

Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media

Yao Liu^a, Cuiqing Jiang^{a,c,*}, Huimin Zhao^b

^a School of Management, Hefei University of Technology, Hefei, Anhui 230009, China

^b Lubar School of Business, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, United States of America

^c Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei, Anhui 230009, China

ARTICLE INFO

Keywords:

User-generated content
Text mining
Competitive advantage
Competing product
Domain-specific sentiment analysis

ABSTRACT

User-generated content (UGC) is becoming increasingly available on social media for a wide range of products and services. Such UGC contains rich information about customer attitudes, opinions, and experiences. We propose a novel method for product competitive advantage analysis, which provides an essential basis for quality management and marketing strategy development, by mining UGC. Compared to traditional product performance analysis methods based on manufacturers' internal data and expert reviews, our method better reflects the perspective of customers. While a few recent methods based on UGC analysis assess the performance of one product in isolation, our method reveals the competitive advantages (and disadvantages) of a target product relative to its competitors. Our method uses supervised learning to identify competitors from UGC and domain-specific sentiment analysis to quantify customer attitudes. A case study in the automotive industry demonstrates the utility of the method.

1. Introduction

Manufacturers and marketers have long been focusing on product market performance, which benefits product marketing strategy development, product quality management decision making, and organizational performance improvement [1–3]. Thus, manufacturers and marketers continually collect product data to analyze product performance in the market. Traditionally, the main data sources for product performance analysis have been manufacturers' internal data, reviews from experts, and off-line customer surveys [1]. Product performance analysis methods relying heavily on traditional data incur significant labor consumption and time lag [4]. In addition, manufacturers and marketers' internal data and expert reviews mainly reflect the perspective of enterprises. However, insights into product performance from the perspective of customers are more essential, especially in a customer-driven market.

Recent years have witnessed the rapid development of social media platforms, such as online forums and online review sites, where more and more customers express their opinions, feelings, and concerns regarding products and services [5–7]. As user-generated content (UGC) becomes increasingly available for a wide range of products and services, it brings new opportunities for product performance analysis. Researchers have started to mine UGC for valuable information

reflecting product performance [4,8].

Compared to traditional data sources, such as manufacturers' internal data, expert reviews, and off-line customer surveys, UGC contains much richer customer opinions and feedback regarding not only a focal product but also its competing products [9]. While it is difficult for a manufacturer to obtain internal data of its competitors, UGC related to its competitors is easily accessible from social media platforms. However, most existing product performance evaluation methods based on UGC (e.g., [4,10–13]) just analyze one product in isolation. The *social comparison theory* suggests that competitions are ubiquitous, and people commonly seek to achieve a superior position over others in various contexts, from daily social situations to organizational settings and market transactions [14]. Nowadays, most markets, excluding the few monopoly markets, are competitive markets. In product purchase processes, customers often compare the pros and cons among the competitors and then choose the most suitable ones, based on their perceptions of multiple products [15,16]. The total market demand is finite, and competing products mutually influence each other in their market performance. Thus, accurate analysis of competitors' strengths and weaknesses is essential for a manufacturer to succeed in product quality management, product redesign, risk management, product marketing, and product pricing [6,16–18]. On the other hand, measuring product performance without consideration of other competing

* Corresponding author at: School of Management, Hefei University of Technology, Hefei, Anhui 230009, China.

E-mail address: jiangcuiqing2017@163.com (C. Jiang).

alternatives, as done in most existing product performance evaluation methods, may lead to incomplete and even misleading conclusions [19].

We strive to bridge this gap in the literature by proposing a novel method for product competitive advantage analysis based on UGC, assessing the competitive advantages (and disadvantages) of a target product relative to its competitors from the perspective of customers. Data on social media are voluminous, and it is unrealistic to understand all the UGC and obtain insightful suggestions manually [20–23]. To help manufacturers reduce labor cost, we develop an automated UGC-based approach to analyzing the competitive advantages (and disadvantages) of a target product via comparing customer attitudes toward the target product versus its competitors. UGC is a kind of unstructured data, making its analysis a daunting task [22,23]. Our proposed method leverages text mining on UGC to obtain insights into product competitive advantages.

The main components of our proposed method for product competitive advantage analysis include competing product identification and customer attitude analysis. Existing competitor identification and analysis methods are mainly based on industry and market structure, expert experience, or off-line surveys [24,25]. Methods based on industry and market structure constantly collect and analyze competitive information from marketing reports, trade journals, newspaper articles, and competitors' websites. However, most of such information is secondary information, whose objectivity can be questionable, and its volume is typically limited [9]. Methods based on expert experience or off-line surveys also have shortcomings, such as subjectivity and prohibitive cost [25]. Recently, some methods have been proposed to leverage UGC to identify competitors. The simplest way is to identify products that co-occur in the same text as competitors [26]. More advanced methods identify competitors from comparative UGC text [15,27,28]. Comparative text is a kind of text where customers compare and discuss multiple products, and these products usually have competitive relationships with each other [15]. Such information reflects the competitive market in the mind of customers. The products mentioned in comparative text have been deemed as competitors. In the existing methods for comparative text identification, rules and keywords predefined by experts have been employed to construct identification models [15,28,29]. However, such expert rule-based methods suffer from high manual cost on experts, the so-called *knowledge acquisition bottleneck*, and limitation in generalizability across application contexts. In our proposed method for product competitive advantage analysis, we identify a set of features characterizing UGC text and apply supervised machine learning techniques to train classifiers for automatically identifying comparative UGC text, thus addressing the shortcomings of expert rule-based methods.

Having identified competitors for a target product, our method quantifies and compares customer attitudes toward the products under comparison in various aspects. The comparative result reveals the competitive advantages (and disadvantages) of the target product relative to its competitors, and also allows the identification of strong competitors in various aspects. The sentiment of customer attitude has been widely used to assess product performance [30]. Sentiment analysis is a class of techniques for extracting and assessing opinions in text [31]. In related previous research, such as customer satisfaction analysis and opinion mining, the main approach is general sentiment analysis [4,30]. However, the performance of sentiment analysis is largely influenced by the application domain. Every domain has its idiomatic words and phrases, and the same word or phrase may have different meanings in different domains [32]. To improve the accuracy of competitive advantage analysis, our method uses domain-specific sentiment analysis by training a domain-specific sentiment lexicon.

2. Related work

2.1. Competitor identification

There are mainly three types of existing competitor identification methods: manager cognition-based methods, industry structure-based methods, and a few competitive sentence-based methods. In manager cognition-based methods, managers develop and apply a mental model of their competitors [25]. To illustrate how managers may apply the cognitive model to identify competitors, Clark and Montgomery [33] developed a process to guide managers' thought. First, managers imagine a profile of target firm as competitor in their mind. Second, managers retrieve competitors from memory and evaluate the similarities between the imaginary profile and the actual firms. Then, managers classify target firms based on the similarities. Although useful sometimes, such methods suffer from substantial subjectivity.

Products in the same market or industry usually serve the same needs or purposes and can therefore be regarded as close substitutes or competing products. Industry structure-based methods are from the perspective of economy. Chen [34] first developed a framework of competitors with two types of industry information: market commonality and resource similarity. Market commonality refers to the degree of market overlap between firms, such as the similarity of product features, functions, and customers they serve. Industry resource similarity has been used to identify firms with similar technologies (or resources) and capabilities as competitors. Firms with similar resource bundles are likely to have similar strategic capabilities, as well as competitive vulnerability, in the marketplace. Peteraf and Bergen [35] pointed out that managers may pay too much attention to rivals with the same types of resources, while neglecting rivals with dissimilar resource bundles that can also satisfy market needs. They classify candidate competitors based on similarities in terms of the market needs served and resource endowments.

These two types (i.e., manager cognition-based and industry structure-based) of methods identify competing firms as competitors. Such competitor identification at the firm level may not be sufficient for fine-grained product competitive advantage analysis. Recently, some methods have been proposed to identify competing products from UGC. The simplest method is to deem the products that co-occur in the same text as competitors [26]. However, there are various reasons for products to co-occur in the same text, even if the products are not really competitors of each other. Competitors of a target product can be considered as the products compared with the target product in comparative text, and hence some methods try to identify competitors from comparative UGC text [15,27,28].

Comparative text identification is still an open problem. There is no mature algorithm to determine whether a piece of text is making a comparison. When it comes to comparative UGC text, the problem becomes even more challenging, as UGC text is often informal and short. In previous research, comparative text has been identified using rules and keywords pre-defined by experts. Park and Blake [28] defined 35 features based on expert rules reflecting lexical and syntactic characteristics of a sentence. Li et al. [15] used eight patterns to cover most comparative text. Jindal and Liu [29] used 83 key words and phrases and 13 rules to identify comparative sentences. These methods based on expert rules demand large amount of labor and time cost of experts. More importantly, expert rule-based methods have limitations in generalizability across various applications and linguistic contexts.

2.2. Customer attitude analysis

Customer attitude is one of the key types of information in UGC that have been used in lots of applications, such as customer preference analysis, customer perception analysis, product recommendation, and sales forecasting. For example, Farhadloo et al. [36] used sentiment analysis to measure customer attitude from UGC text, and model the

aspect-overall customer attitude to analyze customer attitude toward different product aspects. Xiao et al. [30] analyzed customer attitude via sentiment analysis and used ordered choice model to measure customer preference. To mine customer perception from social media, Pournarakis et al. [4] used sentiment analysis and clustering to measure brand awareness and brand meaning. In product recommendation, customer attitude has been used to assess customer preferences toward products [37]. Additionally, customer attitude has served as essential information in sales forecasting [38].

Customer attitude toward products usually includes a positive aspect and a negative aspect, which are two kinds of information reflecting different viewpoints. The *theory of attractive quality* [39] distinguishes these two kinds of information, where positive attitude and negative attitude are analyzed separately to classify product attributes [40]. However, in the aforementioned research, customer attitude has been measured via sentiment analysis as a single metric, which aggregates customer satisfaction and dissatisfaction. Customer satisfaction and dissatisfaction reflect the good side and bad side of a product, respectively, and should be analyzed separately to gain deeper insights.

2.3. Sentiment analysis

Most existing research and applications implement sentiment analysis as classifying the directional state of sentiment, e.g., positive, negative, or neutral [32,41]. It is well known that the sentiment orientation of a word may change with the application domain [32]. To improve the performance in a particular domain, domain-specific sentiment analysis is needed. Supervised machine learning based methods and sentiment lexicon-based methods are two main types of methods for sentiment analysis. Both types have applications in domain-specific sentiment analysis [42–44].

Supervised machine learning based methods mine associations between sentiment classes and text features. The performance of such methods largely depends on the text features used [45]. The most common feature type is bag-of-words, which represents the words, phrases, and their frequency in documents using a vector. Bag-of-words incurs a high dimensional feature space and is prone to overfitting [21]. The weights of features can be adjusted to improve performance in some domains. More complex features have been proposed, but have not shown consistent and significant improvement over bag-of-words [32]. In addition, supervised learning requires a large training dataset. When applied to domain-specific sentiment analysis, it costs substantial time and human labor to mark a large training set, especially on social media, for every domain. This imposes a big obstacle on business practitioners and researchers.

The sentiment lexicon-based approach is another simple and effective approach to sentiment analysis [46]. Compared to machine learning-based methods, sentiment lexicon-based methods work well on short text, which is a major characteristic of UGC [32]. A sentiment lexicon, also called opinion lexicon, is a collection of words or phrases that are commonly used to express feelings [47]. Words or phrases in a sentiment lexicon have corresponding sentiment orientations, such as positive, negative, and neutral. When classifying a sentence, each word in the sentence will be checked against a sentiment lexicon. The simplest way is to just count the number of positive words and negative words as the sentiment score [48]. More sophisticated methods also consider and analyze the syntactic structure of the sentence, e.g., reversing the sentiment orientation based on negation words and adjusting the sentiment strength according to degree words used [49]. Lexicon-based methods are preferred to machine learning-based methods in absence of a large training dataset in a specific domain.

3. Proposed method for product competitive advantage analysis

We propose a novel method for product competitive advantage analysis (outlined in Fig. 1). Given a target product, the method

analyzes social media UGC to identify its competing products and assess its competitive advantages relative to the competitors. The method consists of several steps.

1. The social media data crawling step provides data foundation for the next steps. Customers generate large volumes of data on social media platforms, such as forums, blogs, microblogs, and social networks. Text data are the main type of UGC in which customers express their opinions.
2. Since social media UGC is typically unstructured, data preprocessing is necessary. Typical preprocessing includes useless element (e.g., hashtag, URL, and special symbol) removal, stop word removal, part of speech (POS) tagging, named entity recognition, and fake information filtering. In some special language contexts, e.g., Chinese, some special preprocessing, e.g., word segmentation, may be needed.
3. We intend to analyze product competitive advantages via comparing customer attitudes. Sentiment analysis is widely used to analyze customer attitudes and is chosen as one of the key components of the proposed method. To improve the accuracy of sentiment analysis, we develop a domain-specific sentiment lexicon first.
4. As discussed earlier, competing products can be found in comparative text. However, comparative sentences are rare in text reviews [50]. Thus, another main component of the proposed method is comparative UGC text classification. Competing products are then extracted from the comparative UGC text.
5. UGC regarding the target product and its competing products is then collected and preprocessed. The data crawling and preprocessing are similar to those in steps 1 and 2.
6. Finally, the competitive advantages (or disadvantages) of the target product relative to its competitors are analyzed based on the preprocessed UGC. Sentiment analysis based on the domain-specific sentiment lexicon is used to measure the customer attitudes toward the products in various aspects.

While data crawling and preprocessing (steps 1, 2, and 5) are relatively more common, we discuss the other steps, i.e., domain-specific sentiment lexicon generation (step 3), competing product identification (step 4), and competitive advantage analysis (step 6), in further details.

3.1. Domain-specific sentiment lexicon generation

Using social media UGC, we assess competitive advantages of a target product based on customers' attitude. Customer attitude analysis is a kind of sentiment analysis. Compared to domain-specific sentiment lexicons, general sentiment lexicons have disadvantages of inaccuracy and insufficiency, due to different usage preferences and varying sentiment orientations of words across domains [51]. We therefore generate a domain-specific sentiment lexicon in our method for product competitive advantage analysis. A sentiment lexicon can be labeled manually or generated using automated methods. To avoid the enormous cost of the manual approach, we use an automated method, such as that proposed by Deng et al. [32], to generate a domain-specific sentiment lexicon automatically. The method proposed by Deng et al. [32] starts with a seed sentiment lexicon and gradually expands it with new sentiment words identified from a large unlabeled corpus from a particular domain.

The seed sentiment lexicon is a small set of high-frequency positive sentiment words and negative sentiment words. The sentiment orientation of a candidate sentiment word is determined by its co-occurrence—measured using Pointwise Mutual Information (PMI)—with the two categories (i.e., positive and negative) of sentiment words in the corpus.

The PMI between two words w_1 and w_2 is defined as

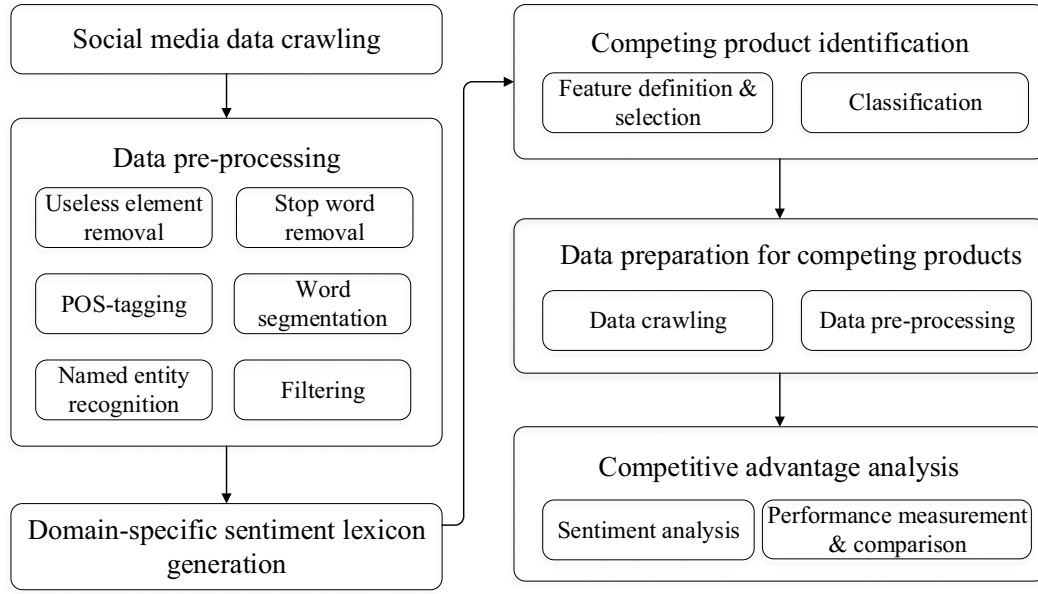


Fig. 1. Proposed method for competitive advantage analysis.

$$PMI(w_1, w_2) = \log_2 \frac{p(w_1, w_2)}{p(w_1)p(w_2)}, \quad (1)$$

where $p(w)$ is the probability that word w occurs in a document and $p(w_1, w_2)$ is the probability that w_1 and w_2 co-occur in a document.

The average PMI between a candidate word w and the set of positive words in the lexicon, Pos , and the average PMI between w and the set of negative words, Neg , are defined, respectively, as

$$AVG_{PMI}(w, Pos) = \frac{1}{|Pos|} \sum_{p \in Pos} PMI(w, p), \quad (2)$$

$$AVG_{PMI}(w, Neg) = \frac{1}{|Neg|} \sum_{n \in Neg} PMI(w, n). \quad (3)$$

Then, the sentiment orientation of w is determined by

$$Orientation(w) = \begin{cases} \text{Positive} & \text{if } AVG_PMI(w, Pos) - AVG_PMI(w, Neg) \geq H \\ \text{Negative} & \text{if } AVG_PMI(w, Pos) - AVG_PMI(w, Neg) \leq -H \end{cases}, \quad (4)$$

where H is a threshold.

Once the sentiment orientation of w is determined, w is added into Pos (Neg) if w is deemed a positive (negative) sentiment word. The procedure is repeated until no more sentiment word can be found.

Table 1
Features used in UGC classification.

Problem	Studies	Feature type				
		Product	Keyword	Sentiment	Linguistic	Social
Product defect identification	[56]	✓	✓	✓	✓	✓
	[21]		✓		✓	✓
Product quality improvement	[48]		✓	✓	✓	✓
Fake review detection	[53]			✓		✓
Toy safety surveillance	[54]		✓			
Helpful review identification	[59,60,63]			✓	✓	
	[55]					✓
	[62]				✓	✓
	[61]				✓	
Stock price movement forecasting	[58]			✓		
Comparative text identification	This study	✓	✓	✓	✓	✓

3.2. Competing product identification

Analyzing competing products from the perspective of customers is more reasonable under market competition. Our method for competing product identification is based on comparative UGC text where customers compare various products during purchase or usage. Specifically, we first identify comparative UGC text and then extract competing products mentioned in the identified comparative text. To overcome the knowledge acquisition bottleneck and limited generalizability of rule-based methods for comparative text identification (e.g., [15,28,29]), we use supervised machine learning to classify whether a piece of UGC text involves comparison of products. Text classification has been widely used in social media text analysis applications, such as online review helpfulness evaluation [52], fake review identification [53], and product defect identification [54]. The classification performance largely depends on the choice of features and classification methods [55].

3.2.1. Feature definition

Previous studies have proposed various types of features for different text classification problems (e.g., [22,56,62]). However, the effectiveness of features changes with the application context. The features used in other text classification problems cannot be directly adopted in the new context, i.e., comparative UGC text classification. We therefore identify potentially useful features for this new problem according to its characteristics, with support from related literature. As

a result, we come up with five types of promising features: product features, keyword features, sentiment features, linguistic features, and social features. Table 1 summarizes the use of these types of features in existing UGC classification studies.

3.2.1.1. Product features. Product features are defined to describe product attributes and product components. They have been used to identify product defects from customer reviews [56]. Product features are potentially useful in the context of this study too. The names of products and product components are typically mentioned when two or more products are compared in UGC text. We can extract the product names and product components mentioned. We can also count the number of product names and the number of product components mentioned.

3.2.1.2. Keyword features. Keyword features have been widely used in text classification. There are obvious differences in the frequencies with which certain words appear in different classes of text [57]. Keyword lists have been extracted to identify product defects mentioned on social media [21,56]. Jiang et al. used keyword features to capture helpful reviews from social media for product quality improvement [48]. In finance lists of negative (“bearish”) words and positive (“bullish”) words have been created [58]. Winkler et al. developed a danger word list from injury and recall text narratives and used it in uncovering potentially dangerous children's toys [54]. Comparative sentences also can be recognized from the use of comparative adjective words e.g. “better”, “more stable”, and “cheaper” [15]. Therefore keywords are potentially useful for identifying comparative UGC text. In this study we extract the words used in a corpus as candidate keywords first. Since an excess of irrelevant features will increase the complexity of a classification model and cause overfitting

we subsequently use feature selection to reduce the dimensionality of the feature space (further discussed later)

3.2.1.3. Sentiment features. Sentiment features measure subjectivity, positivity, negativity, or overall user rating, for words, sentences, or entire documents [56,59]. Sentiment features are commonly used in social media UGC analysis, as they directly reflect the attitude expressed by customers. Cao et al. found that social media reviews expressing either positive or negative opinions convey more information about customer attitude and are more helpful than those expressing neutral opinions in some contexts [60]. Oh and Sheng found that investor sentiment expressed in StockTwits postings has strong predictive ability for future stock price directional movement [58]. Zhang et al. used sentiment features to detect fake online reviews [53]. Sentiment features have also been shown to be strong indicators in helpful review analysis and identification [48,59,63]. When a customer compares competing products, sentiment is necessary to describe the customer's attitudes toward various competing products. Thus, we extract sentiment features for comparative UGC text classification. The domain-specific sentiment lexicon (described earlier) is also useful for deriving such sentiment features tailored to the target domain.

3.2.1.4. Linguistic features. Linguistic features reflect the styles and characteristics of the vocabulary and sentences used. Linguistic features play a significant role in social media text classification [61,63]. They have been used in a variety of applications, including product defect identification [21,56], helpful review identification [59–63], and product quality improvement [48]. UGC text comparing competing products also tends to use particular categories of sentences and vocabulary. Therefore, we extract linguistic features, e.g., the number of (a particular type of) words or sentences, for comparative UGC text classification.

3.2.1.5. Social features. Social features quantify the attention received on social media and social attributes of users [55]. They have found applications in many problems, including product defect identification [21,56], helpful review identification [55,62], product quality improvement [48], and fake review detection [53]. People on social media have different preferences for various topics, while a certain topic may attract customers with particular characteristics. For example, the most popular threads in automotive forums are automotive maintenance and travel, whereas the participants of purchase-related discussions on specific product forums mainly consist of newbies on the platform [21]. Thus, competing product identification from UGC may also benefit from social features. Examples of social features regarding a product review include the number of views and the number of replies. Those regarding the reviewer include the level of the reviewer and the number of reviews/replies posted by the reviewer.

3.2.2. Feature selection

We have identified a comprehensive set of candidate features, inevitably leading to a prohibitively large feature space. It is essential to select a small and effective feature subset. There are many feature selection methods. One method that has been shown to be effective in social media UGC classifications, e.g., product defect identification from online discussion forums [21], is the Correlation-based Feature Selection (CFS) method [64]. CFS takes a heuristic optimization approach. As for the optimization objective, CFS considers not only the correlation between features and the class (i.e., whether a piece of UGC text is comparative or not) but also the correlation among features. It uses a metric called “Merit” to assess the effectiveness of a feature subset. Specifically, the “Merit” of a feature subset S containing k features is defined as

$$Merit_s = \frac{k * \bar{R}_{cf}}{\sqrt{k + k * (k - 1) * \bar{R}_{ff}}}, \quad (5)$$

where \bar{R}_{cf} is the average Pearson's correlation between the features in S and the class, and \bar{R}_{ff} is the average Pearson's correlation between the features in S . Then, best-first search, a heuristic algorithm, is used to search for an optimal feature subset in terms of “Merit”. Best-first search starts with the empty set and expands it with one new feature. The subset with the highest merit is chosen and is expanded again with one new feature. If no expanded subset has higher “Merit”, the algorithm retreats to the next-best unexpanded subset and continues from there. At the end, best-first search returns the best feature subset found.

3.2.3. Classification methods

Many machine learning methods, e.g., naïve Bayes, decision tree, decision table, logistic regression, and k-nearest neighbor, have been investigated for social media text classification. In addition, prior studies (e.g., [21]) have shown that ensemble learning often helps to improve performance for text classification. Ensemble learning is a machine learning paradigm where multiple base models are trained and then combined. Some representative ensemble learning methods include boosting, bagging, and random subspace [31].

3.3. Competitive advantage analysis

To analyze the competitive advantages (and disadvantages) of a target product, we compare the customer attitude toward the target product with that toward its competitors in various aspects. We use the domain-specific sentiment lexicon to analyze the sentiment expressed in UGC text, and then use several metrics defined based on the sentiment scores to measure the customer attitude toward a product in some aspects.

One metric is the *Average Sentiment Orientation* (ASO). With this metric, we treat each customer review equally and only consider its

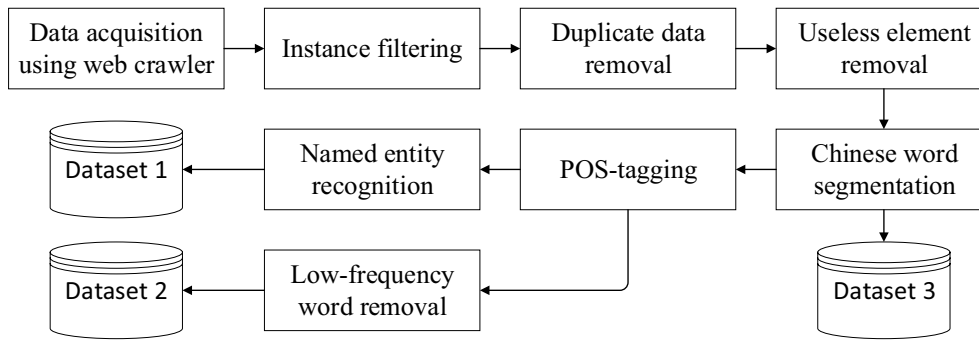


Fig. 2. Data collection and preprocessing steps.

sentiment orientation (positive, negative, or neutral). Specifically, $ASO = \frac{p-n}{N}$, where N is the number of all instances (i.e., customer reviews), p is the number of positive instances, and n is the number of negative instances.

Another metric is *Average Sentiment Score* (ASS). Sentences with the same sentiment orientation may still substantially differ in their strengths (from mild to extreme). Therefore, we assess not only the sentiment orientation but also the sentiment strength of each piece of UGC text. Such sentiment strength analysis requires the consideration of the semantic structure of sentiment, e.g., negation and booster of sentiment degree. Then, $ASS = \frac{1}{N} \sum_{i=1}^N Strength_i$, where $Strength_i$ is the sentiment strength of the i -th instance.

ASO and ASS reflect holistic customer satisfaction aggregating both satisfaction and dissatisfaction. In this aggregation, customer satisfaction and dissatisfaction offset each other. However, as discussed earlier, customer satisfaction and dissatisfaction reflect the good side and bad side of a product, respectively, and should be analyzed separately to gain deeper insights into the customer attitude. Positive (negative) instances reflect customer satisfaction (dissatisfaction). We use the average sentiment strength of positive (negative) instances as a metric for average customer satisfaction (dissatisfaction).

We evaluate the ASS, ASO, satisfaction, and dissatisfaction of the target product and its competitors in various aspects. The competitive advantages (disadvantages) of the target product in comparison with its competitors, i.e., the aspects where the target product is ranked higher (lower) than most competitors, with respect to the metrics can then be identified.

4. Case study

We have applied and evaluated our proposed method in a case study in the auto industry in China. China has been the largest auto market in the world since 2009. In 2017, the auto sales in China reached a historical peak of > 28.88 million, a quarter of the total auto sales in the world, according to the Chinese Industrial Information Web (www.chyxx.com). Fierce competition inevitably comes with the huge market and profit. Thus, competition analysis in the Chinese auto market is very important for auto manufacturers and dealers. In this case study, the Chinese auto market was the context, and Volkswagen Passat, one of the top sellers in this market, was chosen as the target product to be analyzed. Volkswagen Passat is a typical medium-sized vehicle, has a huge customer base, and receives enormous attention in the Chinese auto market. Hence, there is abundant social media UGC regarding this product for analysis.

4.1. Data

We selected autohome.com.cn and bitauto.com, two largest auto product websites in China, as data sources. We collected data in September 2018 from these websites for competing product identification, domain-specific sentiment lexicon generation, and competitive

advantage analysis, respectively.

The dataset for competing product identification consists of postings at the forums of the two websites specifically for the target product, where customers can express any opinion toward or experience with the target product. The postings contain lots of comparisons with various competing products, allowing the identification of major competitors for the target product. However, this kind of data has some disadvantages for aspect-specific customer attitude analysis of competing products, which is the basis for competitive advantage analysis, in that it would require product aspect extraction and corresponding customer attitude measurement.

For aspect-specific customer attitude analysis, we collected another dataset. The two auto websites also host the so-called “word-of-mouth”, besides forums. In the word-of-mouth, automotive features are divided into eight aspects: space, power, control, gasoline consumption, appearance, interior, cost-performance, and comfort. Customers are asked to give their comments regarding each product aspect separately. The word-of-mouth data effectively avoid the disadvantages of the forum data for aspect-specific customer attitude analysis. We therefore collected the word-of-mouth data regarding the target product and the identified major competing products (after they have been identified) from the two websites.

To generate a high-quality domain-specific sentiment lexicon, a large corpus is essential. Since the lexicon is generated for the purpose of aspect-specific customer attitude analysis on the “word-of-mouth” data, we used the same data source. However, the particular products are not essential for the sentiment lexicon. Therefore, we collected even more “word-of-mouth” data regardless of the product from the two websites for the domain-specific sentiment lexicon generation.

Fig. 2 outlines the data collection and preprocessing steps. In the data crawling process, texts with fewer than five words and duplicate data were deemed useless and hence deleted (i.e., instance filtering and duplicate data removal in Fig. 2). Datasets 1, 2, and 3 were used for competing product identification, domain-specific sentiment lexicon generation, and competitive advantage analysis, respectively. Ultimately, we obtained > 20,000 forum postings for competing product identification, > 280,000 word-of-mouth records for domain-specific sentiment lexicon generation, and > 183,000 word-of-mouth records regarding the target product and major competing products for aspect-specific customer attitude analysis.

The common data preprocessing steps include useless element (i.e., hashtag, URL, stop word, and special symbol) removal and Chinese word segmentation. 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. Chinese word segmentation segments a Chinese sentence into a sequence of words, which are the basic meaningful units for processing. Comparative UGC text classification and domain-specific sentiment lexicon generation also require POS-tagging, i.e., tagging the part of speech of each word in a sentence, such as noun, adjective, and verb. In this case study, we used the Jieba Chinese word segmentation program

(<https://github.com/fxsjy/jieba>), a popular Chinese word segmentation program, for Chinese word segmentation and POS-tagging.

In comparative UGC text classification, product names and product components need to be identified (i.e., named entity recognition in Fig. 2). Thus, we constructed an auto product name list and an auto component name list. For the auto product name list, we crawled all the auto product home pages listed by the two auto product websites autohome.com.cn and bitauto.com, and thereby collected 2453 auto product names. For the auto component name list, we extracted the 301 auto component names included in the auto component list published in Wikipedia (titled “list of auto parts”). In sentiment lexicon generation, words with very low (i.e., < 0.01%) frequency were removed.

Most of the data collection and preprocessing steps outlined in Fig. 2 are general over languages, although the particular tools for some of the steps (e.g., POS tagging) may be language-specific. The word segmentation step is unnecessary for some languages that use word separators (e.g., English).

4.2. Domain-specific sentiment lexicon development

Sentiment analysis is needed for both competing product identification and competitive advantage analysis. To improve the performance of sentiment analysis, we trained a domain-specific sentiment lexicon using the method proposed by Deng et al. [32] based on the corpus of > 280,000 word-of-mouth records. We manually selected 60 positive words and 60 negative words with high frequency as the seed lexicon. Specifically, we ranked the words in the corpus by occurrence frequency, and then selected the top 60 positive words and the top 60 negative words tagged by three doctoral students. Only the words with consistent labels from the three taggers were selected as seed sentiment words.

We evaluated the effectiveness of the generated domain-specific sentiment lexicon, in comparison with two general sentiment lexicons from NTUSD (National Taiwan University Sentiment Dictionary, academiasinicanlplab.github.io) and Hownet (www.keenage.com), which have been widely used in Chinese sentiment analysis. We plugged each sentiment lexicon, along with the widely-used booster word list and negation word list from Hownet, into SentiStrength (sentistrength.wlv.ac.uk), a widely used sentiment analysis tool, and tested the performance (classification accuracy) on a set of pre-labeled sentences. SentiStrength was originally developed for English but can be configured for other languages by changing its input files, including sentiment lexicon, booster word list, and negation word list. SentiStrength gives a positive sentiment value and a negative sentiment value for each instance, and the sum of these two values was used as the strength of customer attitude. To construct a testing dataset for sentiment analysis, we employed three doctoral students to label a set of sentences from the word-of-mouth records and selected 600 positive sentences and 600 negative sentences with consistent labelling results from the three taggers.

The domain-specific sentiment lexicon generation method uses a parameter H to balance the sensitivity and specificity of the identified sentiment words. Thus, we tested several values of H to find an appropriate value.

Table 2 summarizes the results. The domain-specific sentiment lexicon generation method yielded the best performance when the parameter H was set to 9. The trained domain-specific sentiment

lexicon ($H = 9$) outperformed both general sentiment lexicons from NTUSD and Hownet; the performance improvement was statistical significance ($p < 0.001$ according to McNemar's test). We therefore used this trained domain-specific sentiment lexicon in the subsequent sentiment analysis.

4.3. Competing product identification

To identify competing products for the target product, we first classified whether each of the forum postings contains comparative text. To train and test classifiers, we manually labeled a subset of 10,000 postings (233 positive cases and 9767 negative cases), which were used as training and testing data. Specifically, we employed three doctoral students to label the postings. Their results turned out to be highly consistent, with Kappa coefficients of 0.77, 0.79, and 0.84, respectively. The final labels were determined via majority voting. To deal with the severe class imbalance, we balanced the two classes by assigning a higher weight to positive cases. The trained classifiers were then used to classify the remaining 10,000 forum postings. From the postings containing comparative text, we identified potential competing products based on the auto product name list.

4.3.1. Feature extraction and selection

As discussed earlier, several types of features, such as product features, keyword features, sentiment features, linguistic features, and social features, are potentially useful for the forum posting classification. We therefore extracted these types of features (listed in Table 3).

After candidate feature extraction, we used CFS (discussed earlier) to select an effective feature set. The final selected feature set consists of 66 features, which were used in the subsequent comparative text identification.

4.3.2. Comparative text identification

Based on the 66 selected features, we classified whether a forum posting contains comparative text. We conducted a series of experiments using Weka (www.cs.waikato.ac.nz/ml/weka/). We compared several standard classification methods and used them as the base classification methods in ensemble learning methods. We selected the best performer among all base and ensemble classifiers as the final classifier. Under every experiment setting, we used 10 independent 10-fold cross validations, resulting in 100 estimates of the classification performance. We gauged the classification performance using the F-measure and Matthews' correlation coefficient (MCC).

First, we tested five standard classification methods: C4.5 decision tree (named J4.8 in Weka), decision table, k-nearest neighbor (kNN), logistic regression, and naïve Bayes, retaining the default parameter settings of Weka. We then tested three representative ensemble learning methods, boosting, bagging, and random subspace [31], with each of the five standard classification methods as the base classifier learner, retaining the default parameter settings of Weka. For boosting, we used AdaBoostM1, one of the widely-used boosting methods. Table 4 summarizes the result. Bagging with logistic regression was the overall best performer. Therefore, we used the classifier trained with bagging and logistic regression as the final classifier for comparative text identification from forum postings.

Table 2
Performance of different sentiment lexicons.

	NTUSD	Hownet	Trained domain-specific sentiment lexicon with $H =$					
			10	9	8	7	6	5
Accuracy (%)	74.25	64.25	86.33	86.58	86.00	85.00	85.08	84.58

Table 3
Candidate features extracted.

Feature type	Feature
Product features	Whether one of the 301 auto components is mentioned
	Number of auto components mentioned
	Whether one of the 2453 auto product names is mentioned
	Number of auto product names mentioned
Sentiment features	Number of positive sentiment words
	Number of negative sentiment words
	Number of booster words at degree 1
	Number of booster words at degree 2
	Number of booster words at degree 3
	Number of booster words at degree 4
	Number of booster words at degree 5
	Number of booster words at degree 6
Keyword features	Words with occurrence frequency over 0.1%
Social features	Number of views
	Number of replies
	The level of the author assigned by the website
	Number of postings written by the author
Linguistic features	Number of replies written by the author
	Number of words
	Number of sentences
	Number of exclamatory sentences
	Number of interrogative sentences
	Number of adjectives
	Number of verbs
	Number of adverbs
	Number of nouns
	Number of modal particles
	Number of mimetic words

Table 4
Classification performance of base and ensemble classifiers.

Ensemble method	Base method	F-measure (stddev)	MCC (stddev)
AdaBoostM1	kNN	0.55 (0.11)	0.58 (0.10)
	J4.8	0.82 (0.07)	0.82 (0.07)
	Decision table	0.83 (0.05)	0.84 (0.04)
	Logistic regression	0.91 (0.05)	0.91 (0.05)
	Naïve Bayes	0.51 (0.04)	0.55 (0.04)
	kNN	0.55 (0.11)	0.58 (0.10)
	J4.8	0.89 (0.05)	0.89 (0.05)
	Decision table	0.85 (0.05)	0.85 (0.05)
	Logistic regression	0.91 (0.05)	0.91 (0.05)
	Naïve Bayes	0.81 (0.06)	0.81 (0.06)
Bagging	kNN	0.54 (0.11)	0.58 (0.10)
	J4.8	0.83 (0.06)	0.84 (0.05)
	Decision table	0.82 (0.06)	0.83 (0.05)
	Logistic regression	0.92 (0.04)	0.92 (0.04)
	Naïve Bayes	0.51 (0.04)	0.55 (0.04)
Random subspace	kNN	0.53 (0.11)	0.59 (0.10)
	J4.8	0.79 (0.08)	0.80 (0.07)
	Decision table	0.76 (0.09)	0.78 (0.07)
	Logistic regression	0.88 (0.06)	0.88 (0.06)
	Naïve Bayes	0.54 (0.05)	0.57 (0.05)

4.3.3. Competing product identification

Using the classifier trained with bagging and logistic regression, we identified forum postings that contain comparative text. From the identified postings that contain comparative text, we identified the auto product names mentioned based on the auto product name list. We deemed the products that are mentioned in these postings with a frequency of > 5% as the main competitors of the target product. Table 5 lists the identified 22 main competing products.

The two auto product websites (autohome.com.cn and biauto.com) also provide some hot competitors for each product. For example, the five competitors of Passat provided by the two websites are Regal, Audi A4 L, Magotan, Camry, and Accord, which also appear among the top six competitors identified by the proposed method, thus showing the validity of the proposed method to some extent. In addition, the proposed method identified more competitors from the perspective of

Table 5
Main competing products of Volkswagen Passat identified.

Order	Product	Frequency mentioned (%)	Order	Product	Frequency mentioned (%)
1	Magotan	48.92	12	Lavida	12.02
2	Mondeo	33.91	13	Citroen C5	9.44
3	Accord	24.89	14	BMW 3 series	9.01
4	Camry	23.18	15	Malibu	8.15
5	Audi A4 L	21.89	16	KIA K5	6.44
6	Regal	18.03	17	Benz C series	6.44
7	LaCROSSE	16.74	18	Focus	6.01
8	TEANA	16.31	19	Sanata	6.01
9	Peugeot 508	14.59	20	Reiz	6.01
10	Tiguan	13.30	21	Atenza	5.58
11	Sagitar	12.45	22	Jetta	5.15

customers. Specifically, Mondeo, the second most frequently compared competitor identified by the proposed method, is missing in the list provided by the two websites. This shows the value of the proposed method in identifying competing products perceived by customers beyond relying on a given fixed list of deemed competitors.

4.4. Competitive advantage analysis

To assess the competitive advantages (and disadvantages) of the target product relative to its 22 main competitors, we measured the customer attitude regarding various aspects of these products. We crawled word-of-mouth data for these products and used domain-specific sentiment analysis to analyze customer sentiment toward each product in each of eight aspects: space, power, control, gasoline consumption, appearance, interior, cost-performance, and comfort. We collected > 183,000 word-of-mouth records from the 46 product forums (for the target product and its 22 main competitors, respectively) on the two auto websites autohome.com.cn and bitauto.com. We used SentiStrength, with the trained domain-specific sentiment lexicon, along with the booster word list and negation word list from HowNet, to measure the sentiment strengths (positive and negative) of the comment under each aspect of a word-of-mouth record. We then computed the ASO and ASS values based on the aspect-specific sentiment analysis results and ranked the 23 products on each of the eight aspects based on these metrics.

Table 6 summarizes the ranking results. According to both ASO and ASS, power, space, and appearance are the top three aspects of the target product, relative to its competitors. The target product performs better than most of its competitors in these three aspects. Thus, these aspects appear to be competitive advantages of the target product. Gasoline consumption is among the bottom aspects of the target product according to both ASO and ASS, and therefore appears to be a competitive disadvantage of the target product. However, high gasoline

Table 6
The ASO ranks and ASS ranks of the target product.

Order	Aspect	ASO rank (percentage)	Aspect	ASS rank (percentage)
1	Power	2 (8.70%)	Space	5 (21.74%)
2	Space	3 (13.04%)	Power	7 (30.43%)
3	Appearance	3 (13.04%)	Appearance	8 (34.78%)
4	Control	7 (30.43%)	Comfort	9 (39.13%)
5	Cost-performance	7 (30.43%)	Interior	10 (43.48%)
6	Comfort	8 (34.78%)	Control	11 (47.83%)
7	Interior	9 (39.13%)	Gasoline consumption	18 (78.26%)
8	Gasoline consumption	16 (69.57%)	Cost-performance	19 (82.61%)

Table 7
ASO values.

Order	Aspect	Target product	Competing products			
			Mean	Stddev	Max	Min
1	Appearance	0.856	0.767	0.064	0.873	0.653
2	Cost-performance	0.628	0.576	0.113	0.761	0.325
3	Space	0.538	0.396	0.144	0.685	0.016
4	Interior	0.537	0.463	0.172	0.792	0.086
5	Comfort	0.521	0.411	0.163	0.671	0.123
6	Control	0.505	0.406	0.122	0.590	0.158
7	Power	0.236	0.092	0.131	0.432	-0.182
8	Gasoline consumption	0.090	0.161	0.098	0.308	0.000

consumption is perhaps inevitable for a vehicle to achieve high performance on power. It is difficult to excel on both power and gasoline consumption simultaneously. Thus, finding an appropriate tradeoff between power and gasoline consumption based on customer feedback is a more realistic goal for the manufacturer. The other bottom aspects are comfort and interior according to ASO and control and cost-performance according to ASS, respectively. These aspects may also be competitive disadvantages of the target product.

To illustrate the difference between our method and previous studies that analyze various aspects of a product alone without comparing with competing products, we list the ASO and ASS values of the target product in Tables 7 and 8, respectively. Some of the aspects with high (low) ASO or ASS values may actually have low (high) ranks when compared with competing products, because many competitors have even higher (lower) values. For example, when looking at the ASO and ASS values of the target product alone, cost-performance appears to be one of the top aspects (second and third highest ASO and ASS values among the eight aspects, respectively). However, when compared with competing products, cost-performance of the target product is only ranked at the fifth and the last among the eight aspects based on ASO and ASS, respectively, and is clearly not a competitive advantage of the target product. On the other hand, power has the second lowest ASO and ASS values among the eight aspects. However, when compared with competing products, power is highly ranked based on both ASO and ASS (first and second based on ASO and ASS, respectively) among the eight aspects and may well be one of the competitive advantages of the target product. Such differences show that examining various aspects of a product in isolation without comparing with its competitors may not be informative and may even be misleading when assessing the competitive advantages (and disadvantages) of the product, as competitive advantages (and disadvantages) are by nature comparative.

The results show that the ASS ranks of the target product are lower than its ASO ranks. ASO measures the average orientation of customer attitude whereas ASS measures the average orientation and strength of customer attitude. The reason for the difference between ASS ranks and ASO ranks lies in the difference between the strengths of satisfaction

Table 8
ASS values.

Order	Aspect	Target product	Competing products			
			Mean	Stddev	Max	Min
1	Appearance	7.737	7.298	1.800	10.807	4.945
2	Interior	3.704	3.818	1.797	7.559	0.858
3	Cost-performance	3.484	4.154	1.734	6.271	2.438
4	Comfort	3.261	3.291	1.434	5.872	0.871
5	Control	3.214	3.022	1.143	5.291	0.980
6	Space	3.189	2.447	0.973	4.014	0.104
7	Power	0.763	0.241	0.989	2.216	-1.874
8	Gasoline consumption	0.327	0.788	0.549	2.046	0.083

and dissatisfaction of customers. Additionally, satisfaction and dissatisfaction are different aspects of customer attitude. Therefore, for finer-grained analysis, we further examined satisfaction (i.e., average sentiment strength of positive comments) and dissatisfaction (i.e., average sentiment strength of negative comments) separately.

Table 9 shows the satisfaction ranks and dissatisfaction ranks of the target product on the eight aspects. There is little correlation between the satisfaction rank and the dissatisfaction rank (Spearman correlation = 0.07). The correlation between the satisfaction rank and the ASS rank and that between the dissatisfaction rank and the ASS rank are also low (Spearman correlations are 0.26 and -0.26, respectively). The low correlations indicate the need to examine satisfaction and dissatisfaction separately, beyond an aggregate metric.

The results show that the satisfaction ranks of the target product in most aspects are lower than its dissatisfaction ranks. That is the main reason why the ASS ranks are lower than the ASO ranks. Customers with positive attitude have lukewarm sentiment toward most aspects except appearance and space. The positive attitude can also be considered as customer satisfaction [30]. Kano proposed the *theory of attractive quality*, which has been widely adopted by researchers and practitioners [39]. In this theory, the quality attributes of a product are categorized as must-be, one-dimensional, or attractive. Must-be attributes are those related to the basic specifications of a product, and customers take for granted that a product will possess them. One-dimensional attributes are related to the quality expected by customers. An attribute at this level satisfies the requirements for customer usage and provides the quality that the customer expects. An attractive quality attribute is understood to be something that delights customers. The reason is that an attractive attribute both satisfies the potential demands of customers and exceeds their expectations. The core of this theory is attractive quality, which can surprise customers and win customer loyalty. A product with enough attractive quality can attract customers more easily, hence helping to generate competitive advantages. The performance of the target product on customer satisfaction indicates that the target product has less attractive quality than some of its competitors. The manufacturer should pay more attention to attractive quality development, which is an effective way to keep competitive advantage and customer loyalty [13].

The satisfaction and dissatisfaction ranks reveal deeper insights, beyond the ASO and ASS ranks, into the competitive advantages (and disadvantages) of the target product. For example, as discussed earlier, power seems to be a competitive advantage of the target product, as it is highly ranked on both ASO and ASS. However, looking at satisfaction and dissatisfaction separately, the truth is not so positive. Power actually has the worst rank in both satisfaction and dissatisfaction, indicating that the average extent of satisfaction regarding the power aspect of the target product among satisfied customers is lower while the average extent of dissatisfaction among dissatisfied customers is higher, compared with competing products. The target product outperforms most competitors in the power aspect in terms of overall ASO and ASS, not because its customers are more satisfied or less dissatisfied, but only because it has a higher proportion of satisfied customers and a lower proportion of dissatisfied customers (i.e., the target product wins by number). In the gasoline consumption aspect, the target product has low ranks in all four metrics. It is clear that customers are largely dissatisfied with the target product in this aspect. Thus, the manufacturer may try to better balance between power and gasoline consumption. In the aspects of control, comfort, cost-performance, and interior, the ASO and ASS ranks of the target product are relatively lower, while the dissatisfaction ranks are relatively higher. There is less praise, but also less criticism, regarding these aspects, comparing to competitors. Thus, it may be easier for the target product to improve in these aspects. The target product has lots of potential to reduce quality gaps from competitors in these aspects.

Table 9
The satisfaction ranks and dissatisfaction ranks of the target product.

Order	Attribute	Satisfaction rank (percentage)	Attribute	Dissatisfaction rank (percentage)
1	Appearance	10 (43.48%)	Comfort	3 (13.04%)
2	Space	14 (60.87%)	Control	5 (21.74%)
3	Control	17 (73.91%)	Interior	7 (30.43%)
4	Gasoline consumption	17 (73.91%)	Cost-performance	11 (47.83%)
5	Comfort	18 (78.26%)	Space	12 (52.17%)
6	Interior	19 (82.61%)	Appearance	12 (52.17%)
7	Cost-performance	20 (86.96%)	Gasoline consumption	14 (60.87%)
8	Power	22 (95.65%)	Power	19 (82.61%)

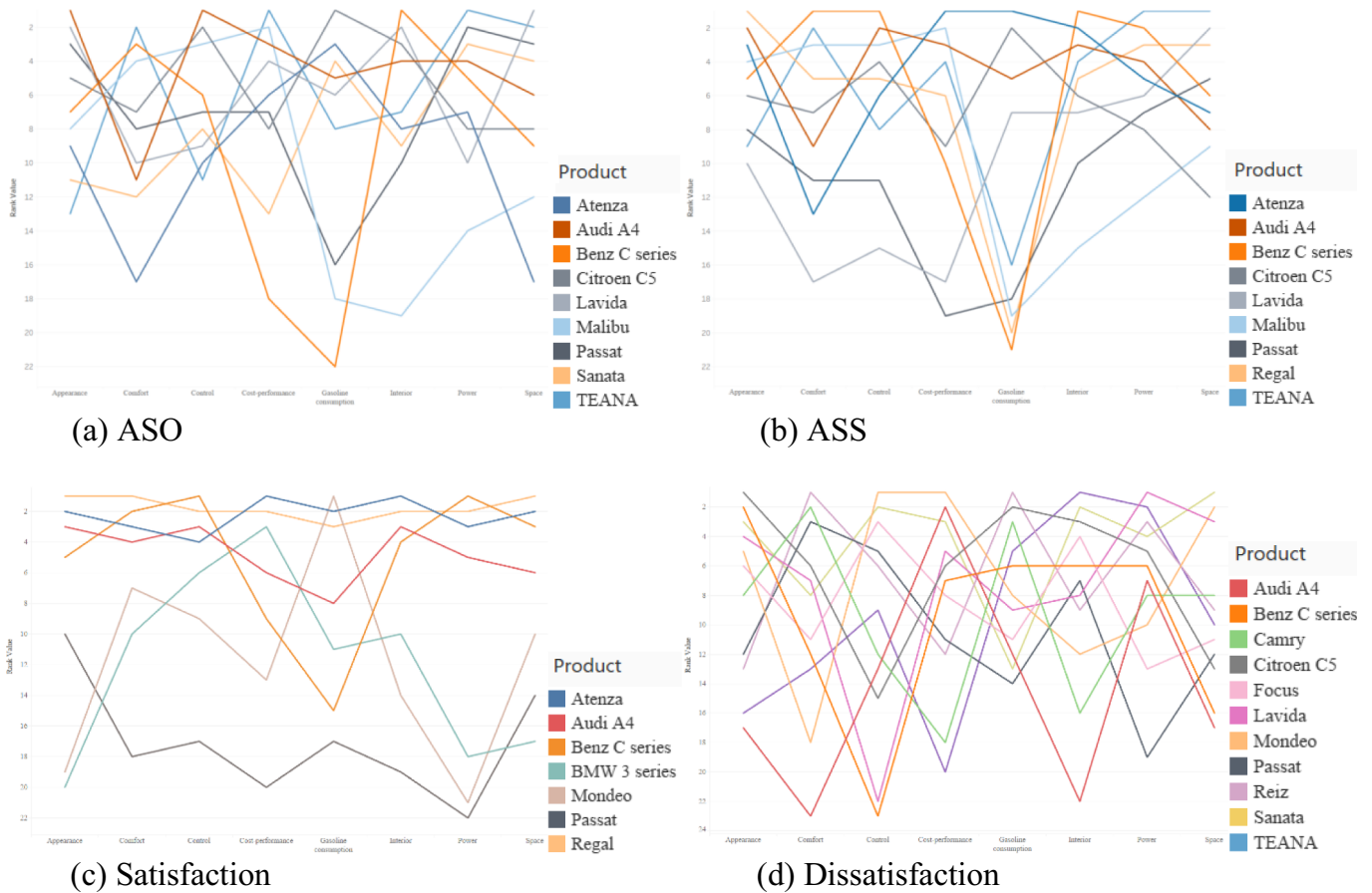


Fig. 3. The target product and strong competitors.

4.5. Identify strong competitors

The ranking results also allow the identification of strong (and weak) competitors for the target product. Fig. 3 shows parallel coordinates along the eight aspects based on ASO, ASS, satisfaction, and dissatisfaction ranks, respectively, including products ranked in the top three in any aspect, along with the target product. The top performers in each aspect in terms of each metric can be directly identified in the figure. Furthermore, the average rank over all eight aspects may be examined to identify overall strong competitors. The top three products in terms of average ASO, ASS, and satisfaction ranks over the eight aspects are all the same, i.e., Atenza, Audi A4, and Benz C series. The top three products in terms of average dissatisfaction rank over the eight aspects are Benz C series, Camry, and Citroen C5.

5. Discussion

Our proposed method potentially adds a new tool to the decision

support systems of manufacturers and marketers. Compared to traditional product performance analysis methods based on manufacturers and marketers' internal data and expert reviews, this method better reflects the perspective of customers. *The gaps model of service quality* indicates that there is a gap between customer expectations of a product and the manufacturer's understanding of those expectations [65]. A primary cause in many firms for not meeting customers' expectations is that the firms lack accurate understanding of their customers. Thus, the proposed method provides a valuable addition on top of traditional methods.

However, we caution that every method has its limitations and weaknesses, and the proposed method is no exception. First, relying purely on user generated feedback makes a managerial decision susceptible to biases and “extra loud” voices that tend to dominate and drive discussions on social media. The proposed method should never be considered a replacement of traditional methods, but a potential complement used in conjunction with traditional methods. Employing multiple methods reflecting the perspectives of various stakeholders

allows managerial decision makers to cross validate these perspectives and to reach a more comprehensive and less biased view of the market.

Second, the proposed method relies on customer comments, which come after the target product is available to the customers, and is therefore not applicable during the initial design and development of the product. Manufacturers still need to rely on extensive market testing with traditional methods early in the design and development of a product to avoid making costly mistakes. Nevertheless, the proposed method may provide valuable insights for manufacturers in their decision-making regarding continuous quality improvement and new model development. For instance, for the vast majority of the vehicles in the Chinese auto market, the manufacturers introduce new models every year and can benefit from a good understanding of the competitive advantages and disadvantages of their products relative to the major competitors in the market.

While we have applied the proposed method in a case study on cars, the method itself is general and can be applied to other products and services where sufficient UGC reflecting customer attitude is available for analysis. Nowadays, UGC is becoming increasingly available online for a wide range of products and services, providing essential inputs for our method. Of course, some of the components of the method may need to be adjusted or extended depending on the data available in the particular context. For example, in our case study, the two websites host the so-called “word-of-mouth” areas, which guide customers to provide comments regarding various aspects separately, allowing direct assessment of customer attitudes toward those aspects. In some domains, however, only totally unstructured comments are available. In that case, more sophisticated techniques for aspect-specific sentiment analysis (e.g., [66–73]) become necessary.

Some of the steps in the proposed method may be simplified or even eliminated in particular applications. For example, while our case study demonstrated the value of the competing product identification module beyond the top competitors recommended by the two websites, in some other contexts, the manufacturers, marketers, or experts may have perfect knowledge of the major rivals of a target product. In that case, the competing product identification step can be skipped and substituted with existing lists of competitors. Also, while the competing product identification module purely relies on customer feedback and reflects customer perception without any limiting criteria, in some contexts, the users may want to narrow the range of competitors based on some product characteristics (e.g., class, price range, and size).

The proposed method uses a variety of machine learning techniques to automate most steps and reduce the amount of manual effort needed to analyze social media UGC. However, manual effort in preparing training and testing data for these techniques (i.e., seed sentiment words and testing sentences for domain-specific sentiment lexicon generation, and training and testing examples for comparative UGC text classification) is still unavoidable. The benefit of using machine learning is that the learned models can be applied on future data automatically, and hopefully the one-time manual effort will pay off in the long run.

6. Conclusion and future work

Product performance analysis from social media has drawn considerable research attention. However, there is limited sense in analyzing the performance of a single product in isolation, as done in previous research. In an open market, the real performance of a product should be evaluated in comparison with its competitors. In this paper, we have proposed a novel method for analyzing competitive advantages (and disadvantages) of a target product relative to its competitors from the perspective of customers based on UGC from social media. Our method comprises a competitor identifier based on comparative text identification, a domain-specific sentiment lexicon learner to more accurately measure customer attitude, and a set of metrics for measuring product performance. A case study in the auto industry

demonstrates the utility of our method and reveals some interesting findings different from previous studies.

Our work contributes to both research and practice. For practitioners, we provide an effective tool for analyzing the competitive advantages (and disadvantages) of their products and identifying strong competitors of their products. It can provide managers with insights from the perspective of customers, which are especially important in a customer-driven market. The success of our case study indicates that it seems promising for manufacturers to actually implement and use our method in practice.

For research, our work highlights the importance of considering competition in product performance analysis. This consideration may be applicable to performance analysis regarding other types of entities (e.g., individuals and organizations) in other contexts, where competition exists and matters, and hence adapted to other research areas. Some of the techniques in our method, such as the competitor identifier and the metrics for customer attitude, may have broader applicability and can be adapted and tested in other contexts, such as customer satisfaction analysis and customer perception measurement.

As part of future work on this topic, several interesting extensions of our work can be explored. First, we only tested the proposed method in a case study in the auto industry. Future research may apply the method to other domains (e.g., cellular phones) to validate its generalizability. Second, we used some particular methods for feature selection, classification, and sentiment analysis in the case study. There are numerous alternative methods developed in the machine learning field, which may be tested in future studies. For example, besides filter methods for feature selection, such as CFS, which select the same feature subset without regard to the target classification method, there are also computationally-intensive wrapper methods, which apply the target classification method on training and testing data to select the best-performing feature subset for that particular classification method. Third, we proposed a supervised learning approach to comparative UGC text identification, avoiding the knowledge acquisition bottleneck and limited generalizability of rule-based methods. Future research may explore a hybrid approach, integrating machine learning and general rules.

Funding

This work was supported by National Natural Science Foundation of China (Grant No. 71731005, Grant No. 71571059, and Grant No. 71771077).

References

- [1] W. He, H. Wu, G. Yan, V. Akula, J. Shen, A novel social media competitive analytics framework with sentiment benchmarks [J], *Information & Management* 52 (7) (2015) 801–812.
- [2] J. Zhan, H.T. Loh, Y. Liu, Gather customer concerns from online product reviews—a text summarization approach [J], *Expert Systems with Applications* 36 (2) (2009) 2107–2115.
- [3] H.L. Chang, Y.C. Chou, D.Y. Wu, S.C. Wu, Will firm's marketing efforts on owned social media payoff? A quasi-experimental analysis of tourism products [J], *Decision Support Systems* 107 (2018) 13–25.
- [4] D.E. Pournarakis, D.N. Sotiropoulos, G.M. Giaglis, A computational model for mining consumer perceptions in social media [J], *Decision Support Systems* 93 (2017) 98–110.
- [5] H. Zhang, H. Rao, J. Feng, Product innovation based on online review data mining: a case study of Huawei phones [J], *Electronic Commerce Research* 18 (1) (2018) 3–22.
- [6] Y.M. Li, H.M. Chen, J.H. Liou, L.F. Lin, Creating social intelligence for product portfolio design [J], *Decision Support Systems* 66 (2014) 123–134.
- [7] J. Gallagher, S. Ransbotham, Social media and customer dialog management at Starbucks [J], *MIS Quarterly Executive* 9 (4) (2010).
- [8] T.S.H. Teo, W.Y. Choo, Assessing the impact of using the Internet for competitive intelligence [J], *Information & Management* 39 (1) (2001) 67–83.
- [9] W. He, X. Tian, Y. Chen, D. Chong, Actionable social media competitive analytics for understanding customer experiences [J], *Journal of Computer Information Systems* 56 (2) (2016) 145–155.
- [10] S. Nazari-Shirkouhi, A. Keramati, Modeling customer satisfaction with new product

- design using a flexible fuzzy regression-data envelopment analysis algorithm [J], *Applied Mathematical Modelling* 50 (2017) 755–771.
- [11] W.G. Kim, H. Lim, R.A. Brymer, The effectiveness of managing social media on hotel performance [J], *International Journal of Hospitality Management* 44 (2015) 165–171.
 - [12] Y.M. Li, T.Y. Li, Deriving market intelligence from microblogs [J], *Decision Support Systems* 55 (1) (2013) 206–217.
 - [13] T. Wang, P. Ji, Understanding customer needs through quantitative analysis of Kano's model [J], *International Journal of Quality & Reliability Management* 27 (2) (2010) 173–184.
 - [14] W. He, J. Shen, X. Tian, Y. Li, V. Akula, G. Yan, R. Tao, Gaining competitive intelligence from social media data: evidence from two largest retail chains in the world [J], *Industrial Management & Data Systems* 115 (9) (2015) 1622–1636.
 - [15] Y. Li, B. Jia, Y. Guo, X. Chen, Mining user reviews for mobile app comparisons [J], *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1 (3) (2017) 75.
 - [16] J. Jin, P. Ji, R. Gu, Identifying comparative customer requirements from product online reviews for competitor analysis [J], *Engineering Applications of Artificial Intelligence* 49 (2016) 61–73.
 - [17] A.A. Taleizadeh, R. Sadeghi, Pricing strategies in the competitive reverse supply chains with traditional and e-channels: a game theoretic approach [J], *International Journal of Production Economics* 215 (2018) 48–60.
 - [18] N.S. Davcik, P. Sharma, Marketing resources, performance, and competitive advantage: a review and future research directions [J], *Journal of Business Research* 69 (12) (2016) 5547–5552.
 - [19] X.L. Shen, S. Ma, Y. Sun, Understanding users' acceptance of competitive social media: a relative model [C], *Pacific Asia Conference on Information Systems*, 2015.
 - [20] S. Zhou, Z. Qiao, Q. Du, G. Wang, W. Fan, X. Yan, Measuring customer agility from online reviews using big data text analytics [J], *Journal of Management Information Systems* 35 (2) (2018) 510–539.
 - [21] Y. Liu, C. Jiang, H. Zhao, Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums [J], *Decision Support Systems* 105 (2018) 1–12.
 - [22] W. Dong, S. Liao, Z. Zhang, Leveraging financial social media data for corporate fraud detection [J], *Journal of Management Information Systems* 35 (2) (2018) 461–487.
 - [23] D. Zhu, T. Lappas, J. Zhang, Unsupervised tip-mining from customer reviews [J], *Decision Support Systems* 107 (2018) 116–124.
 - [24] Y.S. Peng, I.C. Liang, A dynamic framework for competitor identification: a neglecting role of dominant design [J], *Journal of Business Research* 69 (5) (2016) 1898–1903.
 - [25] I. Mohammed, B.D. Guillet, R. Law, Competitor set identification in the hotel industry: a case study of a full-service hotel in Hong Kong [J], *International Journal of Hospitality Management* 39 (2014) 29–40.
 - [26] Z. Zhang, C. Guo, P. Goes, Product comparison networks for competitive analysis of online word-of-mouth [J], *ACM Transactions on Management Information Systems* 3 (4) (2013) 20.
 - [27] K. Chen, P. Luo, H. Wang, An influence framework on product word-of-mouth (WoM) measurement [J], *Information & Management* 54 (2) (2017) 228–240.
 - [28] D.H. Park, C. Blake, Identifying comparative claim sentences in full-text scientific articles [C], *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, Association for Computational Linguistics, 2012, pp. 1–9.
 - [29] N. Jindal, B. Liu, Identifying comparative sentences in text documents [C], *Proceedings of the 29th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2006, pp. 244–251.
 - [30] S. Xiao, C.P. Wei, M. Dong, Crowd intelligence: analyzing online product reviews for preference measurement [J], *Information & Management* 53 (2) (2016) 169–182.
 - [31] G. Wang, J. Sun, J. Ma, K. Xu, J. Gu, Sentiment classification: the contribution of ensemble learning [J], *Decision Support Systems* 57 (2014) 77–93.
 - [32] S. Deng, A.P. Sinha, H. Zhao, Adapting sentiment lexicons to domain-specific social media texts [J], *Decision Support Systems* 94 (2017) 65–76.
 - [33] B.H. Clark, D.B. Montgomery, Managerial identification of competitors [J], *Journal of Marketing* 63 (3) (1999) 67–83.
 - [34] M.J. Chen, Competitor analysis and interfirm rivalry: toward a theoretical integration [J], *Academy of Management Review* 21 (1) (1996) 100–134.
 - [35] M.A. Peteraf, M.E. Bergen, Scanning dynamic competitive landscapes: a market-based and resource-based framework [J], *Strategic Management Journal* 24 (10) (2003) 1027–1041.
 - [36] M. Farhadloo, R.A. Patterson, E. Rolland, Modeling customer satisfaction from unstructured data using a Bayesian approach [J], *Decision Support Systems* 90 (2016) 1–11.
 - [37] H. Li, J. Cui, B. Shen, J. Ma, An intelligent movie recommendation system through group-level sentiment analysis in microblogs [J], *Neurocomputing* 210 (C) (2016) 164–173.
 - [38] R.Y.K. Lau, W. Zhang, W. Xu, Parallel aspect-oriented sentiment analysis for sales forecasting with big data [J], *Production & Operations Management* 27 (10) (2017) 1775–1794.
 - [39] N. Kano, Attractive quality and must-be quality [J], *Hinshitsu (Quality, The Journal of Japanese Society for Quality Control)* 14 (1984) 39–48.
 - [40] Y. Liu, C. Jiang, Y. Ding, Z. Wang, X. Lv, J. Wang, Identifying helpful quality-related reviews from social media based on attractive quality theory [J], *Total Quality Management & Business Excellence* (2017).
 - [41] M. Ghiassi, D. Zimbra, S. Lee, Targeted Twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks [J], *Journal of Management Information Systems* 33 (4) (2016) 1034–1058.
 - [42] M. Ghiassi, J. Skinner, D. Zimbra, Twitter brand sentiment analysis: a hybrid system using n-gram analysis and dynamic artificial neural network [J], *Expert Systems with Applications* 40 (16) (2013) 6266–6282.
 - [43] H. Kang, S.J. Yoo, D. Han, Senti-lexicon and improved naïve Bayes algorithms for sentiment analysis of restaurant reviews [J], *Expert Systems with Applications* 39 (5) (2012) 6000–6010.
 - [44] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches [J], *Expert Systems with Applications* 36 (3) (2009) 6527–6535.
 - [45] A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks [J], *IEEE Transactions on Knowledge and Data Engineering* 23 (3) (2011) 447–462.
 - [46] N. Oliveira, P. Cortez, N. Areal, Stock market sentiment lexicon acquisition using microblogging data and statistical measures [J], *Decision Support Systems* 85 (2016) 62–73.
 - [47] S. Huang, Z. Niu, C. Shi, Automatic construction of domain-specific sentiment lexicon based on constrained label propagation [J], *Knowledge-Based Systems* 56 (2014) 191–200.
 - [48] C. Jiang, Y. Liu, Y. Ding, K. Liang, R. Duan, Capturing helpful reviews from social media for product quality improvement: a multi-class classification approach [J], *International Journal of Production Research* 55 (12) (2017) 3528–3541.
 - [49] M. Thelwall, K. Buckley, G. Paltridge, D. Cai, A. Kappas, Sentiment strength detection in short informal text [J], *Journal of the American Society for Information Science and Technology* 61 (12) (2010) 2544–2558.
 - [50] X. Yang, G. Yang, J. Wu, Integrating rich and heterogeneous information to design a ranking system for multiple products [J], *Decision Support Systems* 84 (2016) 117–133.
 - [51] S. Park, W. Lee, I.C. Moon, Efficient extraction of domain specific sentiment lexicon with active learning [J], *Pattern Recognition Letters* 56 (2015) 38–44.
 - [52] Y. Liu, J. Jin, P. Ji, J.A. Harding, R.Y.K. Fung, Identifying helpful online reviews: a product designer's perspective [J], *Computer-Aided Design* 45 (2) (2013) 180–194.
 - [53] D. Zhang, L. Zhou, J.L. Kehoe, I.Y. Kilic, What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews [J], *Journal of Management Information Systems* 33 (2) (2016) 456–481.
 - [54] M. Winkler, A.S. Abrahams, R. Gruss, J.P. Ehsani, Toy safety surveillance from online reviews [J], *Decision Support Systems* 90 (2016) 23–32.
 - [55] 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 [J], *Decision Support Systems* 56 (2013) 211–222.
 - [56] A.S. Abrahams, W. Fan, G.A. Wang, Z.J. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery [J], *Production and Operations Management* 24 (6) (2015) 975–990.
 - [57] A.S. Abrahams, J. Jiao, G.A. Wang, W. Fan, Vehicle defect discovery from social media [J], *Decision Support Systems* 54 (1) (2012) 87–97.
 - [58] C. Oh, O. Sheng, Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement [C], *ICIS*, 2011, pp. 1–19.
 - [59] D. Yin, S.D. Bond, H. Zhang, Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews [J], *MIS Quarterly* 38 (2) (2013) 539–560.
 - [60] Q. Cao, W. Duan, Q. Gan, Exploring determinants of voting for the “helpfulness” of online user reviews: a text mining approach [J], *Decision Support Systems* 50 (2) (2011) 511–521.
 - [61] S. Krishnamoorthy, Linguistic features for review helpfulness prediction [J], *Expert Systems with Applications* 42 (7) (2015) 3751–3759.
 - [62] H. Hong, D. Xu, G.A. Wang, W. Fan, Understanding the determinants of online review helpfulness: a meta-analytic investigation [J], *Decision Support Systems* 102 (2017) 1–11.
 - [63] S.P. Eslami, M. Ghasemaghaei, K. Hassanein, Which online reviews do consumers find most helpful? A multi-method investigation [J], *Decision Support Systems* 113 (2018) 32–42.
 - [64] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning [C], *Proceedings of the 17th International Conference on Machine Learning*, 2000.
 - [65] M.J. Bitner, V.A. Zeithaml, D.D. Gremler, Technology's impact on the gaps model of service quality [J], *Service Science Research & Innovations in the Service Economy* (2010) 197–218.
 - [66] F. Xianghua, L. Guo, G. Yanyan, W. Zhiqiang, Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon [J], *Knowledge-Based Systems* 37 (2013) 186–195.
 - [67] R.K. Amplayo, S. Lee, M. Song, Incorporating product description to sentiment topic models for improved aspect-based sentiment analysis [J], *Information Sciences* 454–455 (2018) 200–215.
 - [68] B. Jeong, J. Yoon, J.M. Lee, Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis [J], *International Journal of Information Management* (2017).
 - [69] Y. Zhao, B. Qin, T. Liu, D. Tang, Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog [J], *Multimedia Tools and Applications* 75 (15) (2016) 8843–8860.
 - [70] N. O'Hare, M. Davy, A. Birmingham, P. Ferguson, P. Sheridan, C. Gurrin, A.F. Smeaton, Topic-dependent sentiment analysis of financial blogs [C], *International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, ACM, 2009, pp. 9–16.
 - [71] V. Parkhe, B. Biswas, Aspect based sentiment analysis of movie reviews: finding the polarity directing aspects [C], *International Conference on Soft Computing & Machine Intelligence*, IEEE, 2015, pp. 28–32.
 - [72] R. Wang, W. Huang, W. Chen, T. Wang, K. Lei, ASEM: mining aspects and sentiment

of events from microblog [C], Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 1923–1926.

- [73] A. García-Pablos, M. Cuadros, G. Rigau, W2VLDA: almost unsupervised system for aspect based sentiment analysis [J], Expert Systems with Applications 91 (2017) 127–137.

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

Cuiqing Jiang is a Professor at School of Management, Hefei University of Technology. 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).