Research Article

The Effects of Online Information on E-Book Pricing Strategies: A Text Analytics Approach

Kang Li ¹, Lunchuan Zhang ¹, Dianwei Wang,² and Dinglu Pan³

¹School of Mathematics, Renmin University of China, Beijing 100872, China

²School of Information, Renmin University of China, Beijing 100872, China

³School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China

Correspondence should be addressed to Lunchuan Zhang; zhanglc@ruc.edu.cn

Received 19 April 2021; Revised 30 May 2021; Accepted 25 June 2021; Published 6 July 2021

Academic Editor: Giuseppe D'Aniello

Copyright © 2021 Kang Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nowadays, electronic book vendors are increasingly proactive in trying to strategically capitalize on online big data generated by consumers. It will bring great profit for vendors if they can make the most of online reviews to figure out the impact of online data on the selling prices of e-books. In this paper, we complement an emerging body of research to explore how e-book prices could be affected by online information via analyzing the sheer volume of online data from e-book websites, namely, we first employ a domain ontology-based method to select the most discriminative features that may affect e-book prices. Then, the topic modeling method latent Dirichlet allocation and aspect-oriented sentiment analysis methods are applied as a supplement. Using the multiple regression method, we identify the key features that may have effects on the prices of e-books and give the related regression equation. In our results, some factors including paper book prices, paper book pages corresponding to the e-book, and e-book content have significant effects on the price of e-books. The managerial implication is that e-book firms can obtain a reference price for an e-book and may dynamically adjust the price to increase e-book sales according to our data analysis results.

1. Introduction

An electronic book (or e-book) is an electronic version book publication which is available on the screen display of mobile phones, computers, or other electronic devices. In the 2000s, with the growth of e-commerce, there was a trend of print and e-book sales going to the Internet, where customers could buy traditional paper books and e-books on e-commerce websites. Compared to traditional paper books, electronic books have many significant advantages, such as lower prices, increased comfort, and a larger selection of titles. E-books are rapidly gaining popularity around the world in recent years. At the start of 2012 in China, there were more e-books published online than distributed in hardcover. The amount of e-book reading is growing rapidly in China. Data showed that 23% of adults had ever read an e-book in 2013, while this number grew to 28% by the end of 2014, increased by five percentage points. Besides, there is also a significant increase in the use of e-book hardware

devices such as Kindle produced by Amazon. In 2013, only 30% of Chinese adults had an e-reader or a tablet, while this number reached 50% in 2014.

With the rapid development of electronic book publishing and use [1], a lot of practitioners and academic researchers pay more attention to the research related to e-books [2]. The pricing strategies of digital rights are a big challenge in many industries if the transition is made from physical goods to digital products [3]. An electronic book can be seen as a special kind of information goods or services. Information goods differ a lot from physical goods in many aspects. For example, the marginal production cost of the information good is almost equal to 0 [4]. Thus, there is a big difference between e-books and traditional paper books commercially. The research of pricing strategies is becoming a difficult issue for e-book vendors.

Electronic book manufacturers usually adopt these two pricing strategies for their products. The e-books on Amazon can sell at a certain price for perpetual ownership to

customers. Besides, Amazon also launched a special service called Kindle Unlimited, which can be regarded as a kind of leasing model. Kindle Unlimited is a new service that allows users to explore over 1 million e-books and thousands of audiobooks on any device as long as they pay \$9.99 per month. Our research mainly focuses on exploring the first pricing model; that is, consumers can buy e-books at a certain price to obtain permanent reading rights of the book, and the commercial pricing strategy of Kindle Unlimited is not considered in this paper. An interesting business problem for the merchants is how to determine an optimal selling price for each e-book and how to dynamically adjust prices in a series of promotional strategies to increase merchants' profits.

In recent years, there has been a large number of related research in exploring the influences of online information on customers' purchase behaviors. In many e-commerce platforms, there are comment channels to allow customers to give product reviews and rate the products [5]. In this way, it may induce important online information: online word-of-mouth (WOM) effects, helping consumers learn more information about the product. Chevalier and Mayzlin [6] explore the influence of consumer reviews on relative sales of paper version books, and the results suggest that WOM on the Internet have an important impact on customer behavior. Hu et al. [7] suggested that consumers should make the best of online WOM of books for making purchases. On e-book websites, there are also lots of online information which readers may consider, such as related introduction to books, customer reviews, and celebrity recommendation.

For e-books vendors, the pricing strategy differs a lot from traditional paper books. A very significant reason is that e-books, as a kind of product of the Internet era, will have a large amount of online consumer review information. Previous research has shown that products' online information, especially consumer reviews, has a very significant impact on the consumer decision-making process. As a result, the previous research on traditional paper books pricing models is not necessarily applicable to e-books. The online information of e-books will influence consumers' buying decisions largely. Customer review is displayed to all potential paying readers on the e-book website and plays a vital role in the product. Research shows that the vast majority of consumers refer to product reviews during their buying decision process. Good reviews will increase consumers' confidence in buying the product, and bad reviews will make buyers feel disappointed and abandon the product. In addition, consumer review information also affects the e-book's best-selling rankings. Many consumers also treat the best-selling rankings of e-books as a very important reference factor before purchasing. Therefore, the online information of e-books plays a significant role in the purchase intention of consumers and may affect e-book sales. In traditional economic theory, the price of a commodity is determined by the supply and demand in the market. In this way, the e-books' online information may indirectly have effects on the corresponding selling price. If e-book merchants can fully mine the online information and generate the corresponding price strategies, they will

undoubtedly increase the sales of e-books and obtain much greater profits. Thus, research exploring the effects of e-book online data on its selling prices has important management and commercial implications for e-book manufacturers.

An interesting problem is that how e-book vendors mine useful knowledge from their vast online information for their pricing management decisions. More specifically, how to adjust the price of e-books based on massive online data to achieve greater benefits becomes an attractive research question. There are a great number of big data analysis methods applied to the research of e-commerce. Literature in economics and marketing suggested that consumers depend on online product reviews to make purchase decisions [8, 9]. Many researchers provide useful insights into the relationship between online reviews and sales. They showed a positive correlation between the average review ratings and product sales [10–12] or between the volume of reviews and sales [13, 14]. Most of the existing studies explore the impact of online reviews on consumer purchasing behavior or the impact of online reviews on product sales and vendor's profits. There are very few studies to explore the enlightenment of the online information of e-books on the pricing strategies of merchants. This research question has some commercial significance to e-book merchants. Therefore, our research is carried out along such a new path. In addition, most of the previous studies mainly focused on one of the traditional attributes or social attributes. Our research combines the traditional attributes with social attributes of e-books and applies domain ontology and Text Mining methods to integrate numerical data with text-based data in one multiple linear regression model, which is also a supplement to the existing research.

Our research mainly explores the influence of e-book online information on the pricing strategies of e-book merchants. We convert textual data into numerical data by constructing domain ontology and text aspect-oriented sentiment analysis methods, and finally, reveal the main determinants of e-book price through multiple regression models. According to the source of online data of e-book websites, the e-book data can be divided into two categories: traditional attributes and social attributes. Traditional attributes refer to the inherent attributes of e-books themselves as a reading product, such as book prices and book pages. The social attribute mainly refers to a series of text information generated by the consumers' behaviors on the Internet, such as e-book massive consumer review information. From the perspective of data type, the online data of the e-book can be divided into two data types: text data and numerical data. For all the data of e-books, we employ the domain ontology-based method to select key features that may affect e-book prices. In order to mine text information of e-book, latent Dirichlet allocation (LDA) method and aspect-based sentiment analysis are employed in our research. In this way, we convert each aspect of text information into sentiment scores. Then, we use the multiple regression method to combine these two parts of the results into consideration, and we got the final result of the main features that may have effects on the selling prices of e-books.

Our research results can provide a good reference for the pricing strategies of e-book manufacturers commercially. E-book vendors can mine the main factors that may have greater impacts on the selling price through a large amount of online review information for e-books. When applied to business practice, they can identify those attributes that will have significant effects on the price of e-books and then adjust the selling price of e-books accordingly in advance to obtain greater profits.

The remainder of the paper is organized as follows. A brief introduction to previous studies is given in Section 2. Section 3 explains the proposed combined domain ontology and text analytics methodology for e-book pricing. Section 4 illustrates the computational details of the proposed methodology. The experimental procedures and the empirical results are also discussed in this part. Finally, the conclusions and pinpoint future directions of our research work are summarized in Section 5.

2. Literature Review

In our study, we draw mainly on three streams of research in IS economics and marketing: (i) the impact of product prices, (ii) the commercial role of product online information, and (iii) the relationship between online data and prices.

2.1. The Impact of Product Prices. The first stream of literature relevant to our work is related to the research of the impact of product prices in commerce. Previous research has shown that product price has been proven to be an important determinant of customer satisfaction in the manufacturing industry [15], and there are also similar results that have been proven in other industries such as rental cars [16]. The analytical model further examined the impact of price-influenced reviews on firm optimal pricing and consumer welfare [17]. Besides, according to the existing research, when consumers are uncertain about the product information, they tend to regard price as a signal of its quality [18–21]. According to these studies, price plays an important role in consumer choice. Considering these factors, our study investigates how to determine an optimal e-book price based on online information on the websites.

2.2. Commercial Role of Product Online Information. The second stream of extant work relevant to our research is the commercial role of online data. As one of the most important online data, many studies explored the commercial value of customer review, which is also called word-of-mouth effects. Since Rogers, word-of-mouth has been perceived as an important driver of sales in the product diffusion literature [22]. Those models normally assume that consumers' experience with a product is communicated positively through word-of-mouth and, therefore, facilitates product diffusion [23]. With the emergence of large-scale online communication networks and the increase of customers' willingness to evaluate the products online, researchers can observe word-of-mouth over time and,

therefore, obtain a deeper understanding of consumer preferences and decision processes. Recent studies have attached great importance to online customer reviews, which is one of the hottest issues in electronic commerce research. Based on product reviews online or conversation data collected from consumer networks, many researchers can effectively explore the relationship between word-of-mouth and product sales. In the book industry, it has been demonstrated that the differences between consumer reviews posted on the Barnes & Noble site and those posted on Amazon.com were positively related to the differences in book sales on the two websites [6]. Two other studies further found that reviews from the same geographic location [24] or with a higher helpfulness vote [25] have a higher impact on product sales. There are also some researchers who test the relationship between reviews written by customers and customer ratings. Gan and Yu [26] used reviews and ratings to build a multilevel model for showing the sentiment impact on different levels. Chen and Xu [27] studied the determinants of online customer ratings by analyzing customer review data using a combined domain ontology and topic text analytics approach.

However, to our knowledge, research on the relationship between product online data and price is largely missing. We consider this an interesting issue in e-commerce by applying a mixed model combined with domain ontology and topic text analytic in this paper. In particular, our research innovatively tests the effect of online information on e-book prices, which is new compared to previous related work.

2.3. The Relationship between Online Data and Prices. The third stream of extant work relevant to our research is on the relationship between online data and prices. Some related studies consider product reviews and prices together. In most of these existing studies, researchers investigated how product prices have a significant impact on customer reviews and what firms should do to respond to these price effects in reviews. In most above existing literature in the second stream of related work we mentioned, it is implicitly assumed that consumer reviews reflect consumers' perceptions of product quality, and hence, consumers are not affected by the price of products. However, whether this assumption is consistent with consumer behavior is left untested. Bolton and Drew [28] suggested that a buyer will make a trade-off between benefit (quality) and cost (price) before their purchase. Although a lot of people agreed that consumers' purchase decision is determined by the expected utility before purchase, perceived value can also affect the consumer's ex post satisfaction with the purchase [29]. In this way, price should be an important factor influencing consumers' postpurchase satisfaction and, in turn, will influence consumer reviews [30, 31]. Richins [32] also suggested that word-of-mouth may be driven by product price in a purchase. Li and Hitt [17] developed an analytical model that examines the impact of price-influenced reviews on firm optimal pricing and consumer welfare.

The related work mainly discusses the potential impact of price on customers' postpurchase perceptions and

satisfaction and thus influences product reviews. But nearly no studies explore whether customer reviews can affect the price of products. Our research fills this interesting gap and furthermore, the managerial implications for e-book firms are discussed.

3. Research Methodology

In this section, the methodology for e-book price determinants research, which is applied to construct the proposed combined domain ontology and text analytics framework, is briefly described. More specifically, for traditional attributes, we employ the domain ontology-based method to select the most discriminative features that affect e-book prices. For social attributes, which mainly refer to online text information, the topic modeling method latent Dirichlet allocation is applied at first to select feature set, and then we use the aspect-oriented sentiment analysis method to convert text information into numeric data. Using the multiple regression analysis models, we identify the key features that may influence the prices of e-books and give the related regression equation. These analysis methods underpin the development of an effectively combined domain ontology and text analytics framework for studying the effect of online data on e-book price. The overview of the proposed framework is shown in Figure 1.

As described in Figure 1, we crawl e-book prices and e-books attributes consisting of fundamental attributes and social attributes from Amazon websites firstly. Secondly, to capture the professional e-book concepts from traditional attributes, we constructed the e-book ontology. Thirdly, data mining methods are applied to select useful features from traditional attributes and explore the influential impact of these features. Fourthly, for social attributes, which mainly refers to customer reviews, the Latent Dirichlet Allocation method is employed to construct the aspect-related review feature sets, and then, we apply the sentiment analysis method to estimate the sentiment score for each feature obtained by the text mining approach. Finally, we use the multiple regression method, which combines the analysis of the two parts into one model to reveal the main determinants of e-book prices and then the results are discussed. We will discuss the model details in the following subsections.

3.1. Data Collection. The e-book data used in this paper are based on the Amazon website, and we obtain these data by crawling. This is because Amazon, as a leader in the e-book industry, has a massive online e-book resource library and the largest user group. In the e-book industry, Amazon has the largest market share in the world, and both its software and hardware products have the best reputation in the industry among consumers.

This study chooses the data from the best-selling list of electronic books on Amazon.com, including the traditional attributes and social attributes of e-books. The traditional attributes mainly refer to the domain ontologies of the e-book, while the social attributes refer to the online review

information of book readers. In terms of the types of data, our data includes numeric data and text data.

3.2. Traditional Feature Extraction. For traditional attributes, we employ the domain ontology-based method to select the most important features that may have effects on e-book prices.

An ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really exist in a particular domain of discourse. The concept of ontology came from the field of philosophy; ontology is often represented by a hierarchy of concepts and some other semantic information. According to the related work, domain ontology focuses on the concepts and their relationships to a particular domain [33]. Besides, domain ontology represents the axioms (e.g., rules) and constraints that define the foremost features of the domain [34–36]. An ontology compartmentalizes the variables needed for some set of computations and establishes the relationships between them. It is a formal and generic way to represent a set of related concepts of a domain so that different people can reuse and apply this domain knowledge. Ontology is popular in describing domain knowledge due to its distinct advantage of promoting reusability.

In this paper, we identified the semantic relationships among different concepts of the domain of books first. Then, a causal map that can be used as a basis to explain different events (e.g., increasing or decreasing e-book prices) pertaining to the book market is constructed and represented in the form of a semantically rich domain ontology. Finally, the domain ontology is applied to select useful features for e-book price determinants. The proposed e-book price ontology is built and refined using well-known and effective knowledge engineering tools such as Protégé [37]. Book concepts are extracted based on some well-known professional book websites (e.g., douban.com) which are summarized by domain experts.

There are several well-known and effective knowledge engineering tools which can be applied to construct the proposed domain ontology. In our work, Protégé [37], a java-based open-source ontology editor, was adopted as the knowledge engineering tool to construct our domain ontology. We elicit product concepts based on one of the most famous professional book review websites (douban.com), which are written mainly by senior readers who can be regarded as domain experts of this field. On the ontology refinement stage, the Protégé reasoner, which is an external plug-in of Protégé, is applied to check the integrity and the correctness of domain ontology. After completing the refinement step of ontology construction, some prominent concepts related to e-books are encoded in the proposed domain ontology. As a result, several features which may have a direct influence on e-book price are depicted.

3.3. Social Feature Extraction. In this part, the topic modeling method latent Dirichlet allocation (LDA) is applied to extract topics from e-book reviews which can be considered as a supplement to the previous domain ontology method.

FIGURE 1: A text mining and domain ontology framework for e-book price effects.

3.3.1. Data Preprocessing. Before the LDA model setup and further analysis, we should do the data preprocessing work for raw customer reviews. The procedure of data preprocessing consists of the following steps. The first step is Chinese word segmentation, namely, to segment a sequence of Chinese characters into short strings of words. Although the text data of e-book reviews appear in the form of long sentences in different structures, the significant information of the review data is usually distributed across several words in one sentence. Because of this, we should focus on a few keywords instead of the entire sentence. For our study, we use ICTCLAS [38] to complete Chinese word segmentation, which is a widely used Chinese language processing tool and can be implemented in different programming languages. The second step is stop words removal, that is, to remove the words without any specific meanings. For instance, the words like “the,” “of,” “on,” “at” in English are usually regarded as stop words. Stop words can be abandoned during the stage of sentiment analysis. The third step is part-of-speech (POS) tagging. In this way, the property of a certain word in sentences can be recognized. Nouns can usually stand for certain aspect descriptions. Adjectives and adverbs are mainly used for sentiment analysis. After Chinese word segmentation, stop word removal, and POS tagging, the remaining words are what we prepared ready for our following work.

3.3.2. LDA Model. In natural language processing, Latent Dirichlet allocation (LDA) is a generative statistical model that can automatically detect the latent topics among large-scale documents without any prior knowledge or any manual annotation. For our study, we apply the LDA model

to extract the most significant aspect that may directly affect the e-book price from the customer reviews, which can be regarded as another approach for aspect identification.

LDA is a document topic generation model, also known as a three-layer Bayesian probability model, which contains a three-layer structure of words, topics, and documents. LDA is a typical the Bag of Words mode, which means that a document is a collection of a group of words and there is no order or precedence relationship between words. Each document can be viewed as a representation of a probability distribution of many topics, and each topic can be viewed as a representation of a probability distribution of a set of words. What's more, because of the noncorrelation between the components of the random vector in the Dirichlet distribution, the candidate topics are independent of each other [39]. The mathematical notation conventions applied to the topic modeling of the document set by the LDA model are as follows: (1) the v th word in the vocabulary is represented by a V -dimensional vector w , where for any $u \neq v$, $w_v = 1$, $w_u = 0$. (2) The document is a sequence of N words, denoted by $d = \{w_1, w_2, \dots, w_n\}$. (3) The document set is a collection of M documents, $D = \{d_1, d_2, \dots, d_M\}$. Assuming that there are k topics, the probability of the i th word w_i in document d can be expressed as follows:

$$P(w_i) = \sum_{j=1}^T P(w_i|z_i = j)P(z_i = j). \quad (1)$$

For convenience, let $\phi_w^j = P(w|z = j)$ be the multinomial distribution for word w in topic j and $\theta_j^d = P(z = j)$ represents the multinomial distribution for topic T in the

document. The probability of word w that occurs in the document d is as follows:

$$P(w|d) = \sum_{j=1}^T \phi_w^j \theta_j^d. \quad (2)$$

3.3.3. Text Feature Sets. For the proposed topic model in our paper, the Fine-grained Labeled LDA (FL-LDA) method [40] is applied to identify words that are related to e-book aspects. Some seeding words are given before the establishment of social feature sets first. Then, we amend the feature sets of social attributes of e-books by analysis of topics and words which is based on FL-LDA results. In this model, we employ a collapsed Gibbs sampling method [41] to compute the posterior distribution:

$$P(z|w) = \frac{P(w, z)}{\sum_z P(w, z)}. \quad (3)$$

The joint likelihood of a word w associated with an aspect z is given by the following:

$$P(w, z|\alpha, \beta, \gamma) = P(w|z, \beta, \gamma)P(z|\alpha, \gamma). \quad (4)$$

3.4. Sentiment Analysis. In order to determine the sentiment score for different features of each customer review, the proposed approach performs sentiment analysis at the clause level using an aspect-oriented sentiment analysis method. The approach makes use of both aspect-based sentence segmentation and a domain-specific lexicon to decide the sentiment polarities for the aspects we get in the former section. We will give the details in the following sections.

3.4.1. Aspect-Based Sentence Segmentation. For each review sentence in our data which contains multiple aspects, the first task for aspect-based sentiment analysis we should do is to split such a multispect sentence into multiple single-aspect units. Thus, before the sentiment score computing, we will divide a review sentence that contains multiple aspects into multiple segmentations to make sure that each sentence is only related to one single aspect. We first cleaned the raw reviews for the purpose that only the reviews that are related to some aspects are maintained. In this way, unrelated reviews are given away. For our study, to tackle the aspect-based sentence segmentation task, we applied the multi-aspect segmentation (MAS) model that takes a multispect sentence as input and produces multiple single-aspect segments [42].

In order to formulate a multispect segmentation model, we define a criterion function $J(\cdot)$ that aims to evaluate each candidate segmentation U of sentence C . The definition is given as follows:

$$U^* = \arg \max_U J(C, U). \quad (5)$$

Then, we introduce an indicator function $\delta(u_i, u_j)$ to represent whether two adjacent segments express the

meaning of the same aspect, the value of $\delta(u_i, u_j)$ is equal to 1 if two segments u_i and u_j are labeled as two different aspects, and 0, otherwise. That is,

$$\delta(u_i, u_j) = \begin{cases} 1 & \text{if } u_i \text{ and } u_j \text{ are labeled as two different aspects,} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The criterion function $J(C, U)$ can be expressed as follows:

$$\begin{aligned} J(C, U) &= \sum_{1 \leq i \leq k} [\delta(u_{i-1}, u_i) * \text{score}_{a^*}(u_i)] \\ &= \sum_{1 \leq i \leq k} \left[\delta(u_{i-1}, u_i) * \sum_{t \in u_i} \text{score}_{a^*}(t) \right], \end{aligned} \quad (7)$$

where the parameter a^* denotes the aspect in which segment u_i has the maximum probability of happening. The segments that are related to the same aspect will be regrouped together after the procedure of aspect-based sentence segmentation.

3.4.2. Aspect-Oriented Sentiment Analysis. For each review, we cluster the sentence segmentation related to the same aspect together. Then, we can analyze the sentiment polarity for each aspect, respectively. Before analyzing, the reviews do not mention any aspects are discarded, and the default sentiment polarity for the missing aspect is set as neutral. For each feature word in the aspect-related review, we can calculate the polarity. Finally, we acquire the comprehensive aspect sentiment score by accumulating polarity strength for each feature word.

We give the calculation formula of the sentiment score of each feature of the e-book. For each e-book, we assume that in the sentiment analysis polarity corresponding to the i th aspect, the number of sentiment polarity of positive (+1), neutral (0), and negative (−1) reviews in cleaned date sets are N_p^i , N_0^i , and N_n^i , respectively. Then, the sentiment score of the e-book related to this aspect S_i can be calculated as follows:

$$S_i = \frac{N_p^i - N_n^i}{N_p^i + N_0^i + N_n^i}. \quad (8)$$

3.5. Regression Model. The most discriminative features that may affect e-book prices have been selected in the previous sections. In this section, the relationships between electronic book prices and these key features are established based on multiple regression models. We would like to combine the traditional attributes with social attributes which we considered respectively in former sections into one regression model, aimed to capture the main features which may directly affect the e-book prices. We have calculated the sentiment score for each feature generated based on the Ontology method and LDA results, which are regarded as independent variables in the regression model. We treat the e-book prices as dependent variables. Finally, we establish the multiple

linear regression equation by means of a well-known statistics software SPSS.

4. Empirical Analysis

4.1. Data Description. The datasets in our empirical study are crawled from an industry-leading electronic book website Amazon.com, which is the world's largest online bookstore. We ultimately obtain 498 e-books' prices and their related data in the dataset, which are crawled from the best-selling list of e-books in Amazon. On the whole, our crawled data contain three types of data: text data, numerical data, and price data. Overall, our raw datasets consist of 498 e-book price data, 147,362 text data, and the related numerical data in total for the corresponding books. A summary of all the data is intuitively shown in Table 1.

There are a large number of text data we obtained, including e-book reviews and book description about the author celebrity recommendation. However, the number of electronic book text data we can use in our experiment is only a small portion. Because some customer reviews are too short to get the real meanings for our research, or some are only general descriptions for the book. For example, "A very good book, I like it very much," this review is just comprehensive description of the e-book, which do not mention anyone aspect sets we constructed. In this way, we clean the e-book review data in order to strike out the reviews which contained less than 8 words. Those reviews do not make any sense in our experiment results. In the aspect-level data cleaning process, we only maintain the reviews that mention at least the aspect we choose before. Accordingly, there are 26,104 texts left after the aspect-level data cleaning process for customer reviews.

4.2. Feature Extraction. In this section, we will extract the main features that may have a prominent effect on e-book prices. We first apply the domain ontology to facilitate the extraction of traditional features from well-known Chinese book evaluation websites which are summarized by domain experts. As Table 2 shows, there are 6 features extracted from the electronic book ontology. Then, as a supplement, we give some additional features by choosing some LDA topics. We select 4 features in total from online review texts for e-books in which two of them are supplements to the previous results obtained by a domain ontology. We show some representative topic words and the four corresponding features generated from LDA results in Table 3. Finally, by combining these features that we acquired together, we can give an accurate and comprehensive feature set that may affect the price of e-books.

As can be seen in Tables 2 and 3, comparing the features, we extract from domain ontology and LDA results. The LDA results have the following two features in common with the ontology results: e-book content and book packaging, while either of the two parts of results has some features which the other one does not have. There are some topic results which can be regarded as a general description of the electronic book. They just describe a reader's overall

impression of the e-book. Actually, these topics should also have an impact on e-book prices, but in this paper, we cannot obtain any useful information for some specific feature from the general description. As a result, we add the "logistics quality" and "biography" to our results obtained based on product ontology as the supplement, which are two unique aspects we acquire by the LDA model. In this way, we have the following eight dimensions of features that may have significant effects on the prices of electronic books: number of reviews, average ratings of reviews, paper book prices, paper book pages, e-book content, logistics quality, book packaging, and biography.

4.3. Aspect-Oriented Sentiment Analysis. In this section, we carry out aspect-oriented sentiment analysis for electronic books. For each review, we cluster the sentence segmentation related to the same aspect together. In this way, we can analyze the sentiment polarity for each aspect respectively. First, aspect-based sentence segmentation is a significant step of data processing. Next, the segmentation related to the same aspect can be clustered together. In order to explain this process clearly, we give a sample review from a popular book "In the name of people" as follows and show the process results in Table 4.

"In the beginning, I know the book from the TV series. The novel has many differences from the TV drama. The plot on TV was ups and downs, while the characters in the book portrayed vividly and the suspense was exciting. The author of the book, Zhou Meisen, born in 1956, was from Xuzhou, Jiangsu. He is the member of the Seventh, Eighth and Ninth National Committee of the Chinese Writers Association, member of the Presidium, and a professional writer. He has repeatedly won the National "Five One Project" Award, the National Book Award, the National Best-selling Book Award, and the National Outstanding Novelist Award. Li Jingze, a member of the Chinese Writers' Association and secretary of the Secretariat, once recommended this book as follows: literary creation does need to face the general trend of the country and the nation and draw strength from it. Therefore, the rise and fall of a theme work, to a certain extent, itself reflects the changes in the social age. What is called the main melody is high, what is called to effectively reflect the real life, "the people's name" reflects such a theme, it embodies the people-centered creative direction."

In this way, we have finished the aspect-related segmentation and regrouping process. We compute each feature word's sentiment polarity and sum these values up to obtain the aspect-related sentiment polarity for e-book features. In the last, we carry out the calculation of the sentiment score of each feature of the e-book.

4.4. Results' Analysis and Further Discussion

4.4.1. Multiple Linear Regression Results. We will establish a multiple linear regression model to determine the

TABLE 1: Summary of the data.

No.	Data	Description	Type of data
1	E-book reviews	Customer reviews for the e-book	Text data
2	Book description	A brief description of the book	Text data
3	About the author	A brief introduction to the author	Text data
4	Celebrity recommendation	Some celebrities' recommendation	Text data
5	Paper book prices	Corresponding paper book prices	Numerical data
6	Paper book pages	Corresponding paper edition pages	Numerical data
7	Number of reviews	The number of customer reviews	Numerical data
8	Review ratings	Average rating of customer reviews	Numerical data
9	E-book prices	The selling prices of the e-books	Numerical data

TABLE 2: The features which may have effects on e-book prices by ontology.

No.	E-book feature	Type of corresponding data
1	E-book content	Text data
2	Paper book prices	Numerical data
3	Paper book pages	Numerical data
4	Book packaging	Text data
5	Number of reviews	Numerical data
6	Average ratings of reviews	Numerical data

relationships between electronic book prices and the features we selected based on our methods. After the aspect-oriented sentiment analysis process in the last section, the sentiment polarities of the related aspects change to numerical values so that we can use them to execute regression analysis. Firstly, we regard e-book price as a dependent variable and the features we extracted using domain ontology approach as independent variables, which we labeled as regression model I. Secondly, we explore the relationship between the e-book price and aspect-related sentiment scores acquired based on text feature sets and change the independent variables into the sentiment scores of the related aspects we obtain through the text mining approach, which is model II. Thirdly, we treat the ontology feature sets and the text feature sets together as independent variables and conduct model III. In the last subsection, we will carry out a comparison of these three regression models.

4.4.2. Comparison of Three Regression Models. We show the results of the three multiple linear regression models in Table 5. In Table 5, 5(a), 5(b), and 5(c) represent the experimental results of the regression model I, model II, and model III, respectively. By comparing the results of these three models, it can be easily found that the experimental results of model III are the best. Therefore, we choose model III as our final multiple regression model, which best reflects the effects of online data on electronic book prices. Compared with model I, model III has taken two important text attributes-biography and logistics quality into consideration. While in model II, there are some significant traditional features missing in dependent variables compared with model III, such as paper book pages and paper book prices. Therefore, model III is the most accurate regression model and the result of this model is an enhancement to both previous two models. The feature sets in model III consist of

not only concepts based on the e-book ontology but also the aspects selected from LDA topics. The final multiple linear regression results of our experiment are shown in Table 5(c).

As can be seen in Table 5(c), among all eight factors, seven factors have significant effects on the price of e-books, while logistics quality does not have a very significant effect on e-book price because its significance level is greater than 0.05. These seven aspects all have a positive correlation with e-book prices. As the result shows, in terms of the impact on e-book prices, paper book prices and paper book pages corresponding to the e-book are the two most significant aspects. E-book content, average ratings of reviews, and the number of reviews locate the third place to fifth place, respectively. In comparison with these above aspects, biography and book packaging have the least impact on e-book price. In addition, the impact of logistics quality on e-book price is not significant.

4.4.3. Further Discussion. From the results of the multiple regression model, we can have access to see whether these aspects we selected can have an impact on electronic book price and their relative importance. Seven of the eight aspects we selected have a positive association with the e-book price. However, the attribute of logistics quality is irrelevant in terms of e-book price determinants.

In our results, paper book prices and paper book pages corresponding to the e-book are proven to be two important factors which electronic book merchants should consider in deciding the selling price. When a manufacturer sets the price of an e-book commercially, the price and number of pages of the paper version of this e-book could be the two most important reference factors. The price of paper books should be regarded as the most important reference pricing factor. In other words, for those e-books with a higher price of related paper books and more pages, manufacturers can consider setting a higher selling price when pricing them. Conversely, the price of e-books should be set lower.

The third important aspect is e-book content, which reflects the quality of e-books. In other words, if the content of the e-book is more exciting, the price will be correspondingly higher. With the wisdom of big data text mining, we can measure the influence of e-book content on its price importance. For authors, if they can write better content, their books will sell for a higher price. The reference for manufacturers is that for those higher quality e-books, a higher sales price can be set to earn more profit. The e-book's

TABLE 3: LDA results.

No.	Topic words' description	Corresponding feature
Topic 0	Speed, delivery, ship, logistics	Logistics quality
Topic 1	Text, delicate, idea, illustration	E-book content
Topic 2	Plot, characters, story, novel	E-book content
Topic 3	Cover, version, font, set of books	Book packaging
Topic 4	Paper, package, printing, genuine	Book packaging
Topic 5	Recommend, suggest, collect, a good book	General description
Topic 6	After reading, beautiful, satisfied, five-star review	General description
Topic 7	Emotion, narration, realize, principle	E-book content
Topic 8	Author, heart, life, experience	Biography
Topic 9	Chinese version, style, English version, background	E-book content

TABLE 4: Results of aspect-based segmentation.

Aspect	Regroup segmentations' results
E-book content	The plot on TV was ups and downs, while the characters in the book portrayed vividly, and the suspense was exciting. What is called the main melody is high, what is called to effectively reflect the real life, "the people's name" reflects such a theme. It embodies the people-centered creative direction.
Biography	He is a member of the seventh, eighth, and ninth national committee of the Chinese writers association, a member of the Presidium, and a professional writer. He has repeatedly won the national "five one project" award, the national book award, the national best-selling book award, and the national outstanding novelist award.

TABLE 5: The results of multiple linear regression models.

Model	Unstandardized coefficients		Standardized coefficients	<i>t</i>	Sig.
	<i>B</i>	Std. Error			
<i>(a) The results of regression model I</i>					
Constant	37.054	12.086		3.066	0.003
X1 (number of reviews)	0.004	0.003	0.107	1.420	0.158
X2 (average ratings of reviews)	7.113	2.807	0.199	2.535	0.013
X3 (paper book prices)	0.121	0.021	0.767	5.747	0.000
X4 (paper book pages)	0.007	0.002	0.558	4.202	0.000
X5 (e-book content)	28.758	14.131	0.155	2.035	0.044
X6 (book packaging)	1.594	13.666	0.009	0.117	0.907
<i>(b) The results of regression model II</i>					
Constant	48.638	13.789		3.527	0.001
X1 (number of reviews)	0.005	0.003	0.143	1.679	0.001
X2 (average ratings of reviews)	8.992	3.266	0.252	2.753	0.000
X3 (e-book content)	31.108	16.186	0.168	1.922	0.007
X4 (logistics quality)	13.183	52.108	0.022	0.253	0.501
X5 (book packaging)	2.028	15.823	0.011	0.128	0.298
X6 (biography)	8.499	17.225	0.045	0.493	0.012
<i>(c) The results of regression model III</i>					
Constant	37.783	12.443		3.036	0.003
X1 (number of reviews)	0.004	0.003	0.109	1.427	0.036
X2 (average ratings of reviews)	7.423	2.930	0.208	2.533	0.013
X3 (paper book prices)	0.122	0.021	0.775	5.722	0.000
X4 (paper book pages)	0.007	0.002	0.559	4.181	0.000
X5 (e-book content)	28.249	14.452	0.253	1.955	0.043
X6 (logistics quality)	40.934	39.504	0.143	2.368	0.219
X7 (book packaging)	4.726	17.415	0.017	0.271	0.006
X8 (biography)	16.739	18.077	0.053	0.834	0.042

review ratings and the number of reviews are the fourth and fifth most important attributes for pricing strategies, respectively. This mainly demonstrates the important impact

of the word-of-mouth effect on consumer purchases and merchant pricing. Besides, the effect of biography and book packaging on the experimental results is so weak that it can

be ignored in the pricing strategy. The attribute of logistics quality is not significant in our experiment results, so it cannot be regarded as a reference factor for e-book pricing. This reason may be that most of the logistics quality in customer reviews refers to the logistics and delivery of paper books rather than a unique attribute of e-books.

5. Conclusions and Future Work

Our paper proposes a semantic text analytics approach to mine the possible factors from online data which may have effects on the prices of electronic books. In the proposed method, ontology-based product concepts and selected LDA topics are combined together to constitute a comprehensive feature set. The sentiment score for each social feature is estimated by applying aspect-oriented sentiment analysis and thus, we convert all text data into numeric data. Eventually, the multiple regression model is established between the sentiment scores of text features, traditional numeric features of e-books, and the corresponding price. Empirical results reveal that we accurately distinguish the effects of selected features on prices of e-books, which has some realistic references for customers, e-book firms, and industries.

Our research results show that paper book price and the number of paper book pages corresponding to the e-book are the two most significant influencing factors affecting the price of e-books. Among the two, the paper book price has the greatest effect on the price of e-books, followed by the number of paper book pages. The third most important influencing factor that may have effects on the e-book price is the e-book content. This factor is the most essential feature of the e-book and reflects the quality of a book. The average ratings of consumer reviews and the number of reviews are the fourth and fifth most important attributes, respectively. The biography and book packaging are the sixth and seventh significant factors, respectively. These two attributes have relatively weak impacts on the selling price of e-books, which can be ignored when discussing pricing strategies. In addition, the aspect of logistics quality is not significant in the multiple regression equation, so it has little impact on the e-book price.

The results of this paper have certain business implications for the pricing strategies of e-book manufacturers and the e-book industry. When pricing e-books, the two key aspects of paper book prices and the number of paper book pages are two major attributes that merchants should pay the most attention to. The higher the price of the corresponding paper book, or the larger number of paper book pages, the corresponding e-book's selling price should also be set higher. What's more, the three factors of e-book content, e-book's average review ratings, and the number of reviews can also be referred to when considering e-book pricing, and the impact of these three factors on the price is also positively correlated. If the content of the e-book is more exciting, the average review ratings are higher, or the number of customer reviews is larger, the corresponding selling price of the e-book should also be designed higher. The order of the influence of these three factors is e-book content, e-book's average review ratings, and the number of reviews.

This paper is not without some limitations, which may be potential problems for further consideration. First of all, it is critically important that how to effectively construct better feature sets for each attribute. Secondly, our datasets are from only Amazon.com. However, there are many other e-book websites, such as Jingdong.com. Third, in our experiments, we use the same value (e.g., 1) for all of the positive sentiments without differentiating the strength of text sentiment, which may cause some errors in the final results. Future research could pay more attention to these problems we raised to enhance experiments' performance.

Data Availability

The e-book data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] D. Wightman, T. Ginn, and R. Vertegaal, "BendFlip: examining input techniques for electronic book readers with flexible form factors," in *Proceedings of the 13th International Conference on Human-computer Interaction*, pp. 117–133, New York, NY, USA, 2011.
- [2] A. Bergström and L. Höglund, "A national survey of early adopters of e-book reading in Sweden," *Information Research*, vol. 19, no. 2, p. 621, 2014.
- [3] G. O. Singer and A. Sundararajan, "Are digital rights valuable? Theory and evidence from e-book pricing," *SSRN Electronic Journal*, pp. 533–545, 2006.
- [4] C. Shapiro and H. R. Varian, *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston, MA, USA, 1999.
- [5] Y. Dou, M. F. Niculescu, and D. J. Wu, "Engineering optimal network effects via social media features and seeding in markets for digital goods and services," *Information Systems Research*, vol. 24, no. 1, pp. 164–185, 2013.
- [6] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: online book reviews," *Journal of Marketing Research*, vol. 43, no. 3, pp. 345–354, 2006.
- [7] N. Hu, I. Bose, Y. Gao, and L. Liu, "Manipulation in digital word-of-mouth: a reality check for book reviews," *Decision Support Systems*, vol. 50, no. 3, pp. 627–635, 2011.
- [8] P. Chatterjee, "Online reviews: do consumers use them?" *Advances in Consumer Research*, vol. 28, no. 1, pp. 129–133, 2001.
- [9] D. A. Reinstein and C. M. Snyder, "The influence of expert reviews on consumer demand for experience goods: a case study of movie critics*," *Journal of Industrial Economics*, vol. 53, no. 1, pp. 27–51, 2005.
- [10] J. Chevalier and A. Goolsbee, "Measuring prices and price competition online: Amazon and Barnes and Noble," *Quantitative Marketing and Economics*, vol. 1, no. 2, pp. 203–222, 2003.
- [11] C. Dellarocas, N. Azad, and M. Zhang, "Using online ratings as a proxy of word-of-mouth in motion picture revenue forecasting," *SSRN Electronic Journal*, 2005.

- [12] D. Godes and D. Mayzlin, "Using online conversations to study word-of-mouth communication," *Marketing Science*, vol. 23, no. 4, pp. 545–560, 2004.
- [13] W. Duan, B. Gu, and A. Whinston, "The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry," *Journal of Retailing*, vol. 84, no. 2, pp. 233–242, 2008.
- [14] Y. Liu, "Word of mouth for movies: its dynamics and impact on box office revenue," *Journal of Marketing*, vol. 70, no. 3, pp. 74–89, 2006.
- [15] E. W. Anderson, "Customer satisfaction and price tolerance," *Marketing Letters*, vol. 7, no. 3, pp. 265–274, 1996.
- [16] J. McGregor, F. Jespersen, M. Tucker, and D. Foust, "Customer service champs," *Business Week*, vol. 5, 2007.
- [17] X. Li and L. M. Hitt, "Price effects in online product reviews: an analytical model and empirical analysis," *MIS Quarterly*, vol. 34, no. 4, pp. 809–831, 2010.
- [18] W. B. Dodds, K. B. Monroe, and D. Grewal, "Effects of price, brand, and store information on buyers' product evaluations," *Journal of Marketing Research*, vol. 28, no. 3, pp. 307–319, 1991.
- [19] D. Grewal, "Product quality expectations: towards an understanding of their antecedents and consequences," *Journal of Business and Psychology*, vol. 9, no. 3, pp. 225–240, 1995.
- [20] A. Mitra, "Price cue utilization in product evaluations," *Journal of Business Research*, vol. 33, no. 3, pp. 187–195, 1995.
- [21] A. Kirmani and A. R. Rao, "No pain, no gain: a critical review of the literature on signaling unobservable product quality," *Journal of Marketing*, vol. 64, no. 2, pp. 66–79, 2000.
- [22] E. M. Rogers, *Diffusion of Innovations*, Free Press, New York, NY, USA, 1962.
- [23] V. Mahajan and E. Muller, "Innovation diffusion and new product growth models in marketing," *Journal of Marketing*, vol. 43, no. 4, pp. 55–68, 1979.
- [24] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: the role of reviewer identity disclosure in electronic markets," *Information Systems Research*, vol. 19, no. 3, pp. 291–313, 2008.
- [25] P. Chen, S. Dhanasobhon, and M. D. Smith, "All reviews are not created equal: the disaggregate impact of reviews and reviewers at Amazon.com," in *Proceedings of the 28th International Conference on Information Systems*, pp. 9–12, Montreal, Canada, 2007.
- [26] Q. W. Gan and Y. Yu, "Restaurant rating: industrial standard and word-of-mouth: a text mining and multi-dimensional sentiment analysis," in *Proceedings of the 48th Hawaii International Conference on System Sciences*, pp. 1332–1340, Kauai, HI, USA, January 2015.
- [27] R. Chen and W. Xu, "The determinants of online customer ratings: a combined domain ontology and topic text analytics approach," *Electronic Commerce Research*, vol. 17, no. 1, pp. 31–50, 2017.
- [28] R. N. Bolton and J. H. Drew, "A longitudinal analysis of the impact of service changes on customer attitudes," *Journal of Marketing*, vol. 55, no. 1, pp. 1–10, 1991.
- [29] R. A. Spreng, S. B. MacKenzie, and R. W. Olshavsky, "A reexamination of the determinants of consumer satisfaction," *Journal of Marketing*, vol. 60, no. 3, pp. 15–32, 1996.
- [30] S. Varki and M. Colgate, "The role of price perceptions in an integrated model of behavioral intentions," *Journal of Service Research*, vol. 3, no. 3, pp. 232–240, 2001.
- [31] G. B. Voss, A. Parasuraman, and D. Grewal, "The roles of price, performance, and expectations in determining satisfaction in service exchanges," *Journal of Marketing*, vol. 62, no. 4, pp. 46–61, 1998.
- [32] M. L. Richins, "Negative word-of-mouth by dissatisfied consumers: a pilot study," *Journal of Marketing*, vol. 47, no. 1, pp. 68–78, 1983.
- [33] Z. Li, W. Xu, L. Zhang, and R. Y. K. Lau, "An ontology-based Web mining method for unemployment rate prediction," *Decision Support Systems*, vol. 66, pp. 114–122, 2014.
- [34] J. Du and L. Zhou, "Improving financial data quality using ontologies," *Decision Support Systems*, vol. 54, no. 1, pp. 76–86, 2012.
- [35] C.-S. Lee, Y.-F. Kao, Y.-H. Kuo, and M.-H. Wang, "Automated ontology construction for unstructured text documents," *Data & Knowledge Engineering*, vol. 60, no. 3, pp. 547–566, 2007.
- [36] S.-H. Liao, J.-L. Chen, and T.-Y. Hsu, "Ontology-based data mining approach implemented for sport marketing," *Expert Systems with Applications*, vol. 36, no. 8, pp. 11045–11056, 2009.
- [37] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen, "The Protégé OWL plugin: an open development environment for semantic web applications," *The Semantic Web—ISWC 2004*, pp. 229–243, 2004.
- [38] H. P. Zhang and Q. Liu, *ICTCLAS*, Institute of Computing Technology, Chinese Academy of Sciences, Guangzhou, China, 2002.
- [39] D. Blei and J. Lafferty, "Correlated topic models," *Advances in Neural Information Processing Systems*, vol. 18, pp. 147–154, 2006.
- [40] T. Wang, Y. Cai, H.-f. Leung, R. Y. K. Lau, Q. Li, and H. Min, "Product aspect extraction supervised with online domain knowledge," *Knowledge-Based Systems*, vol. 71, pp. 86–100, 2014.
- [41] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, 2004.
- [42] J. Jingbo Zhu, H. Huizhen Wang, M. Muhua Zhu, B. K. Tsou, and M. Ma, "Aspect-based opinion polling from customer reviews," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 37–49, 2011.