Decision Support Systems 105 (2018) 1–12

Contents lists available at ScienceDirect

Decision Support Systems

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

Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums

Yao Liu [a], Cuiqing Jiang [a,*], Huimin Zhao [b]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, China
[b] Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P. O. Box 742, Milwaukee, WI 53201, USA

ARTICLE INFO

Article history:
Received 19 February 2017
Received in revised form 16 October 2017
Accepted 17 October 2017
Available online 20 October 2017

Keywords:
Contextual features
Multi-view ensemble learning
Product defect identification
Social media

ABSTRACT

As social media are continually gaining more popularity, they have become an important source for manufacturers to collect information related to defects on their products from consumers. Researchers have started to develop automated models to identify mentions of product defects from social media, such as online discussion forums. In this paper, we propose a novel method for product defect identification from online forums, addressing two inadequacies in previous studies, namely, the inadequate use of information contained in replies and the straightforward use of standard single classifier methods. Our method incorporates contextual features derived from replies and uses a multi-view ensemble learning method specifically tailored to the problem on hand. A case study in the automotive industry demonstrates the utilities of both novelties in our method.

1. Introduction

Product defects have severe negative effects on product competitive advantage. Identifying product defects promptly and accurately can help manufacturers conduct quality management and improve product competitive advantage [1,2]. Especially in independent industries of developing countries, product defects are more prevalent due to the lack of proven technologies, and enterprises need to pay more attention to product quality management and marketing competitive intelligence.

Traditionally, product defect information collection sources have been mainly quality tests and feedback from after-sales service centers. Product defect information collection modes based on such traditional information sources have the shortcomings of high cost, incomprehensiveness, and hysteresis. Nowadays, more and more consumers share product defects they encounter and express personal opinions using social media [3–5]. Consumers can express their opinions freely without disclosing their true identities and without fear of undesirable consequences on social media [6]. Thus, social media, e.g., online forums, provide novel sources for manufacturers to obtain valuable product defect information. The mode of product defect information collection from social media has the advantages of being low-cost, comprehensive, spontaneous, effective and prompt [7].

Data on social media are unstructured and voluminous. Sifting through the vast volume of social media data to identify the mentions of product defects is a daunting task. In response to the problems of unstructured data and information overload, researchers have applied machine learning techniques to build automated product defect identification models, which can help manufacturers reduce labor costs significantly [8,9]. Specifically, several studies (e.g., [2,9,10,11,12]) have investigated the automatic identification of product defects from online discussion forums. These studies have treated product defect identification as a classification problem, i.e., classifying a discussion thread as defect-related or otherwise based on a set of features characterizing the thread.

We observe two inadequacies in existing studies, in the way features are constructed and the classification methods used, respectively. Previous studies have shown the usefulness of several categories of features, such as linguistic features, social features, and distinctive terms [10]. However, they did not distinguish the replies from the original post in a discussion thread and used the entire thread as a single unit in constructing the features. We posit that it is useful to explore the distinctive information contained in the replies and that contained in the original post. Replies are related to, and also complement, the original post. The replies have strong correlation with the original post in a thread. For example, if a consumer complains about a possible product defect, readers are likely to reply with information about the severity of the defect, possible solutions, suggestions, or similar complaints. If a consumer consults about product price, the repliers will provide price and discount related information. Therefore, the content of replies can reflect whether the original post pertains to a product defect or not to a certain extent. The replies are also different from the original post. They have different roles in a discussion thread, where the repliers express their opinions and suggestions to the original post. Consequently, they tend

* Corresponding author.
E-mail addresses: jiangcuiq2017@163.com (C. Jiang), hzhao@uwm.edu (H. Zhao).

https://doi.org/10.1016/j.dss.2017.10.009
0167-9236/

to exhibit different linguistic characteristics and use different vocabularies. In addition, replies may have some unique features that are not available on the original post.

Previous studies have used standard single classifier methods for product defect identification [2,10]. There are multiple categories of features that can be used in product defect identification, with possible dependence across the categories. The high dimensionality and dependence may cause difficulties for single classifier methods, adversely affecting the performance of defect identification.

We strive to bridge these two gaps by proposing a novel method for product defect identification from online discussion forums. Our method uses features derived from replies, referred to as *contextual features*, to better capitalize on the information contained in the replies, which reinforces and complements the information contained in the original posts. We also propose a multi-view ensemble learning method specifically for product defect identification, to deal with the high dimensionality of the feature space and possible dependence across feature categories. We have applied and evaluated our proposed method in a case study in the automotive industry. The results show that both novelties in our method helped to improve performance.

The rest of the paper is organized as follows. Section 2 reviews related work on social media text classification, especially product defect identification, and multi-view ensemble learning. Section 3 presents our proposed method. In Section 4, we report on the case study and evaluation results. Finally, we conclude the paper by summarizing our contributions and discussing potential future research directions.

2. Related work

In the last few years, several studies (e.g., [2,9,10,11,12]) have investigated the automated identification of product defects from social media, such as online discussion forums. These studies have been conducted in such domains as automotive [2,9,10], consumer electronics [10], appliance [11], and toy [12]. Most of these studies have formulated the problem as a classification problem, sharing some similarities to other social media text classification problems, such as online review helpfulness or usefulness prediction [4,7,13–15]. In this section, we briefly review the features and methods that have been used in social media text classification, with an emphasis on product defect identification. We then introduce multi-view ensemble learning, which inspires our proposed method for product defect identification from online forums.

2.1. Social media text classification

The objective of social media text classification is to classify a piece of social media text (e.g., an online product review and a post at an online discussion forum) into several predefined classes (e.g., useful/useless and defect/non-defect) based on a set of features using a classification method. The classification performance largely depends on the choice of features and classification methods.

Features are quantitative metrics that describe an unstructured piece of qualitative (textual) data [4,16]. Various types of features have been used in social media text classification. Specifically, in the context of product defect identification, Abrahams et al. [10] broadly divided the features into the following seven major categories: lexical features, stylistic features, social features, sentiment features, distinctive terms, product features, and semantic features.

1. *Lexical features* are typically the presence or frequency of unique terms (e.g., words, phrases, or named entities), widely used in any type of text classification. As the number of unique terms is typically large, a feature selection method is needed to reduce the dimensionality [17,18].
2. *Stylistic features* reflect the writing style (e.g., number of unique words, average number of words per sentence, and average number of sentences per paragraph) and readability. Linguistic features that have been used in review helpfulness prediction are closely related and may be considered part of this category [14]. Linguistic features describe the characteristics of the vocabulary and the format of the text content.
3. *Social features* reflect the social characteristics (e.g., activeness, credibility, expertise, and social influence) of the authors of the social media content. Zheng et al. [15] showed that social features are important in deriving better classification results in classifying the quality (useful/useless) of online reviews.
4. *Sentiment features* measure the subjectivity, sentiment polarity (e.g., positive, negative, and neutral), or rating. These may be generated using a simple sentiment lexicon or a sophisticated sentiment analysis tool.
5. *Distinctive terms* occur more prevalently in a particular class of texts in a particular domain. Some examples are the so-called “smoke” words, which are positively associated with defect posts, and “sparkle” words, which indicate consumer satisfaction and product compliance to specifications. Lists of distinctive terms have been crafted for such domains as automotive [2,9,10], appliance [11], and toy [12].
6. *Product features* are structured data characterizing a product. Such product characteristics are presented as tags to a post.
7. *Semantic features* measure the occurrence frequency of concept classes, after mapping words to semantic categories (i.e., concept classes).

There are of course different ways to categorize features. For example, Figueiredo et al. [19] categorized features associated with a social media object into content features, textual features, and social features. The lexical features, stylistic features, sentiment features, distinctive terms, and semantic features in the framework of Abrahams et al. [10] may all be considered textual features. The product features in the framework of Abrahams et al. [10] may be considered content features.

The classification methods used in previous studies on social media text classification, especially product defect identification, have been typically standard single classifier methods. Some methods used for product defect identification are naïve Bayes [2,10], support vector machines [2,10], and logistic regression [10]. As discussed earlier, there are multiple categories of features representing the social media text to be classified. The social media text classification problem, product defect identification in particular, needs to deal with multiple categories of features with high dimensionality and possible dependence across feature categories. Some of the standard single classifier methods used in previous studies, e.g., naïve Bayes, may have difficulty in adequately dealing with such a high-dimensional feature space [20]. As multi-view ensemble learning has been shown to be able to take advantage of multiple groups of features and to be a good solution to the problem of high dimensionality [21], we next briefly review the area of multi-view ensemble learning.

2.2. Multi-view ensemble learning

The multi-view ensemble learning approach aims to exploit multiple views of data (i.e., subsets of features) for improved learning performance [22]. The effectiveness of multi-view learning is ensured by two essential principles, namely, consensus and complementarity [23,24]. The aim of the consensus principle is to maximize agreement among the models corresponding to the different views of the data. Complementarity means that each view may contain some knowledge that other views do not.

Multi-view ensemble learning consists of three steps: view creation, base-classifier construction, and base-classifier ensemble. In the first step, multiple subsets of features, referred to as *views*, are selected. Previous methods for view creation include random view creation [25] and performance-based view creation [26]. Random view creation divides the complete feature set into multiple subsets through random

partitioning. The typical approaches that employ random view creation include Random Subspace [27] and Attribute Bagging [26]. There are also variations of these approaches. For example, Tao et al. [28] proposed a method combining random subspace and asymmetric bagging for support vector machines on small-sized, high-dimensional, unbalanced training datasets to alleviate the problems of classifier instability, majority class bias, and overfitting. Performance-based view creation methods employ search algorithms to ensure diversity of features subsets. Sun et al. [29] proposed a genetic algorithm to search for optimal feature subsets. Di and Crawford [26] used clustering to generate feature views. These methods all have shortcomings. Random partitioning methods cannot ensure a satisfactory outcome [23]. Clustering and genetic algorithms incur high computational complexity and have difficulty in predetermining the optimal number and size of feature views. In addition, random partitioning and clustering-based partitioning ignore the advantage of combining views.

After view creation, multiple base classifiers are constructed based on the different views using some classification methods. This step is typically straightforward. The last step of multi-view ensemble learning is to combine the results of the base classifiers through ensemble rules. The most widely-used ensemble rules are predefined ensemble rules, called un-trainable rules. Typical un-trainable rules include Max, Min, Sum, Product, and Majority vote [30,31]. Un-trainable ensemble rules need to be predefined based on experience or analytical properties, and has unstable performance, especially when there lacks priori knowledge. Trainable ensemble rules learned through machine learning are more flexible. Machine learning methods can be used to determine the ensemble manner and parameters based on training data [32].

3. Proposed method for product defect identification

Following several previous studies [2,9–12], we tackle the problem of product defect identification from an online discussion forum focusing on a particular type of products, e.g., vehicles, consumer electronics, appliances, and toys. Specifically, given a discussion thread on the forum, we classify it into either positive (i.e., the thread discusses a defect) or negative (i.e., the thread does not discuss any defect). We propose a novel method for learning a classifier from labeled training data, which can then be used to classify new discussion threads in the future. The effectiveness of such a classifier largely depends on how to represent a discussion thread (i.e., choice of features) and the classification method. Our proposed method has novelties on both main factors. First, we propose the use of contextual features based on replies in the discussion thread. Second, we propose a novel multi-view ensemble learning method specifically tailored to the product defect identification problem.

3.1. Contextual features based on replies

Every discussion thread consists of one original post and any number of replies. A new thread is originated when someone submits a new post. Following the original post, others may submit replies, which are included in the same thread.

Most previous studies on product defect identification (e.g., [2,9–12]) treated a discussion thread as a single document in constructing features, without distinguishing replies from the original post. We posit that it may be useful to disentangle the information contained in the original post and that contained in the replies, since they may overlap but also complement each other. When the original post and the replies are merged into a single document, the differences between the two are lost. Moreover, since different threads may contain very different numbers of replies—some may contain none, while others may contain many—the influence of replies on the final constructed features varies substantially across threads. When a thread contains many replies, another problem is that the distinctive cues about the nature of the thread get diluted due to possible *topic transferring* [33]. We therefore propose to construct features based on original posts and replies separately. We call the features based on replies *contextual features*.

Apparently, original posts and replies are different. The starter of a thread (i.e., the author of the original post) and the repliers tend to play different roles (e.g., consumer vs. consultants) in the system. The original post and replies tend to have different purposes (e.g., question vs. answers). Features based on original posts and those based on replies may have different effects on defect identification.

While replies are different from the original post, they are also correlated with the original post. The original post directly reflects the intention of a thread (e.g., defect, price, and usage). The replies are related to the original post and also reinforce the intention. While features based on original posts are useful for defect identification, contextual features based on replies may provide additional cues to discriminate defect threads from others.

Figs. 1 and 2 show two threads (written in Chinese, annotated with English translations in red color), related to vehicle defect and price, respectively, from an online forum for the automotive industry. In the first thread, it is obvious that the original post inquired about a possible defect the author faced. In the second thread, the starter consulted about the price of a product. In the defect-related thread, the content of the replies is also related to the defect, such as giving suggestions. These replies are longer and use some special words, such as “check” and “4S store”. In addition, the levels (indicating activeness) of the repliers in the social media platform are higher. This may be because people who paid more attention to product quality are more likely to be senior vehicle owners or amateurs who are more active in social media. In the price-related thread, the users have lower levels in the social media platform, probably because they are potential consumers or new vehicle owners. The sentences in the replies are briefer and always contain price information. These two examples show that replies with different topics (e.g., defect and price) exhibit different characteristics. Features derived from replies may therefore be useful in discriminating defect threads from others.

The examples also show the differences between original posts and replies. For example, the original post in the defect-related thread describes the product defect the user faced, while the repliers express their opinions and suggestions. Due to the different motivations of the original post and its replies, the content and characteristics of the replies are different from those of the original post. For example, the replies contain some vehicle components, such as brake pad, that are related to the defect but not contained in the original post. In addition, the replies have different linguistic features from the original post (e.g., the replies are briefer than the original post). Moreover, replies have some unique features that the original post does not have (e.g., Device used in replies).

Separating replies from original posts also helps to alleviate the adverse effect of topic transferring. Topic transferring refers to the phenomenon that the topic of replies may gradually transfer from the topic of the original post to another topic with the changing of time and repliers' interest. Topic transferring diminishes the correlation between the original post and replies. One way to deal with possible topic transferring is to weight the replies according to the sequence of the replies (i.e., later replies get lower weights). A simpler way is to restrict to the first few replies in constructing contextual features.

3.2. A novel multi-view ensemble learning method

As discussed earlier, there are multiple categories of possible features that can be used for product defect identification, e.g., lexical features, stylistic features, social features, sentiment features, distinctive terms, product features, and semantic features [10]. These features form a high-dimensional feature space, and there may be dependence across feature categories. Adding contextual features based on replies, as we propose, further increases the dimensionality and introduces

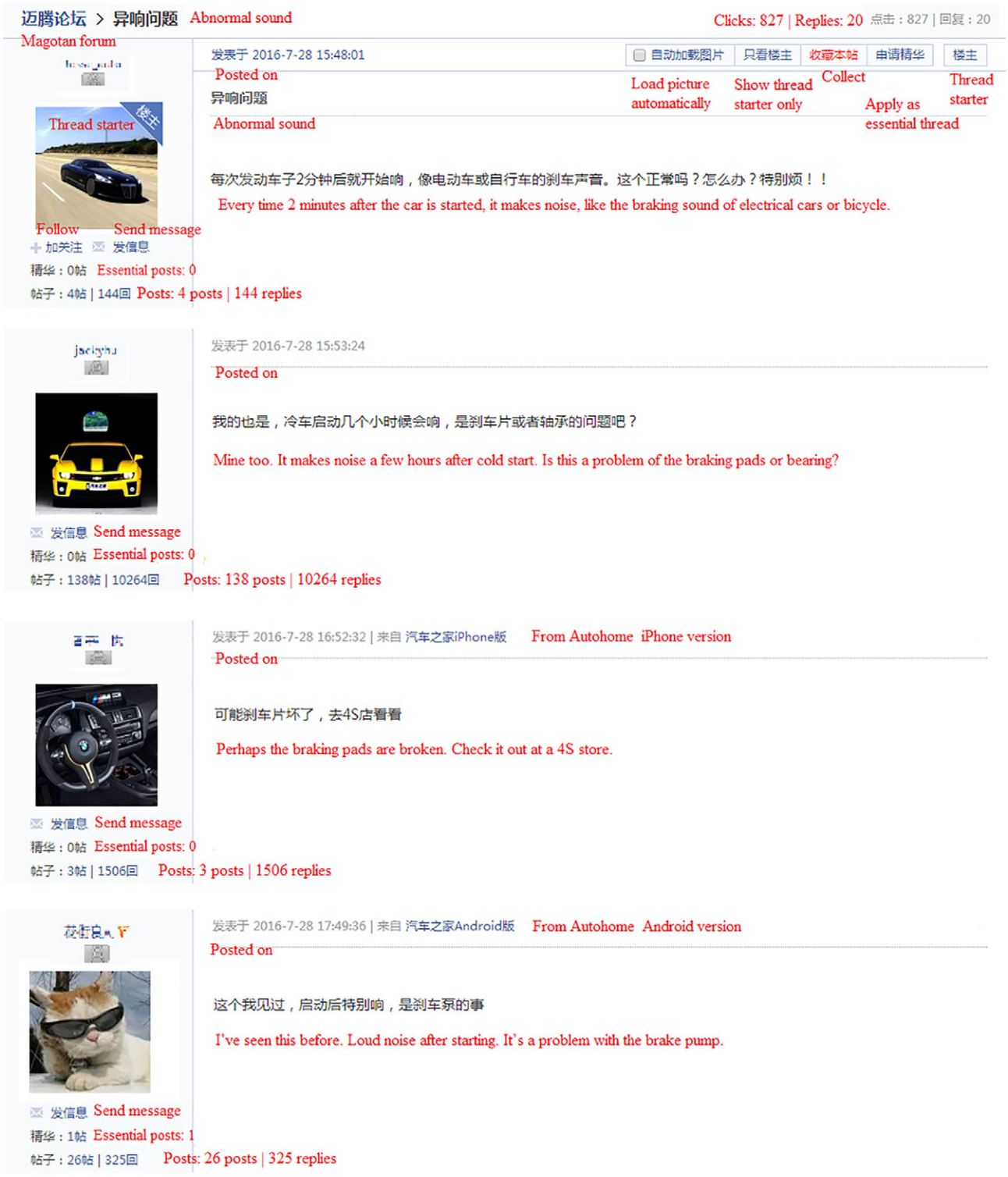

Fig. 1. A thread related to a vehicle defect.

additional dependence between features based on original posts and those based on replies. While the high dimensionality and dependence cause difficulties to many single-view classification methods, such as naïve Bayes, multi-view ensemble learning seems to be suitable for dealing with both problems. By learning multiple base classifiers based on different views (i.e., subsets of features), the dimensionality for each base classifier is reduced, and the dependence across views is avoided.

It seems multi-view ensemble learning is especially suitable for product defect identification incorporating both features based on original posts and contextual features based on replies, as we propose. As discussed earlier, the information contained in original posts and that contained in replies both overlap and complement each other, meeting the essential principals of consensus and complementarity [23,24] desired by multi-view learning. We therefore propose a multi-view ensemble learning method specifically for product defect identification.

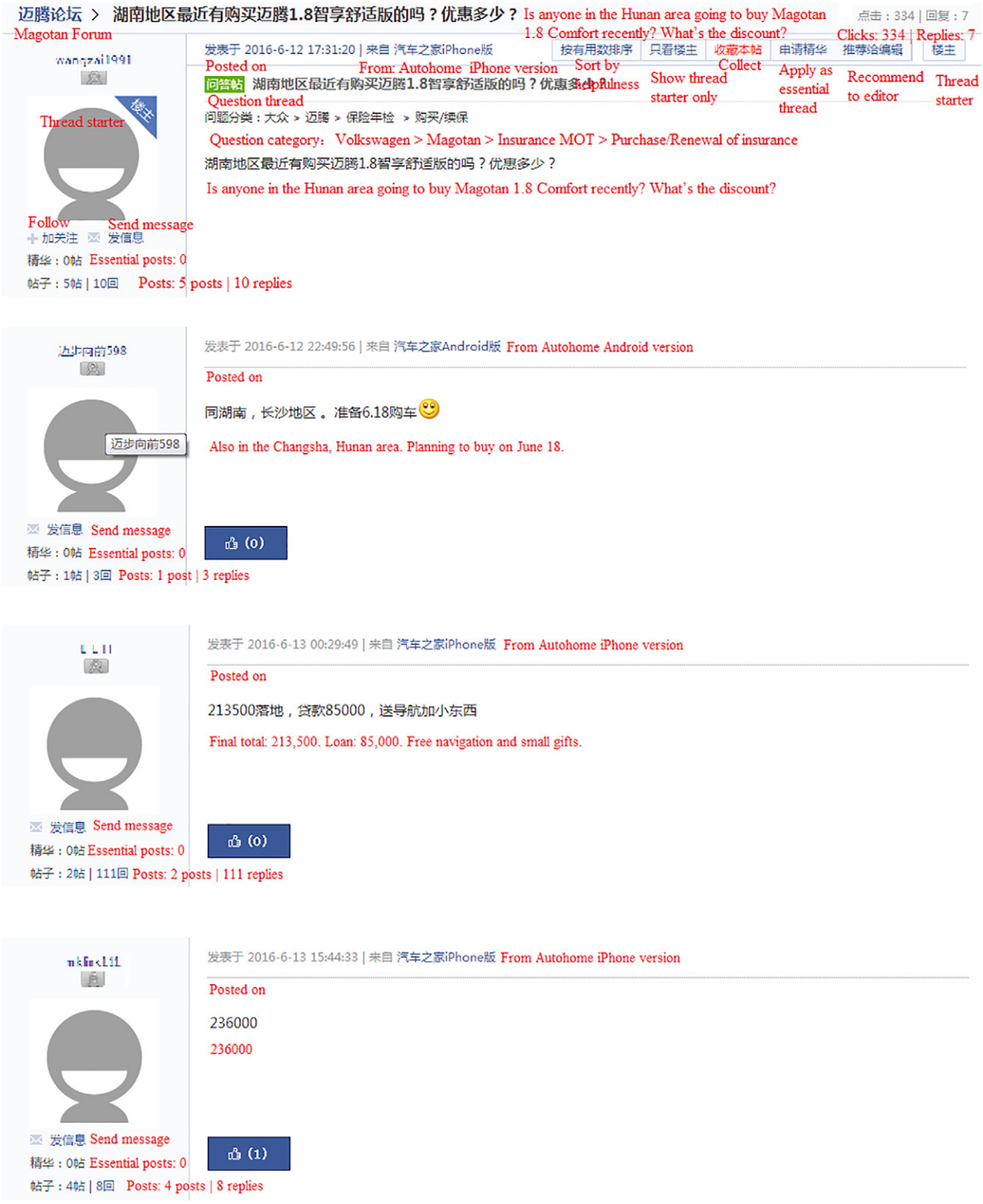

Fig. 2. A thread related to vehicle price.

Fig. 3 outlines the proposed method. The complete feature set is divided into several groups (subsets), each of which is considered an individual view. Another subset of features selected from the complete feature set is used as another view. Each view is used to construct a base classifier using some classification method. Base classifiers 1 to $n-1$ are based the groups of features. Base classifier n is based on the selected features. Finally, a meta classifier is constructed, possibly using a different classification method, to consolidate the predictions of the base classifiers into a final classification result (i.e., the outputs of the base classifiers are the input features of the meta classifier).

View creation and ensemble rule are two key factors of multi-view ensemble learning [34]. Instead of using random view creation methods,

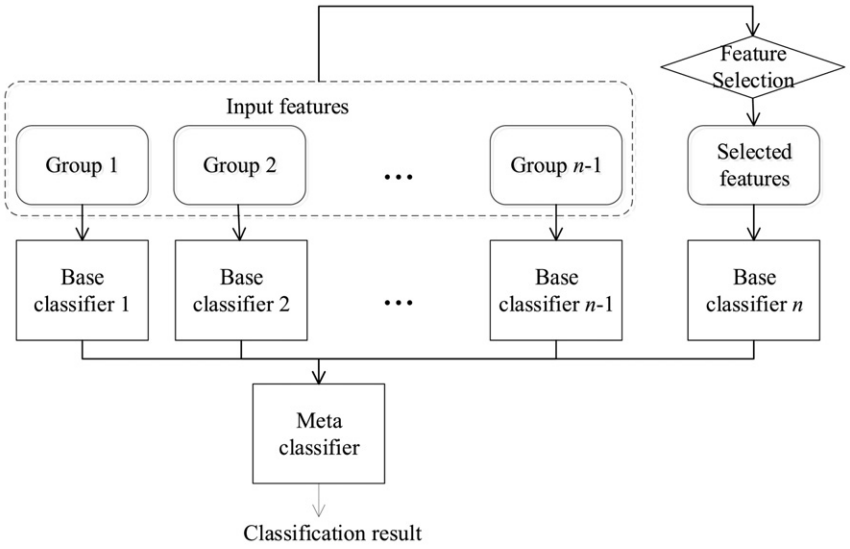

Fig. 3. Proposed multi-view ensemble learning method for product defect identification.

such as Random Subspace [27] and Attribute Bagging [25], or performance-based view creation through genetic algorithm [29] or clustering [26], we see natural divisions of views in the context of product defect identification. Original posts and replies apparently represent different views on discussion threads. In addition, different categories (or combinations of categories) of features naturally form different views. Different categories of features can provide different insights of data for learning [15]. For example, linguistic features extract the information on sentences and words in a post, while social features capture social attributes of the author and the post. Different views consisting of different categories (or combinations of categories) of features ensure the diversity and complementarity of base classifiers, essential for the success of ensemble learning. Such natural division of views also avoids the computational complexity of clustering-based or genetic algorithm-based partitioning.

Besides views based on categories (or combinations of categories) of features, we also select a subset of features from the complete feature set to construct another view. This view can be seen as an overall view combining the individual views. View combination is usually adopted in an attempt to combine the strengths of multiple complementary views, hopefully leading to better performance than any individual view [35–37]. Since the complete feature set has high dimensionality, a feature selection method is needed to reduce the dimensionality. In addition, since there is dependence across categories of features, this feature selection method needs to not only maximize the discriminating power of the selected features (i.e., dependence between the class and the selected features) but also minimize the redundancy in the selected features (i.e., dependence among the selected features). Methods that only consider the dependence between the class and the selected features, without considering the dependence among the selected features, are not appropriate for this purpose.

The ensemble rule is another important factor in multi-view ensemble learning. Un-trainable rules need to be predefined and lack flexibility. In additional, un-trainable rules ignore the relationships among the base classifiers, which are critical for the effectiveness of ensemble learning [32]. We therefore use a trainable ensemble rule. A machine learning method is used to train a meta classifier, which consolidates the outputs of the base classifiers to make the final prediction, similar to the meta classifier in Stacking [38].

4. Case study: defect identification in the automotive industry

We have applied and evaluated our proposed method in a case study in the automotive industry. Vehicle defects have severe consequences and are of great concern to vehicle manufacturers. The USA National Highway Traffic Safety Administration has issued over 90,000 recalls, which have incurred billions of dollars of cost to vehicle manufacturers, dealers, and consumers [9]. In addition to direct cost due to recalls, brand reputation may suffer, and dissatisfied customers are likely to have lower repurchase intention. In each of such recall cases, the cost of the defect is largely proportional to the number of units sold and may be very high. Top-selling cars, such as Volkswagen Magontan, each sells on average in excess of a thousand vehicles per day. Thus, a reduction in defect discovery time by as little as 10 days can keep upwards of 10,000 defective units of that single model off the road, representing a significant saving to the manufacturer.

Identifying product defects is important for product redesign, manufacture redesign, consumer relationship management, and prompt product quality assurance service. For consumers, active product service for defects, once they are identified by the manufacturer, is extremely valuable for smoothly solving safety or performance problems and consequently improving consumers' loyalty and satisfaction [39]. For manufacturers, analyzing more defect feedbacks from consumers is beneficial for product quality improvement, manufacturing technique adjustment, and redesign, hence improving product competitive advantage [7,40]. More importantly, litigation may be initiated by consumers because of product defects, especially if the defects have caused traffic accidents, injuries, or deaths, incurring huge economic losses and brand reputation losses to the manufacturers [9,11]. Identifying product defects with higher accuracy can reduce such risk. Even a small improvement (perhaps as small as 1%) in the performance of product defect identification may be considered practically valuable for the manufacturers.

4.1. Data and pre-processing

We selected autohome.com.cn, a top website for automobile products in China, as our data source. This website contains forums and information of each type of vehicles for user communication. According to official statistics, the average daily number of visitors of autohome.com.cn is >600,000 and the website's average daily number of clicks is >5,000,000.

We crawled 10,000 discussion threads in July 2016 from the Magontan forum, which is one of the most active forums at autohome.com.cn. All of the postings in the threads are written in Chinese. The percentage of threads that contain replies is 95.1%. Thus, contextual features based on replies can be extracted and used. The mean number of

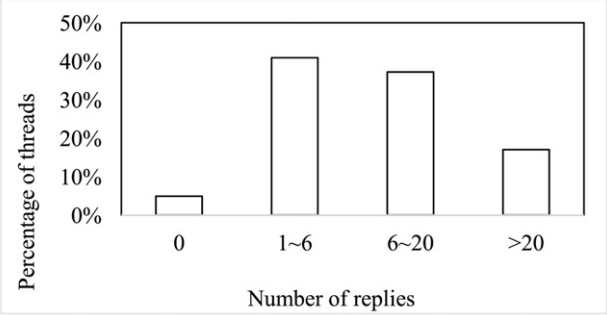

Fig. 4. Distribution of the number of replies in discussion threads.

replies of each thread is 16.56 (standard deviation is 51.98, and the maximum is 1282). The distribution of the number of replies is shown in Fig. 4.

We employed three vehicle engineering master students to tag each of the discussion threads into positive (i.e., a vehicle defect is discussed in the thread) or negative. The final tagging results were determined via majority voting. The kappa coefficients between the tag results of the three taggers are 0.92, 0.949, and 0.875, respectively, indicating satisfactory inter-rater reliability.

The percentage of defect (i.e., positive) reviews is 12.7%. The dataset is a typical imbalanced dataset. The challenge posed by an imbalanced dataset is that standard classification learning algorithms are often biased toward the majority class, leading to a higher misclassification rate for the minority class [41,42]. In order to deal with the problem, we need to re-construct the sample. In the pre-processing stage, we balanced the classes through assigning different weights to the instances in the two classes.

Pre-processing also includes removing duplicates, URL, hash tags, stop words, and special characters (such as emoticons). In addition, we performed word segmentation (since the posts are written in Chinese) and part-of-speech tagging. In English, space is a natural separator of words. Different from English, Chinese does not use any separator of words, each of which may consist of one or more Chinese characters, thus causing difficulty for computers to analyze Chinese text. Chinese word segmentation is to separate a Chinese sentence into a sequence of words, which are the basic meaningful unit for processing. This technique provides the foundation for Chinese text analysis, such as text classification, information retrieval, and machine translation. Take the following sentence as example: "我的车出现了抖动 (My car has jittered)". After word segmentation, the sentence is separated into a sequence of four words of different lengths: "我的\车\出现了\抖动". We used ICTCLAS, a Chinese lexical analysis system, to perform Chinese word segmentation.

4.2. Feature extraction

The features extracted consist of four categories: linguistic features, social features, distinctive terms, and contextual features. The first three categories are based on the original post of each thread. The contextual features are based on the replies following the original post. To alleviate the effect of topic transferring, we selected up to the first six replies for each thread in constructing the contextual features. The specific features extracted in the case study are summarized in Table 1. Note that our linguistic features category also includes the so-called "stylistic features" of Abrahams et al. [10], and our distinctive terms category also includes the so-called "lexical features" of Abrahams et al. [10]. We did not use product features as described by Abrahams et al. [10] because the postings at the forum are not tagged with product characteristics. We did not use semantic features as described by Abrahams et al. [10] because we were not able to find a comprehensive semantic dictionary in Chinese mapping words to semantic categories.

For linguistic features, we used the number of characters, the number of sentences, the number of exclamatory sentences (i.e., sentences ending with the exclamation mark), and the number of interrogative sentences (i.e., sentences ending with the question mark). A longer post contains more useful information so that the consumer can point out a product defect more clearly. In addition, defect posts tend to use some special types of sentences and vocabulary [13]. Inspired by Krishnamoorthy et al. [14], we also used the numbers of adjectives, verbs, adverbs, nouns, modal particles, and mimetic words as linguistic features.

Table 1
The features extracted in the case study.

Category	Feature	Type	Summary statistics
Linguistic features	Number of characters	Numerical	Mean: 77.89, Stddev: 237.26, Max: 10,630, Min: 5
	Number of sentences	Numerical	Mean: 3.05, Stddev: 6.85, Max: 349, Min: 1
	Number of exclamatory sentences	Numerical	Mean: 0.35, Stddev: 1.39, Max: 45, Min: 0
	Number of interrogative sentences	Numerical	Mean: 0.53, Stddev: 1.01, Max: 20, Min: 0
	Number of adjectives	Numerical	Mean: 1.3, Stddev: 1.39 Max: 12, Min: 0
	Number of verbs	Numerical	Mean: 6.75, Stddev: 4.51, Max: 23, Min: 0
	Number of adverbs	Numerical	Mean: 1.73, Stddev: 1.86, Max:12, Min: 0
	Number of nouns	Numerical	Mean: 4.89, Stddev: 3.58, Max: 21, Min: 0
	Number of modal particles	Numerical	Mean: 0.55, Stddev: 0.78, Max: 8 Min: 0
	Number of mimetic words	Numerical	Mean: 0.09, Stddev: 0.39, Max: 8, Min: 0
Social features	Number of views	Numerical	Mean: 2744.2, Stddev: 45,689.67, Max: 4,201,465, Min: 1
	Whether the post contains pictures	Binary	Yes: 35.41%
	Level of the author	Numerical	Mean: 3.57, Stddev: 2.23, Max: 20 Min: 1
	Number of original posts by the author	Numerical	Mean: 25.11, Stddev: 56.66, Max: 1802, Min: 1
	Number of replies by the author	Numerical	Mean: 330.91, Stddev: 1273.8, Max: 45,758, Min: 0
Distinctive terms	Number of smoke words	Numerical	Mean: 1.16, Stddev: 1.28, Max: 12, Min: 0
	Presence of key words	Binary	34 words selected by CFS
Contextual features	Number of replies	Numerical	Mean: 18.38, Stddev: 122.68, Max: 8901, Min: 0
	Number of sentences in replies	Numerical	Mean: 8.97, Stddev: 6.93, Max: 169, Min: 0
	Average number of characters per sentence in replies	Numerical	Mean: 18.38, Stddev: 11.58, Max: 144, Min: 0
	Average level of repliers	Numerical	Mean: 5.01, Stddev: 716.76, Max: 23, Min: 0
	Sentiment of replies	Numerical	Mean: 0.28, Stddev: 0.341, Max: 1, Min: −1
	Device used in replies (Computer or mobile phone)	Binary	Mobile phone: 71.09%
	Number of smoke words in replies	Numerical	Mean: 2.84, Stddev: 3.84, Max: 27, Min: 0
	Presence of key words in replies	Binary	25 words selected by CFS

Social features quantify the social attributes of a post and the author. An active author with better social features is more likely to post more high-quality posts [15]. Reviews with better social features should be related with product quality, because reviews regarding product quality are easier to attract other reviewers' attention [43]. For social features, we used the number of views, the level of the author (a measure of activeness assigned by the website), the numbers of original posts and replies by the author, and whether the post contains pictures.

In the distinctive terms category, we used both smoke words, which appear more prevalently in defect threads, and key words selected using a feature selection method. Both types of words are based on the result of word segmentation using ICTCLAS during data preprocessing. Selected key words are widely used in text classification problems, as different classes of texts tend to have heterogeneous word distributions. Previous studies (e.g., [2,9–12]) have revealed the usefulness of smoke words in product defect identification.

For contextual features based on replies, we selected some typical features to represent linguistic features, social features, sentiment, device used, smoke words, and key words. Note that we used sentiment on replies but not on original posts. Sentiment of entire discussion threads has been shown to be ineffective in product defect discovery [10]. However, the effectiveness of sentiment of replies only, as a contextual feature, is unknown yet. We computed the difference between the number of positive words and negative words, based on the Hownet sentiment dictionary, as the sentiment of replies.

We used a process similar to that of Abrahams et al. [9] in identifying smoke words. We computed the information gain of each word to measure the difference of its distributions in the two types (defect and non-defect) of threads, and selected words that appear frequently in defect threads but barely in non-defect threads manually. The final smoke word list consists of 55 words.

We selected key words for original posts and for replies separately. The number of candidate key words reaches quantities of thousands, and most of them are not effective for the thread classification. Thus, it is necessary to select an effective feature set that has better classification ability. We chose the Correlation-based Feature Selection (CFS) method [18], which has been shown to be an effective feature selection method, as it considers not only the correlation between features and the class but also the correlation among features. The core of CFS is a heuristic for evaluating the effectiveness or merit of a feature subset. The heuristic is that a good feature subset should contain features that are highly correlated with the class while being uncorrelated with each other. Specifically, $Merit_s = \frac{k * \overline{R_{cf}}}{\sqrt{k + k*(k-1)* \overline{R_{ff}}}}$, where $Merit_S$ is the heuristic "*Merit*" of a feature set S containing k features, $\overline{R_{cf}}$ is the average Pearson's correlation between the features in S and the class, and $\overline{R_{ff}}$ is the average Pearson's correlation among the features in S. The numerator can be thought of as an indication of how predictive the feature subset is; the denominator reflects the redundancy among the features in the subset. To reduce the redundancy of the feature subset, some features will be removed if they are highly correlated with other features in the subset. Best-first search is used to search for the best feature subset based on the "Merit" of the feature set. Best-first search starts with the empty set and checks all possible single-feature expansions. The subset with the highest merit is chosen and is then expanded by again adding a single feature. If no expanded subset has higher *Merit*, the best-first search algorithm retreats to the next-best unexpanded feature subset and continues from there. The best-first search algorithm searches the candidate feature subset space in this manner and returns the best subset found when the search terminates. The final key word list for origin posts consists of 34 words, while that for replies consists of 25 words. These two key word lists share 13 common words.

Table 2 shows the key words for original posts and replies. They overlap but also complement each other. Just as mentioned before, the thread starter and repliers have different roles in a discussion thread, where the repliers express their opinions and suggestions to the original post. Consequently, they tend to exhibit different content characteristics, as reflected in their key words. Contextual features contain additional information beyond the features based on original posts. Such complementarity may be beneficial to enhance classification performance in product defect identification.

Table 2
Key words of original posts and replies.

	Replies	Original posts
Exclusive	Refit (改装); Check (检查); Normal (正常); Paint (漆); Claim for Compensation (索赔); Solve (解决); Common problem (通病); Key (钥匙); Money (钱); Situation (情况); Previous (以前); Bluetooth ("蓝牙"的"牙")	Engine (发动机); Reason (原因); Navigation (导航); Drive (开); Idling (怠); Wheel (毂); Gas (油); Obvious (明显); When (时候); Is (是); Problem (问题); Wrong ("怎么回事"的"事"); Sound (响声); Sound (声); Time (时); Has (有); Gear (挡); Seep (渗); Kilometer (公里); Speed (速); Subject (题)
Overlapping	Appear (出现); Noise (噪音); Abnormal (异); Sound (声音); Jitter (抖动); Jerking (顿); Malfunction (故障); Leak (漏); Burn (烧); Find (发现); Discount (优惠); Ring (响); Shake (抖)	

4.3. Performance metrics and estimation method

We used F-measure, Accuracy (ACC), and Matthews Correlation Coefficient (MCC) in assessing the performance of our proposed method, as well as other baseline methods. These metrics are commonly used to measure the performance of classification methods. Given a classification model and a test dataset, the performance of the model on the test dataset can be measured based on the confusion matrix shown in Table 3.

The ACC is simply the proportion of correctly classified examples, i.e., $ACC = (TP + TN)/(TP + TN + FP + FN)$.

The F-measure is the harmonic mean of precision and recall, i.e., $F-measure = 2^* P^* R/(P + R)$, where precision $P = TP/(TP + FP)$ and recall $R = TP/(TP + FN)$.

MCC attempts to measure the accuracy on each of the two classes with good balance between the two classes. $MCC = \frac{TP^* TN - FP^* FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

We repeated 10-fold cross validation 10 times to estimate the performance of a classification method. During each 10-fold cross validation, the entire dataset is randomly split into 10 roughly equal-sized subsets (folds). Each fold is kept for testing the model trained on the other nine folds. This is repeated for all 10 folds in each 10-fold cross validation. The entire 10-fold cross validation is repeated 10 times, with different random splitting of folds each time. The performance results reported later are the averages over the $10 * 10$ estimates.

4.4. Experiments

We conducted a series of experiments using Weka (http://www.cs.waikato.ac.nz/ml/weka/) extended with our proposed method. First, we compared several standard classification methods and chose the top performer as the base classification method in our proposed multi-view ensemble learning method. We then evaluated the effectiveness of the two major novelties in our proposed method: the use

Table 3
Confusion matrix.

	Predicted as positive	Predicted as negative
Actually positive	True positives (TP)	False negatives (FN)
Actually negative	False positives (FP)	True negatives (TN)

Table 4
Performance of standard classification methods.

Classification method	F-measure (Stddev)	ACC (Stddev)	MCC (Stddev)
Naïve Bayes	**0.839 (0.0006)**	**0.825 (0.0005)**	**0.661 (0.0012)**
Random Forest	0.755 (0.0090)	0.794 (0.0060)	0.623 (0.0103)
Logistic Regression	0.737 (0.0052)	0.781 (0.0036)	0.596 (0.0058)
Support Vector Machine	0.384 (0.0024)	0.618 (0.0014)	0.365 (0.0023)
k-Nearest-Neighbors	0.371 (0.0054)	0.604 (0.0019)	0.308 (0.0047)

The bold means that Naive Bayes gets best performance.

of contextual features based on replies and the use of multi-view ensemble learning for combining base classifiers based on different categories of features. We compared our proposed multi-view ensemble learning method with several other ensemble methods, such as Random Subspace, Bagging, and Boosting, using all features. We evaluated different categories of features and different combinations of features to identify the effectiveness of adding contextual features. We also compared our proposed method with one that derives features from the entire threads (i.e., merging original posts and their replies), as has been done in previous studies [2,9–12]. For all experiment settings, 10 runs of 10-fold cross validation were used to estimate the classification performance.

4.4.1. Comparison of standard classification methods

First, we tested five widely used classification methods: Naïve Bayes, Logistic Regression, Support Vector Machine, Random Forest, and k-Nearest-Neighbors. We used all the features (Table 1) in this experiment. We retained the default parameter values of Weka for all the methods. Table 4 shows the average and standard deviation of the performance of each method. Naïve Bayes outperformed all other methods in terms of all three performance metrics; *t*-test shows that the difference between Naïve Bayes and every other method in terms of every performance metric is statistically significant ($p < 0.01$). We therefore selected Naïve Bayes as the base classification method for our proposed multi-view ensemble learning method and other baseline ensemble methods in the subsequent experiments.

4.4.2. Effectiveness of our proposed multi-view ensemble learning method

We evaluated the effectiveness of our proposed multi-view ensemble learning method, in comparison with other ensemble methods. We used four ensemble methods—Random Subspace, Bagging, Adaboost M1, and Logitboost—as baselines. We used Naïve Bayes as the base classification method for all the ensemble methods. We retained the default parameter values of Weka for all the baseline methods. We used all the features (Table 1) in this experiment. In Random Subspace, each base classifier is based on a random subset of the features. In Bagging and Boosting (Adaboost M1 and Logitboost), all base classifiers are based on all the features.

In our proposed multi-view ensemble learning method, each base classifier is based on one of the four categories of features, which are considered different views of the data. In addition, another base classifier is trained based on a subset of all the features selected using CFS. This subset is considered an overall view of the data. We chose CFS because it considers not only the correlation between the selected features and the class but also the correlation among the selected features. Finally, the five base classifiers are combined using a logistic regression model. We chose logistic regression for the meta classifier because it is suggested, in the context of Stacking, that "relatively global, smooth" classifiers, e.g., logistic regression, are better choices for meta classifiers [38]. Fig. 5 illustrates the structure of our proposed method, as instantiated in this case study.

Table 5 shows the average and standard deviation of the performance of each ensemble method. Our proposed multi-view ensemble learning method outperformed all other ensemble methods in terms of all three performance metrics; *t*-test shows that the difference

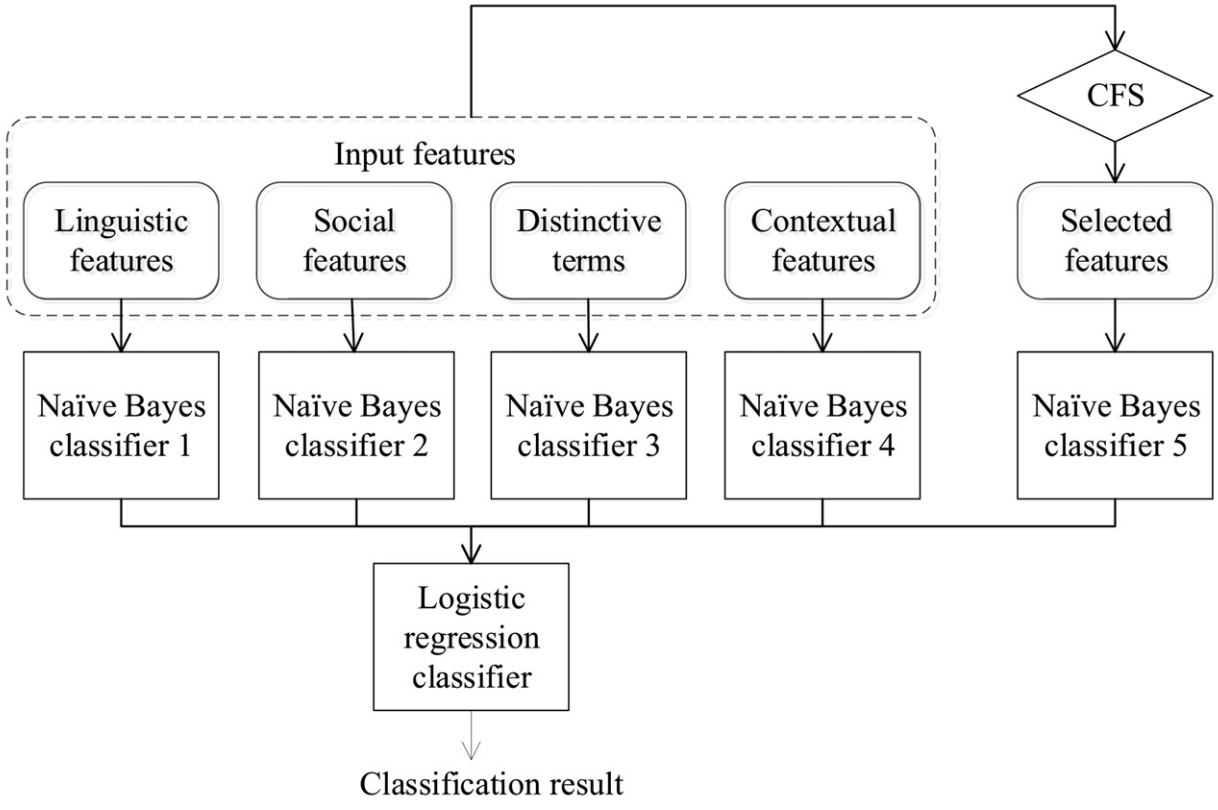

Fig. 5. Structure of proposed method as instantiated in the case study.

Table 5
Performance of different ensemble methods.

Method	F-measure (Stddev)	ACC (Stddev)	MCC (Stddev)
Proposed method	**0.893 (0.0016)**	**0.896 (0.0013)**	**0.793 (0.0026)**
Random Subspace	0.842 (0.0022)	0.828 (0.0022)	0.658 (0.0047)
Adaboost M1	0.865 (0.0015)	0.858 (0.0015)	0.720 (0.0029)
Bagging	0.839 (0.0022)	0.826 (0.0022)	0.660 (0.0046)
Logitboost	0.879 (0.0026)	0.886 (0.0021)	0.780 (0.0041)

The bold means that proposed method gets the best performance.

between our proposed method and every other ensemble method in terms of every performance metric is statistically significant ($p < 0.01$).

4.4.3. Effectiveness of contextual features

In another experiment, we examined the effectiveness of the contextual features based on replies. We tested the performance of different categories of features and several combinations of feature categories, using Naïve Bayes or our proposed multi-view ensemble learning method. Table 6 summarizes the results.

Among the four categories of features, distinctive terms led to the best performance, showing the importance of distinctive terms in defect identification. This finding is consistent with previous studies [10–12]. Contextual features based on replies also gave close performance, much higher than that based on linguistic features or social features, showing that information contained in replies is also valuable in defect identification.

Simply combining multiple categories of features using Naïve Bayes did not lead to improved performance over that given by distinctive terms or contextual features alone. One possible reason is that combining multiple categories of features increases the dimensionality of the feature space, and thus the likelihood of overfitting. Another possible reason is that there is dependence across the different categories of features, violating the conditional independence assumption of Naïve Bayes. The dependence between distinctive terms and contextual features is more obvious. The same smoke word list was used on the original posts and their replies. The key word lists selected by CFS for the original posts and the replies share many common words.

However, using our proposed multi-view ensemble learning method, combining multiple categories of features led to substantial performance improvement. Combining the three categories of features based on original posts improved the performance of any single category of features; the performance improvement in terms of every performance metric is statistically significant ($p < 0.01$). Adding our proposed contextual features further improved the performance; the performance improvement in terms of every performance metric is again statistically significant ($p < 0.01$). In our proposed method, each base classifier is based on one category of features, avoiding the increase in dimensionality of the feature space and the adverse effect of the dependence across different categories of features.

The results show that our proposed contextual features based on replies are indeed valuable and can contribute to performance improvement over other features based on original posts. However, the effectiveness of adding contextual features also depends on the way they are added. Simply combining them with other features in a Naïve Bayes classifier did not improve performance, but adding them into our proposed multi-view ensemble learning model substantially improved performance.

4.4.4. Comparison with previous studies

Some previous studies used standard single classifier methods in product defect identification [2,10–12]. They also used the information contained in replies, but in a different manner. They simply merged the replies with the original post of each discussion thread to derive various categories of features (i.e., there was no distinction between the original post and the following replies), instead of adding additional contextual features based on just replies separately. To compare our proposed method with these previous studies, we tested the performance of either merging the replies with the original post or adding our proposed contextual features, using either a single Naïve Bayes classifier or our proposed multi-view ensemble learning method.

The results (Table 7) show that our proposed multi-view ensemble learning method substantially outperformed the single Naïve Bayes classifier method in terms of every performance metric, no matter how replies were used (either merged with original post or used to derive additional contextual features). Adding additional contextual features also outperformed merging replies with original post, using either Naïve Bayes or our proposed multi-view ensemble learning method.

Overall, our proposed method substantially outperformed the method used in previous studies (i.e., using a standard single classifier method and merging replies with original post); the difference in terms of every performance metric is statistically significant based on *t*-test ($p < 0.01$). The performance improvement was over 5% (0.840 to 0.893, 0.825 to 0.896, and 0.661 to 0.793 in terms of F-measure, ACC, and MCC, respectively), and considered practically valuable, given the potentially dire consequences of vehicle defects and the large volume of discussion threads on the forum. Both novelties (adding contextual features and using our proposed multi-view ensemble learning method) contributed to the performance improvement.

4.5. Defect discussion thread analysis

After identifying defect discussion threads from the online forum using our proposed method, the manufacturer may further analyze such threads using various text analysis methods. For example, topic analysis can be used to extract topic words from the threads. These can help the manufacturer quickly get a rough sense of what defects are mentioned on the forum.

Table 6
Results of different feature combinations.

Classification method	Feature set	F-measure (Stddev)	ACC (Stddev)	MCC (Stddev)
Naïve Bayes	Linguistic features (a)	0.655 (0.0025)	0.671 (0.0019)	0.343 (0.0037)
	Social features (b)	0.663 (0.0023)	0.623 (0.0029)	0.350 (0.0081)
	Distinctive terms (c)	0.850 (0.0007)	0.850 (0.0006)	0.701 (0.0011)
	Contextual features (d)	0.834 (0.0004)	0.840 (0.0003)	0.690 (0.0007)
	(a) + (d)	0.845 (0.0011)	0.853 (0.0009)	0.708 (0.0018)
	(b) + (d)	0.836 (0.0004)	0.845 (0.0003)	0.693 (0.0007)
	(c) + (d)	0.845 (0.0006)	0.845 (0.0007)	0.691 (0.0011)
	(a) + (b) + (c)	0.834 (0.0009)	0.820 (0.0009)	0.650 (0.0019)
	All features	0.853 (0.0007)	0.844 (0.0008)	0.693 (0.0014)
Proposed method	(a) + (b) + (c)	0.882 (0.0009)	0.883 (0.0008)	0.767 (0.0015)
	All features	**0.893 (0.0016)**	**0.896 (0.0013)**	**0.793 (0.0026)**

The bold means that our method gets the best performance.

Table 7
Results of different uses of replies.

Classification method	Use of replies	F-measure (Stddev)	ACC (Stddev)	MCC (Stddev)
Naïve Bayes	Merged with original post (corresponding to previous studies)	0.840 (0.0008)	0.825 (0.0006)	0.661 (0.0014)
	Contextual features	0.853 (0.0007)	0.844 (0.0008)	0.693 (0.0014)
Multi-view ensemble learning	Merged with original post	0.882 (0.0015)	0.880 (0.0012)	0.761 (0.0024)
	Contextual features (i.e., proposed method)	**0.893 (0.0016)**	**0.896 (0.0013)**	**0.793 (0.0026)**

The bold means that proposed method gets the best performance with contextual features.

To illustrate the use of such text analysis methods, we analyzed the threads in our dataset that were identified as defect threads by our proposed method. We used Latent Dirichlet Allocation (LDA) topic analysis model to analyze the topics of the threads. For the purpose of contrasting, we also performed the same analysis on non-defect threads and on all the threads.

The top topics mentioned in non-defect threads focus on "buy", "discount", "maintenance", "navigation", "price", "lube", "VW", "engine", "market", and so on. These topics involve different aspects in automobile purchase and usage. However, the top topics mentioned in defect threads are substantially different. They mainly involve specific symptoms of automotive defects, such as "sound", "jerking", "jitter", and "alarm", and pertinent parts, such as "engine", "gearbox", "brake", "accelerator", "door", and "gas". In addition, the topic words "refit", "check", and "maintenance" indicate that vehicle users may often check the defective parts and repair them.

The top topics for all threads are very similar to those for non-defect threads. This is expected as majority of the threads are non-defect. Topic words indicating defects are buried in the numerous other popular topic words. It is difficult to directly identify discussions of defects from all threads. It is therefore important to narrow down to a small portion of defect threads first. The topic analysis results generated based on our thread classification results confirm that our method can effectively identify defect threads from all threads, which mainly consist of non-defect threads.

5. Conclusion and future work

Product defect identification from social media, especially online discussion forums, has drawn considerable research attention. However, existing product defect identification methods have not fully exploited the information contained in replies and have not adequately addressed the high dimensionality of feature space and dependence across feature categories. In this study, we have proposed a novel method, aiming to bridge these two gaps specifically. The method incorporates contextual features based on replies to better exploit the utility of replies in reinforcing and complementing the original posts. A multi-view ensemble learning method is proposed specifically for the focal problem to deal with the high dimensionality and dependence problems. Results from a case study in the automotive industry show that our method improved defect identification performance, as compared to existing methods, and both novelties in our method contributed to the performance improvement. Although the method is proposed for product defect identification, it is quite general and can be adapted and applied to a broad range of social media text classification problems, such as online review helpfulness or usefulness prediction [4,7,13–15].

The findings of this study have implications for both practitioners and researchers. For practitioners, our case study shows that automated defect identification models can achieve satisfactory performance (our proposed method achieved an accuracy close to 90%). Follow-up analyses show that defect discussion threads identified by our method can indeed reveal mentions of defects, which would otherwise be buried in the vast volume of irrelevant threads. Thus, it seems promising for manufacturers to actually implement and use our method in practice, substantially saving manual labor in sifting through the voluminous data on their online forums. Besides cost saving, an automated tool also shortens the time needed to identify defects, allowing prompt remedial actions.

For research, our work contributes a new way to effectively explore the complementarity of different components associated with a social media object in social media analytics. Our study shows the effectiveness of contextual features derived from replies in a discussion thread and multi-view ensemble learning in product defect identification. Similar ideas may be explored and tested in other contexts, such as review helpfulness prediction [4,7,13–15], sentiment analysis [6,44], and opinion leader identification [3].

As part of future work on this topic, several interesting extensions of this work can be explored. First, more comprehensive evaluations in multiple industries may be conducted to test the generalizability of our findings. Second, novel features, such as social networking features and semantic features, may be constructed and tested. Third, the current version of our method requires a large amount of manually tagged training data, which can incur difficulties in practical applications. Semi-supervised learning techniques may be used to reduce the manual labor required and take advantage of the high volume of social media data. Fourth, our current method runs in a static, batch manner. Social media data have high velocity and the characteristics of product defect discussions are constantly changing. Frequently retraining models in a batch manner to catch such changes promptly is infeasible. Future research may extend our method with dynamic and incremental learning.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 71731005 and Grant No. 71571059), Fund for Humanities and Social Science Fund Research Planning of the Ministry of Education (Grant No. 15YJA630010) and the key grant of educational commission of the Anhui province (No. KJ2016A525).

References

[1] Y.M. Li, H.M. Chen, J.H. Liou, L.F. Lin, Creating social intelligence for product portfolio design, Decis. Support. Syst. 66 (2014) 123–134.
[2] A.S. Abrahams, J. Jiao, W. Fan, G.A. Wang, Z. Zhang, What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings, Decis. Support. Syst. 55 (2013) 871–882.
[3] S. Liu, C. Jiang, Z. Lin, Y. Ding, R. Duan, Z. Xu, Identifying effective influencers based on trust for electronic word-of-mouth marketing, Inf. Sci. 306 (2015) 34–52.
[4] S. Lee, J.Y. Choeh, Predicting the helpfulness of online reviews using multilayer perceptron neural networks, Expert Syst. Appl. 41 (2014) 3041–3046.
[5] C.Q. Jiang, K. Liang, H. Chen, Y. Ding, Analyzing market performance via social media: a case study of a banking industry crisis, SCIENCE CHINA Inf. Sci. 57 (2014) 1–18.
[6] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
[7] Y. Liu, J. Jin, P. Ji, J.A. Harding, R.Y.K. Fung, Identifying helpful online reviews: a product designer's perspective, Comput. Aided Des. 45 (2013) 180–194.

[8] X. Zhang, Z. Qiao, L. Tang, W. Fan, E. Fox, G. Wang, Identifying product defects from user complaints: a probabilistic defect model, Proceedings of the Americas Conference on Information Systems, 2016.
[9] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media, Decis. Support. Syst. 54 (2012) 87–97.
[10] A.S. Abrahams, W. Fan, G.A. Wang, Z.J. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery, Prod. Oper. Manag. 24 (2015) 975–990.
[11] D. Law, R. Gruss, A.S. Abrahams, Automated defect discovery for dishwasher appliances from online consumer reviews, Expert Syst. Appl. 67 (2017) 84–94.
[12] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews, Decis. Support. Syst. 90 (2016) 23–32.
[13] H. Almagrabi, A. Malibari, J. Mcnaught, A survey of quality prediction of product reviews, Int. J. Adv. Comput. Sci. Appl. 6 (2015) 49–58.
[14] S. Krishnamoorthy, Linguistic features for review helpfulness prediction, Expert Syst. Appl. 42 (2015) 3751–3759.
[15] X. Zheng, S. Zhu, Z. Lin, Capturing the essence of word-of-mouth for social commerce: assessing the quality of online e-commerce reviews by a semi-supervised approach, Decis. Support. Syst. 56 (2013) 211–222.
[16] A. Fernández, V. López, M. Galar, M.J. Del Jesus, F. Herrera, Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches, Knowl.-Based Syst. 42 (2013) 97–110.
[17] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (2014) 16–28.
[18] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, Proceedings of the Seventeenth International Conference on Machine Learning 2000, pp. 359–366.
[19] F. Figueiredo, H. Pinto, F. Belém, J. Almeida, M. GonçAlves, D. Fernandes, E. Moura, Assessing the quality of textual features in social media, Inf. Process. Manag. 49 (2013) 222–247.
[20] H. Elghazel, A. Aussem, Unsupervised feature selection with ensemble learning, Mach. Learn. 98 (2015) 157–180.
[21] G. Chao, S. Sun, Multi-kernel maximum entropy discrimination for multi-view learning, Intell. Data Anal. 20 (2016) 481–493.
[22] S. Zhu, X. Sun, D. Jin, Multi-view semi-supervised learning for image classification, Neurocomputing 208 (2016) 136–142.
[23] V. Kumar, S. Minz, Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification, Knowl. Inf. Syst. 49 (2016) 1–59.
[24] V. Kumar, S. Minz, Multi-view ensemble learning for poem data classification using SentiWordNet, Advanced Computing, Networking and Informatics-Volume 1, Springer 2014, pp. 57–66.
[25] R. Bryll, R. Gutierrez-Osuna, F. Quek, Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recogn. 36 (2003) 1291–1302.
[26] W. Di, M.M. Crawford, View generation for multiview maximum disagreement based active learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 50 (2012) 1942–1954.
[27] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844.
[28] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1088–1099.
[29] S. Sun, F. Jin, W. Tu, View construction for multi-view semi-supervised learning, Advances in Neural Networks-ISNN 2011 2011, pp. 595–601.
[30] Y. Li, H. Guo, X. Liu, Y. Li, J. Li, Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data, Knowl.-Based Syst. 94 (2016) 88–104.
[31] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, Y. Zhou, A novel ensemble method for classifying imbalanced data, Pattern Recogn. 48 (2015) 1623–1637.
[32] M. Woniak, M. Graña, E. Corchado, A survey of multiple classifier systems as hybrid systems, Inform. Fusion 16 (2014) 3–17.
[33] Q. Zhou, G. Wang, H. Chen, A topic evolution model based on microblog network, Lect. Notes Electr. Eng. 260 (2014) 791–798.
[34] Z. Wang, S. Chen, D. Gao, A novel multi-view learning developed from single-view patterns, Pattern Recogn. 44 (2011) 2395–2413.
[35] M. Liu, L. Zhang, H. Hu, L. Nie, J. Dai, A classification model for semantic entailment recognition with feature combination, Neurocomputing 208 (2016) 127–135.
[36] J. Hou, M. Pelillo, A simple feature combination method based on dominant sets, Pattern Recogn. 46 (2013) 3129–3139.
[37] M. Swain, S. Sahoo, A. Routray, P. Kabisatpathy, J.N. Kundu, Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition, Int. J. Speech Technol. 18 (2015) 387–393.
[38] D.H. Wolpert, Stacked generalization, Neural Netw. 5 (1992) 241–259.
[39] W. Reinartz, M. Krafft, W.D. Hoyer, The customer relationship management process: its measurement and impact on performance, J. Mark. Res. 41 (2004) 293–305.
[40] W.D. Hoyer, R. Chandy, M. Dorotic, M. Krafft, S.S. Singh, Consumer cocreation in new product development, J. Serv. Res. 13 (2010) 283–296.
[41] J.F. Díez-Pastor, J.J. Rodríguez, C. García-Osorio, L.I. Kuncheva, Random balance: ensembles of variable priors classifiers for imbalanced data, Knowl.-Based Syst. 85 (2015) 96–111.
[42] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.
[43] O. Emir, Customer complaints and complaint behaviours in Turkish hotel restaurants: an application in Lara and Kundu areas of Antalya, Afr. J. Bus. Manag. 5 (2011) 4239–4253.
[44] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowl.-Based Syst. 89 (2015) 14–46.

Yao Liu is a doctoral student at the School of Management, Hefei University of Technology, China. His research interests include online reviews analysis, product quality management, and machine learning.

Cuiqing Jiang is a Professor at the School of Management, Hefei University of Technology, China. He received his PhD degree in 2007 from Hefei University of Technology. His research interests include knowledge management, business intelligence, management information systems, and IT project management.

Huimin Zhao is a Professor of Information Technology Management at the Lubar School of Business, University of Wisconsin-Milwaukee. He received the B.E. and M.E. degrees in Automation from Tsinghua University, China and the Ph.D. degree in Management Information Systems from the University of Arizona, USA. His current research interests include data mining and healthcare informatics. He has published in such journals as MIS Quarterly, Communications of the ACM, ACM Transactions on MIS, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Systems, Man, and Cybernetics, Information Systems, Journal of Management Information Systems, Journal of the AIS, and Decision Support Systems. He has served as a senior editor for Decision Support Systems and an associate editor for MIS Quarterly. He served as a co-chair of the 19th Workshop on Information Technologies and Systems (WITS), the 5th INFORMS Workshop on Data Mining and Health Informatics, and the 9th China Summer Workshop on Information Management (CSWIM).