nearest neighbor classification (KNN). The optimal classifier selected is then used as the classifier for non-relevant message identification and noise reduction.

3.4.1 Naïve Bayes

The Naïve Bayes is a probabilistic classifier based on the Bayes' theorem with an assumption of independence. The assumption is that a particular feature presenting or not in a class is unrelated to any other feature.

3.4.2 J48

J48 is an implementation of C4.5 release 8 (Quinlan, 1993) that produces decision trees. It adopts a greedy search technique and generates decision trees from top to bottom.

3.4.3 SVM

The SVM is a supervised learning algorithm. With the nonlinear mapping algorithm, the SVM transforms the low-dimensional nonlinear samples into high-dimensional feature space, to make the samples linearly separable. Then it constructs the optimal partition hyper plane in the feature space based on the structural risk minimization theory. The SVM classification is carried out by using a linear kernel with the Sequential Minimal Optimization algorithm (Platt, 1998).

3.4.4 KNN

The KNN classifies examples based on the nearest training examples, which is an instance-based learning method (Aha, 1992). An example is classified based on the voting result of the k nearest examples whose correct classes are known.

Precision and recall are used to evaluate the performance of the text classifiers for relevant messages identification (Sokolova and Lapalme, 2009). We adopt the F-measure$_{REL}$ (the F-measure of the relevant class) to evaluate the RMIS, in which we pay more attention to the recall and the precision of the valuable messages. All the experiments are done via tenfold cross-validation to insure the robustness of performance.

4. Experiment

In the previous section, we have described how the RMIS can be used for identifying topic relevant messages. The success of this approach will depend on the proportions of noise messages removed and relevant ones reserved. In this section, we will examine the performance of the RMIS as described above.

4.1 Data

4.1.1 Data preparation

We choose the financial Web forum *Eastmoney*[1] as our data sources. There are many UGCs for stocks, funds, bonds, etc. that could provide valuable information for investors, companies, the government, etc. Since anybody could become a reader and a contributor for the UGC, messages on any topics are published without any limitation. There is a wide variety of noise data that should be removed to get the valuable information we really care. Thus, getting the quality data with minimal noise is the first step for a successful research study. Only after removing the noise messages from the collection, can we perform meaningful decision-making and data mining tasks, such as extracting public opinions from news media, extracting trading signals about investors' sentiments and behaviors from UGC to predict the stock price trend.

The data set used for evaluating the system we propose comprises messages related to Shanghai Stock Exchange (SSE) Composite Index[2], which is the most important index in the Chinese stock market. We collect over 30,000 messages from the *Tangulunjin* forum of *eastmoney* (http://bbs.eastmoney.com) which provides a virtual space for investors to share information about SSE Composite Index. The messages cover one and a half years period, between April 1, 2009 and September 30, 2010. All data extracted are stored in a repository for later processing.

Samples of the extracted messages are shown as below. The first two messages express extreme optimism about stock market and encourage others to buy the stocks. They provide the information for investors' opinions to the index. While the last two talks about something non-relevant with the SSE Composite Index. The third one talks about a teleplay and the fourth one shows an advertisement of the speaking zone. The last two are non-relevant to the topic of SSE Composite Index, which we do not care in studying SSE Composite Index.

- 九月将是艳阳天.今天大盘走的死气沉沉，只不过是为了做月线收盘价而已, 明天将在权重股的带领下重拾升势, 开启金九升途。
- 预测会沿着 5 天均线上行，反弹预计会创 3123 点以上的新高。
- 关公面前耍大刀，来呀，温酒。。。最近在读三国，新三国拍的真不好看。
- 大碗茶聊吧-----请进　　把“快乐投资、快乐投机、快乐友情、快乐人生”作为办吧主题思想，为朋友们提供一个敞开心扉、互相交流、诉说甜酸苦辣的平台。

We download the posts with java programs and store the messages in the MySQL for further experiment.

4.1.2 Data labeling

The messages are unstructured and expressed in a casual way. There is no obvious symbol of the Composite Index non-relevant or relevant for each message. We have to distinguish the meaning of the messages and mark the messages as the SSE Composite Index relevant or not by ourselves. Considering the workload and precision of calibrating, we randomly select 1,000 messages from the 30,000 messages. Then we use the majority voting method to label the messages. Five experts with financial knowledge background are invited to label every message we selected with as the SSE Composite Index "relevant" or "non-relevant". If at least four experts mark a message in the same class, we will adopt it. Otherwise, we will give up the message and randomly select a new one from the rest of messages.

The detail criteria for messages labeling are listed below. Messages talking about *SSE* Composite Index are marked with "relevant"; it may be bullish, bearish, neutral, hearsay, news or talking about the past index. Messages not talking about *SSE* Composite Index are marked with "non-relevant", which may include anything like jokes, advertisement.

In the end, there are 363 messages identified as non-relevant messages, indicating a high noise ratio (more than 36 per cent) in the forum data.

4.2 Feature generation and selection

4.2.1 Chinese word segmentation and POS tagging

Considering the limitation of the ICTCLAS in identifying new words and the need of identifying topic relevant messages, a dictionary of commonly used Web words from Sogou Lab and a lexicon from finance domain are combined in the process of word segmentation. Words in the finance lexicon are significant in identifying whether a message is talking about the stock market, which is our topic of concern. The finance lexicon contains over 20,000 financial field words and was previously created by Liang (2009). Some of the words in the lexicon are shown below:

4.2.1.1 Finance lexicon. 利多, 利空, 多头, 空头, 反弹, 盘整, 死多头, 多翻空, 短多, 斩仓, 割肉, 套牢, 多杀多, 热门股, 对敲, 筹码, 踏空, 跳水, 诱多, 骗线, 阴跌, 停板, . . .

We compare the results of word segmentation before and after the usage of the finance thesaurus that we have mentioned. We can see that, when combining the finance thesaurus, some financial words can be segmented correctly. See Table II for an example.

Words marked with "DICT" are included in the finance thesaurus. We can see that "均" and "线" are combined as one word "均线" after the usage of the finance thesaurus, which word is frequently used in the financial domain. "预测", "反弹", "上行" and "预计" are included in the thesaurus, are still segmented correctly as whole words after we use the thesaurus. The thesaurus helps a lot in identifying the topic related words and further for the topic-related documents.

4.2.2 Represent the messages

We represent every message with VSM. After segmenting words with ICTCLAS, we count the frequency of every word that appears in the message set. There are 19,904 distinct words appearing 224,371 times in total in the 1,000 messages. Then we calculate the DF of every word and list the words in a descending order. There are several financial words in Table III, which also appear at the top of DF ranking list. In this paper, we choose the top 30 words of the DF list with removing words that are included in the domain dictionary. Table IV shows the final stop words list. They appear 39,324 (17.79 per cent of all the words) times in total and have little information.

We also delete punctuations and words that have a frequency less than three. Then we have 5,811 words as initial attributes. Overall, we remove 14,403 (account for 72.36 per cent of the total) words, which appear 68,032 times. We effectively accomplish the goal of dimension reduction.

To further improve the classification performance and reduce the time consumed in training the model, we have to select an optimal subset of features using feature selection methods. We use IG, MI and CHI to conduct feature selection, and build the selected feature sets. We can get three ranked lists of all the attributes with the feature selection

Table II An example of the messages

	The content
The message	预测会沿着5天均线上行反弹预会创3123点以上的新高
Before combining the finance thesaurus	预测/v 会/v 沿着/p 5/b 天/n 均/a 线/n 上行/v, /n 反弹/v 预计/v 会/v 创/v 3123点/t 以上/f 的/u 新高/n
After combining the finance thesaurus	预测/DICT 会/v 沿着/p 5/b 天/n 均线/DICT 上行/DICT, /n 反弹/DICT 预计/DICT 会/v 创/v 3123点/t 以上/f 的/u 新高/n

Notes: Every word and its POS are split with "/"; "v" is short of verb, which represents the word is marked as a verb; "p" is short of preposition; "b" means the word is marked as distinguishing word; "n" is short of noun; "t" is short of time; "f" means the word is marked as locative; "u" means the word is marked as auxiliary word; "DICT" means the word is included in the finance thesaurus

Table III Words at the top of the DF descending order list

Words	DF	Words	DF	Words	DF	Words	DF
大盘	276	市场	271	股	258	涨	225

methods. We construct 20 feature sets with the number of attributes from 100 to 2000 on each feature selection method respectively. We use *tf-idf* as the weight of words to represent the information of each word contained effectively. We use the *Weka* package for all classification tasks. Tenfold cross-validation is used to train and test the performance of the classifiers.

5. Experimental results and discussion

5.1 Classification results and analysis

We followed the standard text mining process in the experiments (Fan *et al.*, 2006). We performed the Chinese word segmentation and removed some stop words, low-frequency words and punctuations. The feature selection was accomplished with IG, CHI and MI. Then we represented the training data with *tf-idf* as the weight of words in VSM. With 20 selected feature sets per feature selection method, 60 selected feature sets were gotten. Finally, the four classifiers were applied to all selected feature sets in turn.

5.1.1 Performance comparison of different classifiers

The performance results (measured by F-measure$_{REL}$) of the four classifiers under different experimental conditions are summarized in Figures 2-5. From the four figures, we can see that the classifiers based on different feature selectors show similar trends. We can get several observations from Figure 2:

- The three feature selectors get similar F-measure$_{REL}$ based on the four classifiers respectively. IG is a little better for SVM and J48 than the other two feature selectors. CHI is more suitable for Naïve Bayes than other feature selectors, and MI is more suitable for KNN.
- The number of features has an impact on the F-measure$_{REL}$ of the four classifiers. As the number of features increases, the F-measure$_{REL}$ of SVM, Naïve Bayes and J48 reach the highest point, then begins to decline. However, the F-measure$_{REL}$ of KNN drops rapidly as the number of attributes increases.
- The SVM performs better with any given feature selection method than without. For Naïve Bayes, IG and CHI could help improve the performance a little bit, but MI could not. As we can see in Figure 2(b), it seems the feature selectors could not provide any help for J48. Figure 2(d) shows that the feature selectors could help improve the F-measure$_{REL}$ of KNN when the number the features is less than 600.

We can see from the results that the first 100 features contain the most information for classification. SVM could improve its performance with more features, and still keep effective as the number of the features increase to 2,000. Naïve Bayes gets the maximum F-measure$_{REL}$ when the number of features reaches 300. However, the number of features nearly has no effect on J48. The first 100 features could provide J48 enough information for classification; more features do not help and even confuse the J48 classifier. The number of features has the greatest effect on KNN, as the number of features increase, the classifier continually drops in performance. Among the four classifiers, SVM and J48 are the most robust ones.

5.1.2 Performance comparison of different feature selection methods

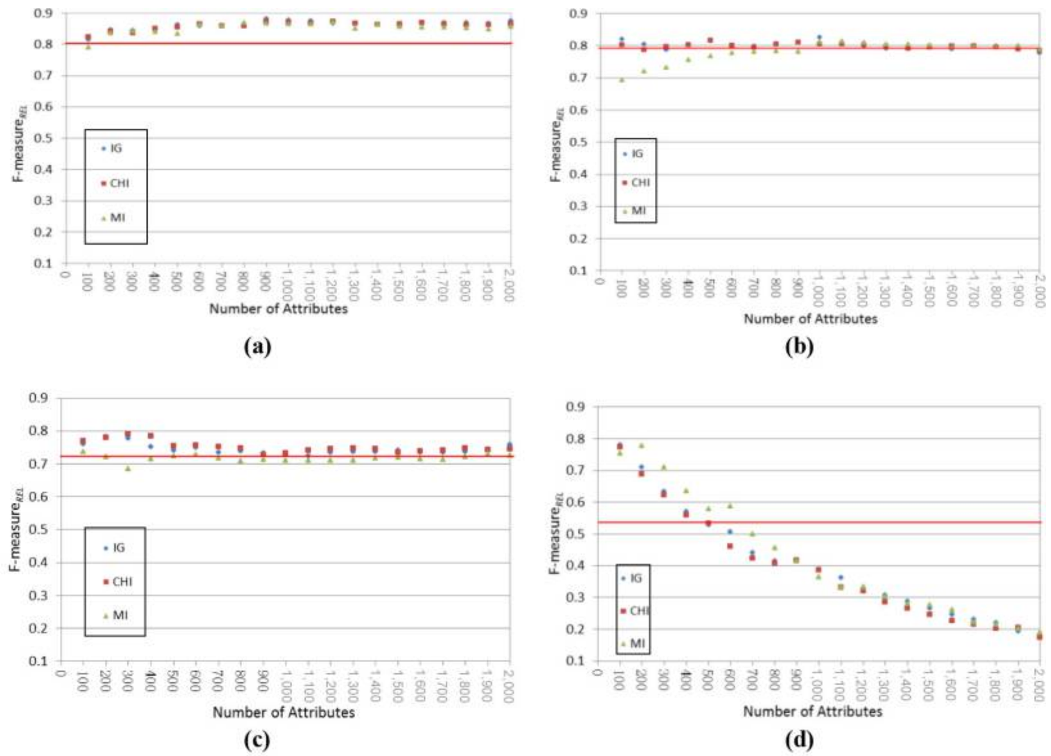
The performance (measured by F-measure$_{REL}$) of the three feature selectors combined with different classifiers is summarized in Figures 3. From the figures, we can see that various classifiers indeed produce different performance for RMIS.

- Clearly, SVM gets the best performance among the four classifiers based on all the feature selectors, followed by J48, Naïve Bayes and KNN. As the number of attributes

Table IV Stop words

Stop words	DF	Stop words	DF	Stop words	DF	Stop words	DF	Stop words	DF
的	676	一	386	上	315	到	276	要	253
是	548	也	343	将	294	这	271	为	253
了	507	大	338	和	293	后	270	但	253
不	486	会	325	下	283	个	265	而	250
在	475	就	321	中	281	都	262	从	230
有	410	多	315	看	281	还	254	人	222

Figure 2 F-measure$_{REL}$ for RMIS with classifiers



Notes: The additional red horizontal line is the F-measure$_{REL}$ base on 5,811 features (without the three feature selectors mentioned above); (a) SVM; (b) J48; (c) Naïve Bayes; (d) KNN

increases, the performance list ordered by F-measure$_{REL}$ stays the same.

- As the feature set size increases, the performance of SVM, Naïve Bayes and J48 improve initially, and then decline after reaching a peak. However, the performance of *KNN* declines directly as the feature set size increases.
- The SVM and J48 get a higher F-measure$_{REL}$ with larger feature set size. The Naïve Bayes and KNN, on the other hand, perform best with a very small feature set size, which means that more features would confuse the classifier. We can conclude that the feature set size we need mainly depends on the classifier, and little on the feature selectors.

5.1.3 The combination of classification algorithms and feature selection methods

In Table V, we list the best performance for every combination of the classifier and the feature selector. In each cell of the table, two values are reported. The first value is the maximum F-measure$_{REL}$ of the combination among the 20 features sets. The second one is the size of feature set obtaining the maximum F-measure$_{REL}$.

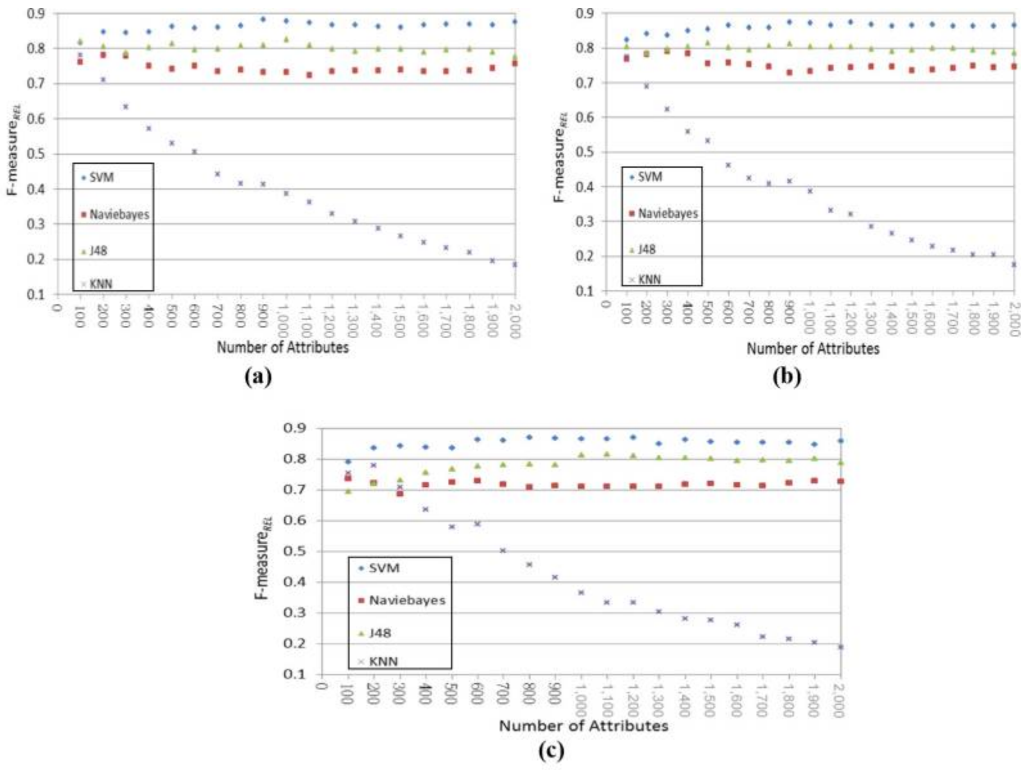
- This table clearly shows that SVM is the best among all classification algorithms. With 84.9 per cent non-relevant messages identified and 90.8 per cent relevant messages saved, the SVM could remove the noise most effectively. The next ones are J48, KNN and Naïve Bayes.
- *IG* is the best for SVM, J48 and KNN; CHI is best for Naïve Bayes.
- The best feature set size depends on the chosen classification algorithm. A feature set with smaller size is more suitable for Naïve Bayes and KNN, but SVM and J48 need a much larger feature set. With a greater N, the SVM and the J48 could take advantage of more information of the train messages.

Our results show that SVM + IG yields the best combined performance. This is consistent with the result from supervised classification result comparison (Du *et al.*, 2014). Although most of the relevant messages could be identified from the messages, there are still messages that are classified into the wrong class. This depends on many factors, such as the feature selectors we choose, the number of attributes we keep or the classifiers we apply. There are also other reasons that could affect the results. First, the Web content in UGC contains many new words and expressions invented by users, which could limit the training classification performance if we do not get enough training data that cover the entire corpus. Second, the system we proposed does not consider the context or semantics of the words, and treats each message as a bag of words. This could potentially lead to misinterpretation of the messages and an inaccurate classification result.

5.2 Message length effect on classification results

From the results of last section, we know that the SVM classifier with IG gets the best performance. Generally, longer messages contain more information. If the message contains more information, it will be easier to identify it into the right

Figure 3 F-measure$_{REL}$ for RMIS with feature selectors



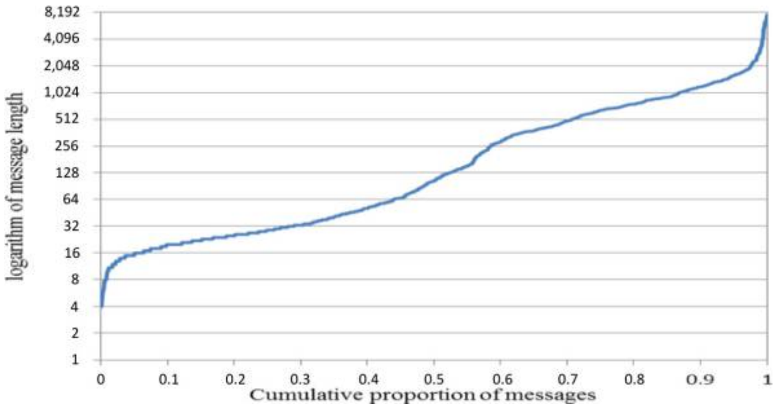
(a)

(b)

(c)

Note: (a) IG; (b) CHI; (c) IG

Figure 4 The distribution of length of all the messages



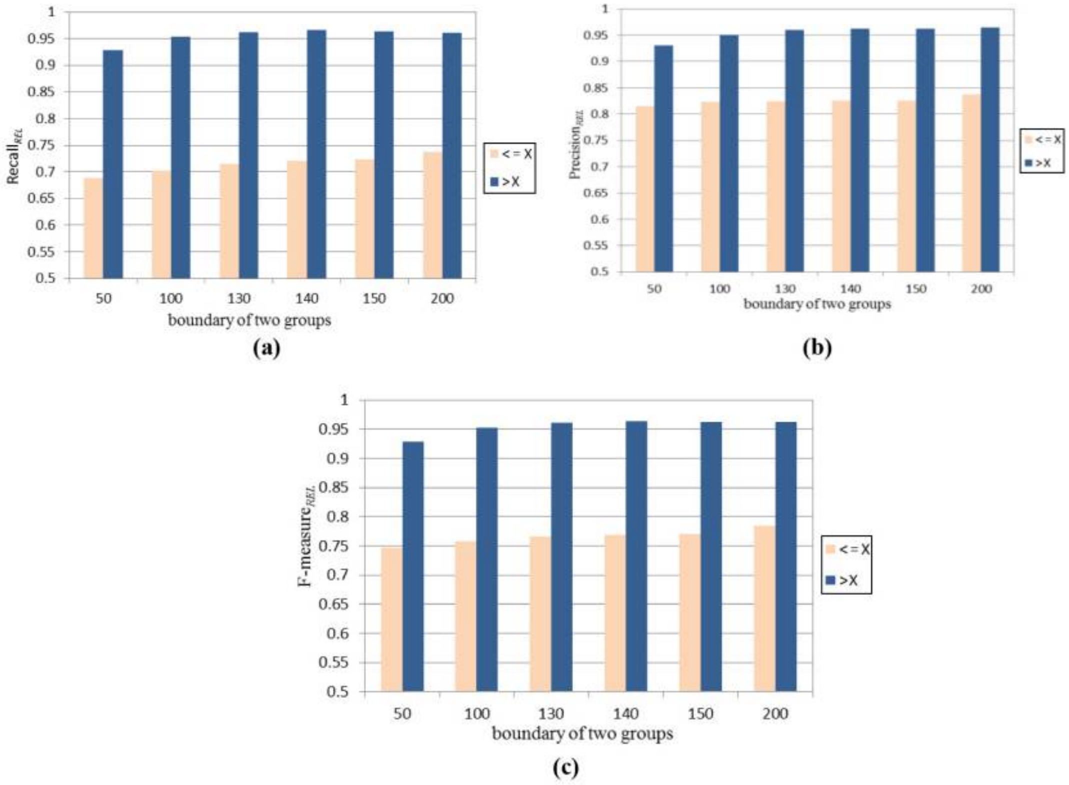
class. In this section, we will discuss how message length affects the performance of classifiers.

First, we display the distribution of the message length in Figure 4. Nearly all the messages are shorter than 2,000 words. We can see that more than 80 per cent of all the messages have a length of less than 1,000 Chinese characters. About 50 per cent of messages are shorter than 100 characters. The distribution of the message length is very uneven.

To study whether the length of the message has an effect on the performance of the classifier, we compare the result of classification based on the different message length. All the messages are divided into two groups according to the length: messages longer than x (x is defined as the boundary of the two groups) are in one group ($G_{long,x}$); and the remaining messages are in another group ($G_{small,x}$).

- In Figure 5(a), the recall of relevant class (Recall$_{REL}$) is shown as the length x changes. The Group $G_{long,50}$ could get a Recall$_{REL}$ more than 0.92 and have little change as the x increases. The groups $G_{small,x}$ get the Recall$_{REL}$ no more than 0.75. The largest gap (0.25) between the $G_{small,x}$ and $G_{long,x}$ happens when x is 130. We can conclude that messages shorter than 50 cause the small Recall$_{REL}$.
- Figure 5(b) shows the precision of relevant class (Precision$_{REL}$). We can see that $G_{long,x}$ gets Precision$_{REL}$ as high as 0.95 as x increases to 100. Also the Precision$_{REL}$

Figure 5 Evaluation on messages sets with different lengths



Notes: (a) Recall_{REL}; (b) Precision_{REL}; (c) F-measure_{REL}

Table V The maximum $\text{F-measure}_{\text{REL}}$ for every combination

	SVM	Naïve Bayes	J48	KNN
IG	0.882	0.782	0.825	0.781
N	900	200	1,000	100
CHI	0.874	0.790	0.816	0.775
N	1,200	300	500	100
MI	0.871	0.737	0.815	0.779
N	1,500	100	1,100	100

of $G_{\text{small,x}}$ is higher than 0.8. That is to say, the length has more effect on the Recall_{REL} than the Precision_{REL}.

- As Figure 5(c) shows, the groups $G_{\text{long,x}}$ appear to be identified much more easily; with 140 words, we get the highest F-measure_{REL} of 0.9625. In contrast, the $G_{\text{small,x}}$ gets very low F-measure_{REL}, which is no more than 0.8.

We can see from the above results that the message length does have a great effect on the result of classification, especially the recall of relevant messages. Longer messages get a better classification performance. The shorter messages contain less information about the topic we care, and then it is difficult to identify them from the noise.

6. Conclusions, limitations and future works

In this paper, to answer the call for more research in dealing with low value density issue in big data, we propose to view the noise as the concerned topic non-relevant messages and build a systematic framework for identifying domain relevant data from UGC. We provide sufficient evidence that our system can identify most of the useful information from the UGC. The framework is the preprocessing for the new knowledge discovery from the big data both in research and business work.

There is no apparent difference among the feature selectors based on the results of the classification. However, the feature selectors could really help improve the result as the feature size changes. Performance from different classifiers tends to be different. The empirical result shows SVM is better than other classifiers for identifying domain relevant data. We need to choose the right feature selector for the right classifier. We found that SVM combined with IG is the best combination in our system.

We also study how the message length affects the classification performance. We find that longer messages get better classification performance. Shorter messages lower the quantity of valuable information we get, but they do not affect the quality of the information we get. Messages shorter than 50 words are the real reason of the poor performance. How to improve the classification of the short messages remains a challenging problem for future research.

Limitations do exist in our study. In our experiment, only 1,000 messages were included because of the needed accuracy and the heavy workload of human annotation. We studied only one large online Web forum. Messages from diversified forums could be used to perform a more comprehensive study. More research is certainly needed to deal with the ever-growing internet slang or new words invented by users. In addition, the results from the short messages were not as good as from the long messages, and more in-depth research needs to be done in this area.

Notes

1 Eastmoney (www.eastmoney.com/), China's most visited and most influential internet financial media according to Alexa.

2 SSE Composite Index is an index of all stocks (A shares and B shares) that are traded at the Shanghai Stock Exchange.

References

Abrahams, A.S., Jiao, J., Wang, G.A. and Fan, W. (2012), "Vehicle defect discovery from social media", *Decision Support Systems*, Vol. 54 No. 1, pp. 89-97.

Aha, D.W. (1992), "Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms", *International Journal of Man-Machine Studies*, Vol. 36 No. 2, pp. 267-287.

Antweiler, W. and Frank, M.Z. (2004), "Is all that talk just noise? The information content of internet stock message boards", *The Journal of Finance*, Vol. 59 No. 3, pp. 1259-1294.

Bagozzi, R.P. and Dholakia, U.M. (2002), "Intentional social action in virtual communities", *Journal of Interactive Marketing*, Vol. 16 No. 2, pp. 2-21.

Basili, R., Moschitti, A. and Pazienza, M.T. (1999), *A Text Classifier Based on Linguistic Processing*, *International Joint Conference on Artificial Intelligence*, *Sweden*.

Bertaglia, T.F.C. and Nunes, M.D.G.V. (2016), "Exploring word embeddings for unsupervised textual user-generated content normalization", *Proceedings of the 2nd Workshop on Noisy User-generated Text*, *Osaka, Japan*.

Cao, Y., Cao, Y., Jin, M. and Liu, C. (2006), "Information retrieval oriented adaptive Chinese word segmentation system", *Journal of Software*, Vol. 17 No. 3, pp. 356-363.

Clemons, E.K., Gao, G.D. and Hitt, L.M. (2006), "When online reviews meet hyper differentiation: a study of the craft beer industry", *Journal of Management Information Systems*, Vol. 23 No. 2, pp. 149-171.

Conneau, A., Schwenk, H., Barrault, L. and Lecun, Y. (2016), "Very deep convolutional networks for text classification", *arXiv preprint*, arXiv: 1606.01781.

Crăciun, M., Cocu, A., Dumitriu, L. and Segal, C. (2006), "A hybrid feature selection algorithm for the QSAR problem", *International Conference on Computational Science*, *No.* 1, pp. 172-178.

Cunningham, P., Nowlan, N., Delany, S.J. and Haahr, M. (2003), "A case-based approach to spam filtering that can track concept drift", *International Conference on Case-Based Reasoning*, *Trondheim, Norway*.

Das, S.R. and Chen, M.Y. (2007), "Yahoo! For Amazon: sentiment extraction from small talk on the web", *Management Science*, Vol. 53 No. 9, pp. 1375-1388.

De Valck, K., Van Bruggen, G.H. and Wierenga, B. (2009), "Virtual communities: a marketing perspective", *Decision Support Systems*, Vol. 47 No. 3, pp. 185-203.

Devi, K.N. and Bhaskarn, V.M. (2012), "Text sentiments for forums hotspot detection", *International Journal of Information Sciences and Techniques*, Vol. 2, pp. 53-62.

Du, M., Pierce, M., Pivovarova, L. and Yangarber, R. (2014), "Supervised classification using balanced training", *International Conference on Statistical Language and Speech Processing*, pp. 147-158.

Egger, M. and Schoder, D. (2017), "Consumer-oriented tech mining: integrating the consumer perspective into organizational technology intelligence-the case of autonomous driving", *Proceedings of the 50th Hawaii International Conference on System Sciences*, *Hilton Waikoloa Village, Hawaii*.

Fan, W. and Gordon, M.D. (2014), "Unveiling the power of social media analytics", *Communications of the ACM*, Vol. 57 No. 6, pp. 74-81.

Fan, W., Gordon, M.D. and Pathak, P. (2005), "Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison", *Decision Support Systems*, Vol. 40 No. 2, pp. 213-233.

Fan, W., Wallace, L., Rich, S. and Zhang, Z. (2006), "Tapping the power of text mining", *Communications of the ACM*, Vol. 49 No. 9, pp. 76-82.

Feng, Z., Guo, X., Zeng, D., Chen, Y. and Chen, G. (2013), "On the research frontiers of business management in the context of Big Data", *Journal of Management Sciences in China*, Vol. 16 No. 1, pp. 1-9.

Godes, D. and Mayzlin, D. (2004), "Using online conversations to study word-of-mouth communication", *Marketing Science*, Vol. 23 No. 4, pp. 545-560.

Guansong, P. and Shengyi, J. (2012), "A summary of research on automatic text classification technologies", *Information Studies: Theory & Application*, pp. 123-128.

Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection", *The Journal of Machine Learning Research*, Vol. 3 Nos 7/8, pp. 1157-1182.

Hu, W., Du, J. and Xing, Y. (2016), "Spam filtering by semantics-based text classification", *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, pp. 89-94.

Jain, A. and Mandowara, J. (2016), "Text classification by combining text classifiers to improve the efficiency of classification", *International Journal of Computer Application*, Vol. 6 No. 1, pp. 126-129.

Joachims, T.(1997), "DTIC document: A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", *14th International Conference on Machine Learning*, *Nashville, TN*, pp. 143-151.

Joachims, T. (1999), "Transductive inference for text classification using support vector machines", in *Proceedings of the Sixteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers, pp. 200-209.

Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. (2017), "Bag of tricks for efficient text classification", *Conference of the European Chapter of the Association for Computational Linguistics*, *Valencia, Spain*.

Lewis, D.D. and Ringuette, M. (1994) "A comparison of two learning algorithms for text categorization", *Third Annual Symposium on Document Analysis and Information Retrieval*, pp. 81-93.

Li, N. and Wu, D.D. (2010), "Using text mining and sentiment analysis for online forums hotspot detection and forecast", *Decision Support Systems*, Vol. 48 No. 2, pp. 354-368.

Liang, X. (2009), *Foundation of Internet Financial Information Intelligent Mining*, Peking University Press, Beijing.

Liu, G. and Yang, F. (2012), "The application of data mining in the classification of spam messages", *International Conference on Computer Science and Information Processing (CSIP), IEEE*, pp. 1315-1317.

Liu, Q., Zhang, H., Yu, H. and Cheng, X. (2004), "Chinese lexical analysis using cascaded hidden Markov model", *Journal of Computer Research and Development*, Vol. 41 No. 8, pp. 1421-1429.

Liu, Y. (2006), "Word of mouth for movies: its dynamics and impact on box office revenue", *Journal of Marketing*, Vol. 70 No. 3, pp. 74-89.

Liu, Y. (2011), "Word-of-mouth for movies: its dynamics and impact on box office revenue", *Journal of Marketing Research*, Vol. 70 No. 3, pp. 74-79.

Liu, Y., Huang, X., An, A. and Yu, X. (2007), "ARSA: a sentiment-aware model for predicting sales performance using blogs", in *The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 607-614.

Luss, R. and D'aspremont, A. (2012), "Predicting abnormal returns from news using text classification", *International Workshop on Advances in Machine Learning for Computational Finance*, London, UK.

Haddoud, M., Mokhtari, A., Lecroq, T. and Abdeddaïm, S. (2016), "Combining supervised term-weighting metrics for SVM text classification with extended term representation", *Knowledge and Information Systems*, Vol. 49 No. 3, pp. 909-931.

Madnick, S.E., Wang, R.Y., Lee, Y.W. and Zhu, H. (2009), "Overview and framework for data and information quality research", *Journal of Data and Information Quality (JDIQ)*, Vol. 1 No. 1, p. 2.

Moe, W.W. and Schweidel, D.A. (2012), "Online product opinions: incidence, evaluation and evolution", *Marketing Science*, Vol. 31 No. 3, pp. 372-386.

Ouzzani, M., Papotti, P. and Rahm, E. (2013), "Introduction to the special issue on data quality", *Information Systems*, Vol. 38 No. 6, pp. 885-886.

Pang, B. and Lee, L. (2004), "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", in *The 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, p. 271.

Pengyu, L., Jingchuan, P., Du Mingming, L.X. and Lijun, J. (2014), "A lexicon-corpus-based unsupervised Chinese word segmentation approach", *International Journal on Smart Sensing & Intelligent Systems*, Vol. 7 No. 1, pp. 263-282.

Platt, J. (1998), "Sequential minimal optimization: a fast algorithm for training support vector machines", Technical Report 98-14, Microsoft Research, Redmond, WA.

Quinlan, J.R. (1993), *C4. 5: Programs for Machine Learning*, Morgan Kaufmann.

Rinser, D., Lange, D. and Naumann, F. (2013), "Cross-lingual entity matching and infobox alignment in Wikipedia", *Information Systems*, Vol. 38 No. 6, pp. 887-907.

Sabherwal, S., Sarkar, S.K. and Zhang, Y. (2008), "Online talk: does it matter?", *Managerial Finance*, Vol. 34 No. 6, pp. 423-436.

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. (1998), "A Bayesian approach to filtering junk e-mail", *Proceeding of the AAAI-98 Workshop on Learning for Text Categorization*, AAAI Technical Report WS-98-05, Madison, WI.

Salton, G., Wong, A. and Yang, C.S. (1975), "A vector space model for automatic indexing", *Communications of the ACM*, Vol. 18 No. 11, pp. 613-620.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006), "Effects of age and gender on blogging", *The AAAI Spring Symposium Computational Approaches to Analyzing Weblogs*, Menlo Park, CA, pp. 191-197.

Schmunk, S., Höpken, W., Fuchs, M. and Lexhagen, M. (2014), "Sentiment analysis – extracting decision-relevant knowledge from UGC", *Information and Communication Technologies in Tourism*, Dublin, Ireland.

Schütze, H., Hull, D.A. and Pedersen, J.O. (1995), "A comparison of classifiers and document representations for the routing problem", *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM*, pp. 229-237.

Sebastiani, F. (2002), "Machine learning in automated text categorization", *ACM Computing Surveys (CSUR)*, Vol. 34 No. 1, pp. 1-47.

Shi, L., Mihalcea, R. and Tian, M. (2010), "Cross language text classification by model translation and semi-supervised learning", *The Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA*, pp. 1057-1067.

Silva, C. and Ribeiro, B. (2003), "The importance of stop word removal on recall values in text categorization", in *The International Joint Conference on Neural Networks*, 20-24 July 2003, *IEEE*, Vol. 3, pp. 1661-1666.

Smith, K. (2009), "The wisdom of crowds", *Nature Reports Climate Change*, Vol. 3, pp. 89-91.

Sokolova, M. and Lapalme, G. (2009), "A systematic analysis of performance measures for classification tasks", *Information Processing and Management*, Vol. 45 No. 4, pp. 427-437.

Wang, B. and Pan, W. (2005), "A survey of content-based anti-spam email filtering", *Journal of Chinese Information Processing*.

Wang, X. (2005), "Dictionary-free Chinese words acquisition method based on bigram", *Computer Engineering and Applications*, pp. 177-179.

Wiener, E., Pedersen, J.O. and Weigend, A.S. (1995), "A neural network approach to topic spotting", *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, Citeseer*, pp. 317-332.

Wu, J. and Kang, Y. (2006), "Text classification based on improved mutual information feature selection", *Journal of Computer Applications*, Vol. 26, pp. 172-173.

Xue, N. (2003), "Chinese word segmentation as character tagging", *Computational Linguistics and Chinese Language Processing*, Vol. 8 No. 1, pp. 29-48.

Yang, Y. and Pedersen, J.O. (1997), "A comparative study on feature selection in text categorization", in *The Fourteenth International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, pp. 412-420.

Yuan, M., Ouyang, Y.X. and Xiong, Z. (2013), "A text categorization method using extended vector space model by frequent term sets", *Journal of Information Science and Engineering*, Vol. 29 No. 1, pp. 99-114.

Zhang, J., Peng, Q. and Kang, X. (2011a), "Research on tendency classification algorithm for online movie comment", *Computer Engineering and Applications*, Vol. 47 No. 11.

Zhang, W., Yoshida, T. and Tang, X. (2011b), "A comparative study of TF* IDF, LSI and multi-words for text classification", *Expert Systems with Applications*, Vol. 38 No. 3, pp. 2758-2765.

Zhang, Y., Dang, Y. and Chen, H. (2011c), "Gender classification for web forums", *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, pp. 1-10.

Zheng, R., Li, J., Chen, H. and Huang, Z. (2006), "A framework for authorship identification of online messages: writing style features and classification techniques", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 3, pp. 378-393.

Zou, F., Wang, F.L., Deng, X. and Han, S. (2006), "Automatic identification of Chinese stop words", *Research on Computing Science*, Vol. 18, pp. 151-162.

Corresponding author

Yumei Li can be contacted at: lym_27@126.com