Short or Long Review? - Text Analytics and Machine Learning Approaches to Online Reputation

Xiangming Samuel Li
Department of Management Sciences, University of Waterloo, Ontario, Canada
MBA Academic Department, University Canada West, Vancouver, British Columbia, Canada

*Correspondence: Xiangming Samuel Li, Email: samuel.li@ucanwest.ca

ABSTRACT: This paper first constructs a numerical text review score by applying text analytics and machine learning techniques to more than three million text reviews collected from the Airbnb platform. Next, we employ the text review score to analyze the effect of review length on text review score and obtain insights on the interplay between the text review length and online reputation. The main contributions of this paper include: experimenting with advanced text analytics and machine learning approaches to assess online reputation; constructing an innovative text review score as a new online reputation measure; building a large knowledge-based review corpus with labels; and obtaining important insights about the effects of text review length on online reputation. Further, it has managerial and business implications for all internet platform markets and the sharing economy players seeking to build more effective online reputation systems.

Keywords: Internet platform, Sharing economy, Online trust and reputation systems, Text review score, Review length, Text analytics, AI, Machine learning, Online purchase decision making

1. INTRODUCTION

The Internet platforms enable global online users to connect, communicate, and trade with each other, that is, conduct peer-to-peer (P2P) transactions [1, 2]. For example, eBay provides an online marketplace to enable people to exchange used goods. The Uber mobile software platform enables a driver to share a ride and the Airbnb platform allows travelers to share a host’s home and enjoy local experiences. The sharing economy that these examples constitute has witnessed an exponential growth. Statistics Canada estimates that an approximate 10% of adult Canadians (2.7 million) patronized an online marketplace to enable people to exchange used goods. The Uber mobile software platform enables a driver to share a ride and the Airbnb platform allows travelers to share a host’s home and enjoy local experiences. The sharing economy that these examples constitute has witnessed an exponential growth. Statistics Canada estimates that an approximate 10% of adult Canadians (2.7 million) patronized Airbnb and Uber and spent USD 1.31 billion from November 2015 to October 2016. Globally, the sharing economy is estimated to grow to USD 335 billion by 2025, if the rapid growth of Airbnb and Uber is taken as indicative [3].

However, the critical issue on such an Internet e-commerce platform is trust, which allows a stranger to share services without being cheated. There are cases of Airbnb hosts canceling a reservation at the last minute, overcharging the guest through false accusations of property damage, and providing poorly maintained space. Therefore, the online platform trust system, including online rating and text review, plays a critical role in mitigating such problems and facilitating platform market success. First-generation e-commerce platforms, such as eBay and Amazon, thrive on anonymous trust mechanisms, such as star rating and text review. They implemented a one-sided reputation system in which only buyers could rate and comment on issues, such as product quality, online payment, and customer services [4]. In contrast, second-generation platforms, such as Taskrabbit, Uber, and Airbnb, require both trading parties to rate and review each other. However, there are several problematic issues in a two-sided reputation system. Tadelis [5] points out a significant proportion of users (termed as “silent” users) rely on others’ rating and reviews, without contributing their own feedback. Some users provide only positive ratings owing to concerns about retaliation by the other party [6, 7]. Zervas et al. [8] find that nearly 95% of Airbnb listings boast an average user-generated rating of over 4.5 out of the possible 5. In addition, there is the problem of using a fake review to inflate product quality and service [4]. These problems in the online reputation systems may lead to the failure of the burgeoning platform economy.

Existing studies use star rating as a measure of the online trust system because it is simple and intuitive [9]. However, because star rating cannot offer fine-grained product information, it would become redundant if text reviews can be processed and scored by using advanced natural language processing (NLP) and artificial intelligence (AI) technology [10].

Thus, this paper presents an innovative approach to quantifying text reviews using advanced text analytics and machine learning approaches. Briefly, this research proposes a text analytics approach to analyze over three million text reviews on Airbnb and build a large labeled Airbnb review corpus; this corpus is used to train top supervised machine-learning algorithms to predict new text reviews. This new reputation measure is then used to investigate the effects of review length on online reputation. The study provides managerial insights on important factors impacting online
reputation; this can help all online platforms to build more effective trust systems.

2. THEORETICAL BACKGROUND

2.1 User-generated Information and Online Reputation Systems

Studies show that user-generated information or word-of-mouth (WoM) plays an important role in a consumer’s purchase decision making [11]. WoM is defined as “any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet” [12]. King et al. [13] report that WoM has higher impacts on consumers’ purchase decision than other forms of marketing communication.

Online reputation systems, such as star ratings and text reviews, are components of WoM and play a crucial role in facilitating successful online transactions [5]. Several studies find that sellers with better reputations attract more buyers, obtain higher prices, and achieve higher sales and profits. For example, Cabral and Hortacsu [14] use eBay data to show that sellers with positive feedback are more likely to stay in the market and charge higher price. Chevalier and Mayzlin [15] find that improvements in book reviews result in an increase in sales on Amazon.com and Barnesandnoble.com, while Jabr and Zheng [16] show that product referrals through online reviews have contributed to higher sales on Amazon.com. A few other studies discover a somewhat different relationship between reputation and performance. Resnick and Zeckhauser [17] find that better reputations lead to rapid sales or fast sales turnover, but no significant price hikes on eBay. Ye et al. [18] find different effects of seller reputation on price and sale volume for eBay US and Taobao China.

Whereas the role of online reputation systems has proven to be important, there are several issues in interpreting the results. One problem is the positive bias, wherein users give excessively high ratings relative to what they have experienced [5, 8]. For example, Filipas et al. [19] report that nearly 90% of Uber’s Chicago trips received a perfect 5-star rating; Chevalier & Mayzlin [15] also show that online book reviews on Amazon and Barnes and Noble are all overwhelmingly positive. Recently, Fradkin et al. [7] investigated online review bias in Airbnb by a field experiment and found that the users, who were offered an incentive coupon for submitting their reviews, reported more negative experiences; otherwise, they would have still remained “silent,” which results in overall upward positive ratings on Airbnb. A few explanations are suggested for these excessively high ratings. One is “reciprocity” bias, wherein one party rates the other party high if it has received a high rating [20]. Another is the “under-reporting” bias, wherein users with less than satisfactory experience do not bother to write a review [21]. Another explanation is the retaliation bias in which users are reluctant to write negative reviews, fearing retaliatory reviews by seller because of reciprocity [17].

2.2 Star Rating versus Text Review

The majority of studies on online reputation systems use star rating because it is simple and intuitive and can be quantified easily [9]. Star rating can serve as a cue for reading the text review content during online product searching and purchase decision process [22]. Hu et al. [23] argue that star rating is used to reduce the consideration of alternatives because it can easily capture people’s attention during online searching. However, several authors have pointed out limitations of the current system of star rating. Archak et al. [24] argue that the mean of numeric product ratings is insufficient to extract all relevant information to make a final purchase decision. Further, Pavlou and Dimoka [10] claim that star ratings cannot offer fine-grained information regarding a seller and would become redundant if text reviews can be summarized and scored.

Given the limitations of star rating, attention has shifted to utilizing text reviews to capture sellers’ reputation information [25]. Several studies show that text reviews can capture multiple features of the products and complex sentiments of reviewers, and help consumers find preferred products and increase confidence of their choices [27]. Hu et al. [28] find that text review plays a significant role in generating product sales, and Moon et al. [29] also show that text review can convey information on why viewers like or dislike products.

However, analyzing text reviews and extracting subjective opinion and related information from text reviews requires a sophisticated approach and is prone inaccurate results [30]. Because text reviews often contain a large amount of information and are coupled with conflicting opinions, such reviews can make it difficult for online consumers to identify and consider those product attributes to their final purchase decisions [31]. Assessing a review’s sentiment requires computation at the review or document level, which can be subjective [32, 33]. Further, because the text analytics approach is still relatively nascent [5], it lacks a mature methodology for dealing with unfair or faked reviews, bootstrapping mechanisms, and online trust dynamism [34].

2.3 Text Analytics Approach

Recently, some studies have tried to analyze text review data using text analytics and machine learning techniques. Text analytics or text mining is a process of discovering patterns or trends from large unstructured text database using computing algorithms [35, 36]. Archak et al. [24] employed text mining techniques to decompose textual reviews into product segmental features using an Amazon dataset and found that textual data can provide consumer’s preferences for product features. Taboada et al. [37] experimented with a lexicon-based text-mining approach and developed a semantic orientation calculator (So-Cal) to extract sentiment scores from review text, aimed to analyzing consumer opinions. Salehan and Kim [38] used the SentiStrength software as a big data analytics tool and investigated readership and helpfulness of online consumer reviews. Interestingly, Gan et al. [39] presented a text mining and sentiment analysis of online restaurant reviews; and they found that consumers’ sentiments
regarding food, service, ambience, price, and context can explain the variance in restaurants' star ratings.

It is still at an early stage to explore various text analytics techniques to analyze user-generated review texts [5, 34]. Early studies used manual text mining in which user-generated messages, such as “damaged,” “annoyed,” “late,” and “refund” are manually classified and analyzed [33]. However, manual-based text mining can be labor intensive, cumbersome, and prone to human error [40, 41]. Recent studies use the lexicon-based approach that automates the manual-based text analytics tasks by using sentimental analysis algorithms and lexicon dictionaries [37, 42]. In the field of management, Ghose and Ipeirotis [43] used linguistic analysis tools to study features of review texts and reviewers; while Netzer et al. [44] adopt semantic network analytics tools to analyze online user-generated content, including blogs, forums, and chats. However, these lexicon-based tools are designed to tackle specific domain problems; thus, they cannot be easily generalized, and often suffer from lower accuracy in labeling decisions than human experts [30, 40]. Nevertheless, it can be used to automate manual-based text-mining labor and annotate a limited number of labels for other supervised machine learning algorithms.

The most recent wave of research effort focuses on experimenting with machine learning approaches, including AI and deep learning techniques; this is also called supervised learning. The supervised learning approach can achieve much higher model accuracy than lexicon-based approaches [41, 45]. Das and Chen [46] used machine learning algorithms to classify online stock messages into three types of investment sentiments: bullish (+), bearish (-), and neutral. Rivas et al. [47] evaluated several text classification algorithms and found that deep learning method tends to outperform traditional machine learning models. However, the machine learning approach often requires a large number of training datasets (i.e., the corpus) and intensive computing resources. In addition, it is not easy to interpret the results and generalize or transfer learning to other new domains [48].

### 3. DATA AND TEXT ANALYTICS PROCESS

#### 3.1 Source of Data

Airbnb is a global lodging sharing platform for travelers in which hosts rent their spare rooms as short-term accommodation. Founded in 2008 in San Francisco, Airbnb currently has over 7 million listings in 220 countries globally with 500 million guests’ arrivals by the end of 2019; further, it generated USD 41 billion as revenue for Airbnb hosts. Airbnb provides three types of room rentals: entire home, private room, and shared room. Based on the data from New York Airbnb, the entire home constitutes 54% of all rentals, followed by private rooms (44%) and shared rooms (2%). After their stay, guests can voluntarily provide their feedback on their experiences with the hosts and their properties with star ratings and text reviews on the Airbnb website. The star rating assigns numerical values from 1 to 5 on six questions that cover host-related issues (communication, check-in, process, etc.) and property-related issues (cleanliness, location, etc.), in addition to the “overall experience.” The text review includes information about guest experiences in free format, ranging from a few words to several paragraphs (in the data used in this study, the average length is 50 words; minimum is 1; and maximum is 500 words). Airbnb used to have a review system in which reviews by hosts or guests are immediately made public so that one party can retaliate against, or reciprocate to, the other party after reading its review. Since July 10, 2014, a new system has been introduced wherein a party cannot read the other party’s review before submitting its reviews.

Studies show about 67% of Airbnb guests provided star ratings, text reviews, or both [7]; 73.5% of Uber riders provided feedback to their drivers in New York City; about 62.42% of freelance workers received feedback reviews from their clients on Taskrabbit; and about 50% of buyers left reviews on eBay [17]. In contrast, only 2% people shared their dining experiences on the public review platform, Yelp.com [49]. The difference can be explained by the degree of social interactions, presence of consolidated platform to leave a comment, and incentives for leaving a review [50, 51].

A total of 3.16 million text reviews for 182,069 listings have been collected from Inside Airbnb (http://insideairbnb.com) for the four major English-speaking Airbnb locations (Toronto, New York, London, and Sydney) in December 2018. A single host can have multiple listings. For example, if a host rents two rooms in her house, there are two listings, and professional hosts often rent out multiple listings. The 182,069 listings are the currently available ones, as of December 2018, and both star ratings and text reviews left by guests for each listing are collected. The data also include host attributes, such as super-host, casual host, professional host, the date on which the host was last reviewed, and property attributes, such as property type, room rental, parking availability, number of bedrooms and bathrooms, as well as other attributes, such as gym availability and disability accommodation. While all the individual text reviews are available for each listing, only the average value of all previous star ratings is available for a listing. After removing data with missing values and reviews written in a language other than English, a total of 2.48 million text reviews for 182,069 listings are utilized in this analysis.

Table 1 summarizes the data. Both New York and London cover more than 70% of the text review data owing to the large size of their market. However, the average number of

---

2 A semantic network is a graphical representation of sentences, where terms appear close each other if they occur together frequently in sentences.


4 Inside Airbnb is an independent open-source data collection of the Airbnb listings for public use, and it scraps data quarterly from Airbnb.com for major global cities, including Toronto, New York, London, and Sydney.
years of service of all the listings is about 3.2 years; this points to a high turnover in this business. The average reviews per listing are relatively higher in Toronto and New York, compared to London and Sydney, because Airbnb was founded in the Silicon Valley in 2008 and it first penetrated North America before expanding overseas.

<table>
<thead>
<tr>
<th></th>
<th>Toronto</th>
<th>New York</th>
<th>London</th>
<th>Sydney</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of text reviews</td>
<td>456,821</td>
<td>1,086,800</td>
<td>1,176,072</td>
<td>446,708</td>
<td>3,166,401</td>
</tr>
<tr>
<td>No. of listings</td>
<td>19,255</td>
<td>49,056</td>
<td>77,096</td>
<td>36,662</td>
<td>182,069</td>
</tr>
<tr>
<td>No. of hosts</td>
<td>10,455</td>
<td>29,618</td>
<td>37,214</td>
<td>18,041</td>
<td>95,328</td>
</tr>
<tr>
<td>Average reviews per listing</td>
<td>25</td>
<td>22</td>
<td>16</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Average listing price (USD)</td>
<td>99.21</td>
<td>140.48</td>
<td>136.92</td>
<td>127.51</td>
<td>126.00</td>
</tr>
</tbody>
</table>

Table 1: Summary of Airbnb data.

3.2 Text Analytics

Text analytics, or text mining, refers to the process of deriving useful information from text by identifying patterns, including classification, clustering, association, and detection [24, 41]. In the context of the Airbnb data, this study allocates numerical values to each text review by using text analytics and machine learning techniques, including review text normalization, feature engineering, lexicon-based unsupervised learning, supervised learning algorithms, and prediction or transfer-learning techniques. This process of identifying the polarity (e.g., positive, negative, or neutral sentiment) of each review from the subjective information contained in the reviews is also called sentiment analysis or opinion mining [30, 52]. The sentiment or polarity is extracted after the creation of a vectorized matrix for each text review.

The main process of deriving sentiment scores from raw text review is summarized in figure 1. The unstructured raw text review data from three cities (Toronto, London, and Sydney) are first converted into a structured format for text analytics; this is called feature engineering. Then, a large training data (“corpus”) is created by using several lexicon-based tools and manual checking. The rationale of using Airbnb raw data from different global cities to build the corpus is to include various heterogeneous characteristics of training data, such as different Airbnb listing properties, hosts, cities, and cultural aspects; the aim is to achieve better model-learning performance [32, 41]. The next step-called supervised learning—is to use the corpus data to train a few text analytics algorithms to learn the patterns in the text. The algorithm or pattern-matching rule learned from the training is then applied to the raw review dataset of New York to predict and derive the numerical text review score of each listing in Airbnb New York.

![Figure 1: Overview of text analytics and machine learning process.](image)

3.2.1 Text Pre-processing

The first step in text analytics is to convert the unstructured raw text review data into a structured matrix format using various NLP tools [41, 53]. This process involves tokenization (splitting text into smaller meaningful linguistic tokens), lemmatizing (sorting words by group), stop words (e.g., a, the, like, etc.) removal, and parsing text (e.g. converting texts into a tree format). Once the raw text data are normalized into linguistic components, they are converted into matrix forms using the term frequency-inverse document frequency (TF-IDF) vectorization that measures the relative importance of words in each text review and in the whole corpus dataset [32,41]. More specifically, TF-IDF is an algorithm to identify the relative importance of words in a document, and it attaches higher weights to those words that occur more frequently in a single review and inverse weights to those that occur frequently in a large selection of reviews (corpus); in the assumption that is nouns, adjectives, and verbs are more important than words, such as “the,” “and,” “e.g.,” “i.e.,” “etc.,” “it,” and “they.” Mathematically,

\[ TF-IDF = TF * IDF = \frac{tf}{\log (N/df)} \]

Where tf stands for term frequency—the number of times that the term occurs in a text review; N is total size of the corpus—the total number of text reviews in it; and df is document frequency—the number of text reviews in which the term appears.

3.2.2 Lexicon-based Labeling and Creation of Training Data

After the creation of a vectorized matrix for each text review, several lexicon-based tools are utilized to allocate a positive or negative sentiment label to each review; this is also called unsupervised learning. Lexicons are specialized dictionaries or lists of vocabularies with associated sentiment labels of positive or negative for each word, and they are constructed to analyze text sentiments. For example, a lexicon-based tool of SentiWordNet categorizes 8,427 words; Pattern NLP has 5,750 words; and VADER has 7,500 words, emoticons, and...
slang terms, all on a scale between −1.0 and +1.0. In addition, each lexicon tool has unique techniques to analyze words/phases/sentences and compute an aggregated sentiment score [54, 55].

This study uses three tools. SentiWordNet [56] is the most widely used English lexicon for sentiment analysis and opinion mining, whereas PatternNLP is a complete NLP package and uses its own subjectivity-based sentiment lexicon, containing scores for polarity, subjectivity, intensity, and confidence, along with other tags, such as part of speech to which a word belongs. Last, VADER is a rule-based sentiment analysis framework, specially built for sentiment analysis engines across social media resources, including social media text, the New York Times’ editorials, product reviews, and movie reviews [48]. Because each of these three lexicons has a unique vocabulary database, different linguistic focus, and different sentiment scoring methods, they are all “employed” to label review data in this study.

These three tools are applied to a total of 1,649,386 text reviews of listings in Toronto, London, and Sydney to label each of them as either positive or negative. The shares of positive sentiment label are 92.6% for SentiWordNet, 96.5% for PatternNLP, and 97.5% for VADER. Comparing the labels by the three lexicon-based tools, all three tools assign matching labels to 90.2% of the reviews; two tools do so for 99.9% of the reviews. This study selects the text reviews whose labels are matched by all three tools, resulting in a total of 1,487,302 reviews with 1,477,016 positive labels and 10,286 negative labels.

From the selected review data of 1,487,302, we need to create a training dataset, called a knowledge-based “corpus” that includes an equal share of positive and negative labels. The main issue in selecting corpus data is the optimal size of the dataset. In general, machine learning techniques can find a pattern in a big dataset even if there is no pattern in the real world; this is because computing algorithms have “an almost unlimited capacity to find patterns in data” [35]. This type of overfitting problem can be handled by selecting a limited amount of data. Liu [30] estimates that about three thousand sentiment words are used in most online reviews and argues that the optimal size of training data is about three to six times the number of sentiment words. Thus, this study intends to select about 10,000 to 18,000 reviews for the corpus dataset. Another issue of selecting a corpus dataset is the balance between negative and positive reviews in the dataset. Because the majority of labels are positive, the size of the corpus data is limited by the number of negative labels, that is, the 10,286 reviews resulting from the lexicon-based unsupervised learning and final human manual verification [36, 41].

Based on the number of negative labels, this study first selects a total of 20,572 reviews as potential candidates for a corpus dataset, with 10,286 positive reviews and 10,286 negative reviews. Studies show that the allocation of sentiment labels through automated lexicon-based tools is prone to errors, and manual verification is necessary [42, 44]. From the manual verification of 20,572 potential review data, only 53 are found to be false positives (positively labeled although the reviews are negative) and 1,351 are false negative (negatively labeled although the reviews are positive), leading to an error rate of 6.8% in the potential dataset. The higher rate of false negative implies that the lexicon-based tools do a better job in catching positive sentiments than negative ones in the Airbnb review text reviews. By removing the incorrectly allocated false negative reviews, we obtain 8,935 true negative reviews. We randomly selected an equal number of true positive reviews and created a corpus dataset of 17,870 text reviews.

### 3.2.3 Training with Supervised Learning

After the training dataset is created, the corpus data is randomly split into a training dataset (70% of 17,870) and a testing dataset (30% of 17,870). The share of a training dataset depends on the size of the data: a higher share, say 80% or 90%, of data can be allocated to the training dataset if the size of the database is small, that is, less than thousand observations [36]. Several supervised machine learning algorithms, or text classifiers, are trained to detect patterns in the corpus training data. Among several algorithms, this study initially used three algorithms: multinomial naive Bayes (MNB), logistic regression (LR), and support vector machine (SVM). MNB is an intuitive and fast text classifier, suitable for discrete features-based classification, and is often used for a reliable baseline to gauge other high-accuracy classifiers. LR is a classic statistical model to compute probability of output of a binary dependent variable taking values 1 (positive) or 0 (negative sentiment) by using a logistics function; it is based on input independent variables as the review text’s features. Last, SVM is a non-probability classifier, operating by separating data points in space using hyperplanes, that is, centerlines for dividing different classes, and it is well-known for solving complex unstructured data problems involving text [57].

Each of these three models is trained by the training dataset and then evaluated by comparing their predicted sentiment labels with the reserved testing dataset. By comparing the actual and predicted labels, the performance of each algorithm is compared; the results are shown in Table 2. While MNB shows significantly lower model performance than the others, both LR and SVM algorithms demonstrate very similar and outstanding results between 96% and 98% performances in terms of accuracy, precision, recall, and F1 Score. Thus, the

---

6 We experimented with the following supervised-learning algorithms: naive Bayes (MNB), Bernoulli naive Bayes (BNB), stochastic gradient decent classifier (SGDC), linear support vector classifier (LSVC), and recurrent neural networks, and long short-term memory (LSTM). LSTM can achieve very high binary classification accuracy, but it is not suitable for sentiment scoring owing to its non-linear polarized activation function.

7 F1-score is a harmonic measure of model accuracy by averaging both precision and recall, which reaches its best
following analysis uses the SVM score as text review score because of its reputation for tackling unstructured text problems and its best recall performance.

<table>
<thead>
<tr>
<th>Model performance metrics</th>
<th>MNB</th>
<th>LR</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (A): Fraction of correct classification</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Precision (P): Fraction of correct positive labels</td>
<td>0.93</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>Recall (R): Fraction of correct positives over all</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>F1 score: Weighted mean of P and R</td>
<td>0.94</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 2: Machine-learning model performance metrics: MNB, LR, and SVM.

### 3.2.3 Predicting with Transfer Learning

The last process is to apply the trained models to predict new Airbnb New York (NY) data; this is called model prediction or transfer learning for a new dataset. Each of the 836,348 raw text reviews of NY Airbnb is first pre-processed into a structured matrix format and is assigned text review scores through the trained models. The mean value of the review scores is 4.15 and standard deviation is 0.77; the mean for the star rating is 4.7 and standard deviation is 0.39. Figure 2, which compares the distribution of the two measures, shows that the distribution of the star ratings (on the left) is J-shaped; the text review scores (on the right) approximate a bell-shaped distribution. This indicates that the text review score has wider distribution, compared to that of star rating. Worth noting is that the spike between intervals 1 and 2 in the text review score is due to an automatically generated review by the system, such as “The host cancelled the booking xx days ahead of the my arrivals;” this is regarded as very negative review and approximately 2.1% of the reviews text reviews in this study are system-generated.

Moreover, Table 3 shows the analysis results of a further comparison between text review score and star rating, based on Airbnb NY listings with 37,624 observations. Star ratings match with text review scores for 10.77% of the observations (row 1). On the other hand, in rows 2 through 5, we can see that a majority (about 95%) of the matches between text review score and star rating are at the 4-5 star interval. However, only 0.46% of them match in the 3-4 star interval (row 3), and merely 0.01% of them match in the 1-2 star interval (row 5); this shows that the text review score differs substantially from the star rating, and they only match at high star rating interval (4–5). More specifically, only 8.58% of the text review scores are higher than star ratings (row 6); however, in sharp contrast, as many as 91.42% of the star ratings are higher than the text reviews scores (row 7). This suggests that a star rating is much more “upward positive” than a text review score. To gauge the magnitude of the gap between the star rating and text review score, row 8 shows that the star rating is 1 star higher than the text review scores in 11.39% of the observations; it is 2-stars higher in only 2.25% of them (row 9), and 3 stars higher in as low as 0.49% of the observations (row 10). By calculating the percentage difference between row 7 and 8, one finds that about 80% of the text review scores differ from star rating by less than 1. Over and above that, text review score is evaluated as the review text sentiment probability prediction from 0 to 1 as infinite continuous value, which enables normalized reputation scores (from 1 to 5 or even from 1 to 100 or whatever applicable) to catch up more find-grained sentiment information from the text reviews than the star ratings, which only rate the target from 1 to 5 integrally and discretely. Hence, the star ratings have much more upward positives than text review scores; this suggests that text review score can be a more effective measure of online reputation.

<table>
<thead>
<tr>
<th>Evaluation Questions</th>
<th>Text review score vs. star rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) What percentage of observations match overall (0.1 range)?</td>
<td>10.77%</td>
</tr>
<tr>
<td>(2) What percentage of observations match in the 4–5 star interval?</td>
<td>10.23%</td>
</tr>
<tr>
<td>(3) What percentage of observations match in the 3–4 star interval?</td>
<td>0.46%</td>
</tr>
</tbody>
</table>

value at 1, which implies perfect precision and recall, and the worst at 0.
Extensively in longer textual content. Consequently, unhappy situations are more likely to occur and reminding others to avoid them; further, more negative experience involves recounting the incidents that took place. There is substantial evidence to show that talking about a negative experience involves recounting the incidents that occurred and reminding others to avoid them; further, more negative experiences are more likely to occur and reminding others to avoid them. On the other hand, a longer text review is more likely to be written by a dissatisfied consumer who shares her negative consumption experience with others, a lower reputation score [68, 69]. Godes and Silva [68] investigated the relationship between online rating and review length on the Amazon online reviews, and found that longer reviews are associated with lower ratings. Further, they explained that in line with the SSE theory, online consumers are less motivated to post additional positive reviews, but more willing to write longer negative review to convey their negative feedback. More recently, Eslami et al. [25] also found that the most impacting factor in the model prediction is review length, which is negatively associated with the review rating.

However, the previous studies might have overlooked the impacts of review extremity (RE) on such an overall negative relationship between review length and reputation score. This is because the extreme reviews, including extreme positive and negative ones, have unique qualitative characteristics [15, 22]. For example, RE is extreme either when, out of a reputation score of 5, it is near 1 (extremely negative) or 5 (extremely positive). Further, the extreme reviews can be very short positive or very short negative (RE literatures), or very long negative (SSE theory). More specifically, RE, including both extremely positive and extremely negative reviews, indicate the need for topical relevancy (centroid) to measure a quality dimension of online user reviews [22]. For example, Forman et al. [69] argues that extreme reviews are unequivocal and, therefore, have more impacts on consumers’ purchase decision making, whereas the equivocal reviews or moderate reviews, such as those with a star rating of 3, are less informative than the extreme reviews having endpoint ratings of 1 or 5 stars. On the other hand, Korfitatis et al. [59] argue that extreme review need not be lengthy because they deal only with one-sided views on issues—either very good (e.g., great host) or extremely bad (e.g., dirty room); a moderate review is expected to shed more light on two-sided issues, thereby leading to longer review length [60, 70].

To summarize, a short review length can be interpreted as a very low or high text review score in the RE literature; according to the SSE theory, a long review indicates a negative sentiment. Overall, it might indicate that an inverted-U-shaped correlation exists between review length and text review score. Thus, to gain more insight into the complex relationship between review length and online reputation score, the following is hypothesized:
Hypothesis: An inverted U-shaped correlation exists between review length and text review score (reputation).

5. METHODS AND RESULTS

Next, we present both statistical and econometric analysis methods to study the effects of review length on text review score based on a cross-sectional dataset of Airbnb NY text reviews on December 7, 2018; it has 778,281 observations. More specifically, the statistical distributions of the newly computed text review score, based on the text analytics and machine learning approaches (Section 3), and their corresponding review lengths, are closely examined by using statistical techniques; further, the nonlinear correlation between review length and text review score are thoroughly studied by using IV regression analysis model.

5.1 Statistical Analysis

The histogram distribution of text review length based on the Airbnb NY text reviews is shown in figure 3. The average length is approximate 50 words (49.77); the minimum number of words in a review is 1 and maximum is 500.

To investigate the statistical relationship between review length and text review score, we performed a statistical analysis of 100 groupings of review length (1% of the reviews in each), its corresponding text review score (the mean), and standard deviation. Figure 4 shows the scatter graph and the linear fitted line on the correlation between review length and text review score based on the above statistical analysis results. An overall negative linear relationship is noticeable between review length and text review score. However, the far left of the scatter plot (the short review length) clearly shows that a non-linear relationship exists between review length and text review score. It starts with a spike at the shortest review length of 1–2, drops steeply at the review length of 3–4, and then increases steadily from the review length of 5 and reaches its peak (highest text review score of 4.44) at a review length at about 30 words. After that, it exhibits a clear downward trend with an increase in review length. This complex relationship needs to be further examined by using a more advanced econometric analysis technique that is presented next.

5.2 Econometric Analysis

The ordinary least square (OLS) regression model, equipped with an instrument variable (IV) and a quadratic term, is proposed to further examine the nonlinear correlation between review length and text review score. The descriptive statistics of the model’s variables and control variables (cohorts) are presented in Table 4. A total of 778,281 Airbnb NY text reviews since 2014-7-10 are extracted for this OLS IV regression study to alleviate the concerns of platform retaliation and reciprocity biases prior to 2014-7-10. The mean of the dependent variable, text review score, is 4.36 and standard deviation (s.d.) is 0.83; further, the average review length for NY Airbnb is 49.77 words. Moreover, we include cohorts of listing price, and Airbnb host/room/property types as control variables. Because Airbnb guests evaluated and rated the listing prices (value), host interactions and services, and property attributes in their text reviews, a text review score is computed for such Airbnb home-sharing experiences. In addition, because there is potentially endogeneity issue between the dependent variable TextRevScore and listing price, that is, a higher reputation score leads to higher listing price for the Airbnb listing, and vice versa [71], an instrumental variable (IV) called distance CC is introduced; this is the geographical distance to an interesting central location, such as the Time Square in New York, by using each Airbnb listing’s latitude and longitude variables from our collected dataset. An IV Test of endogeneity suggests that the variable “distance CC” is exogenous and effective, and thus, two-step IV regression model is employed to correct such an endogeneity.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text review score</td>
<td>Platform reputation score, dependent variable</td>
<td>4.36</td>
<td>0.83</td>
</tr>
<tr>
<td>TextRevScore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Review length</td>
<td>Length of text reviews (total words/symbols of comments)</td>
<td>49.77</td>
<td>47.53</td>
</tr>
<tr>
<td>revlen</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8 It is a total number of clean Airbnb NY text reviews after removing both non-English and missed valued records from the raw NY text reviews: 1,086,800 shown in Table 1.

9 New York Time Square is selected as the city center because it can meet both tourists and business travelers’ requirements while selecting Airbnb accommodations.
Short or Long Review? - Text Analytics and Machine Learning

The regression results, obtained by using the statistical software STATA, are presented in Table 5. It is worth noting that three continuous variables-text review score, review length, and listing price—are all converted into their logarithmic values as a response to the data skew and for facilitating an easy interpretation of the percentage change or multiplicative factors [71].

Table 5 (1)-the IV regression without the quadratic term of review length-shows that the coefficient on review length is -0.009***; this indicates that each percentage increase in the review length decreases its corresponding text review score by 0.9% at a highly significant level (P-value < 0.01). Further, it suggests a negative correlation between review length and text review score—this is in line with the previous findings by Esrami et al. [25] and Godes & Silva [68].

However, Table 5 (2)-the IV regression with quadratic term of review length-shows that the coefficient on the review length is +0.1175***; this indicates that each percentage increase in the text review length increases its corresponding text review score by 11.75% at a highly significant level (P-value < 0.01). This suggests a positive effect of review length on text review score at the beginning (the short review length). However, the highly significant and negative quadratic effect (-0.019*** ) suggests that an inverted-U-shaped relationship exists between review length and text review score at the turning point, that is, the 31st word, followed by negative effect. More interestingly, figure 5 plots and illustrates such an inverted-U-shaped relationship between review length (X-axis) and text review score (Y-axis), with a turning point at the 31st word; this, arguably, matches closely its statistical scatter plot in figure 4.

To summarize, the statistical analysis method demonstrates an overall negative relationship between review length and text review score; however, short review length appears to have a nonlinear correlation with text review score. Further, the econometrics analysis method, using the IV regression model and quadratic effect term, reveals an inverted-U-shaped relationship between review length and text review score with a turning point at the 31st word. Therefore, both statistical and economic analyses provide consistent evidence about long

---

**Table 4: Descriptive statistics of variables for Airbnb NY (N=778,281).**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE</th>
<th>N (1)</th>
<th>N (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text Review Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log (revlen)</td>
<td>-0.009***</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log (revlen)</td>
<td>-0.019***</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logprice</td>
<td>-1.073***</td>
<td>0.084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Host_type</td>
<td>-0.1013***</td>
<td>0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room_type</td>
<td>1.299***</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Property_type</td>
<td>0.108***</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DistanceCC</td>
<td>5.576***</td>
<td>0.319</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

***significance: p< 0.001, **significance: p< 0.01, *significance: p<0.05.

**Table 5: Model IV Regression Results: (1) without quadratic term; (2) with quadratic term (Airbnb NY, N=778,281).**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE (1)</th>
<th>Coefficient</th>
<th>SE (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log (TextRevScore)</td>
<td>Dependent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log (revlen)</td>
<td>-0.009***</td>
<td>0.001</td>
<td>-0.1175***</td>
<td>0.003</td>
</tr>
<tr>
<td>Log (revlen)</td>
<td>-0.019***</td>
<td>0.001</td>
<td>-0.019***</td>
<td>0.001</td>
</tr>
<tr>
<td>Logprice</td>
<td>-1.073***</td>
<td>0.084</td>
<td>-1.047***</td>
<td>0.083</td>
</tr>
<tr>
<td>Host_type</td>
<td>-0.1013***</td>
<td>0.006</td>
<td>-0.099***</td>
<td>0.006</td>
</tr>
<tr>
<td>Room_type</td>
<td>1.299***</td>
<td>0.1</td>
<td>1.269***</td>
<td>0.098</td>
</tr>
<tr>
<td>Property_type</td>
<td>0.108***</td>
<td>0.008</td>
<td>0.105***</td>
<td>0.008</td>
</tr>
<tr>
<td>DistanceCC</td>
<td>5.576***</td>
<td>0.319</td>
<td>5.289***</td>
<td>0.315</td>
</tr>
</tbody>
</table>

***significance: p< 0.001, **significance: p< 0.01, *significance: p<0.05.

Figure 5: An inverted U-shaped relationship between review length and text review score with a turning point at the 31st word (X-axis: review length; Y-axis: text review score; 95% confidence interval).

---

10 The turning point for the quadratic equation is the absolute value of $B_1 / 2B_2$.
11 Margins, at (revlen = (1/200)) marginsplot.
review length indicating a negative text review reputation score (SSE theory), but a short text review length being either very positive or very negative (RE literature). After all, the above-mentioned significant evidence supports our hypothesis in Section 4: an inverted U-shaped relationship exists between review length and text review score (reputation), which is a new finding.

6. CONTRIBUTION, LIMITATION, AND FUTURE RESEARCH
The first main contribution of this paper is that we have successfully constructed an innovative text review score as a new online reputation measure by using advanced text analytics and machine learning approaches; these can be employed to build or enhance the online reputation systems of any internet platform.

Second, we have built a new and large knowledge-based text review corpus with labels (i.e., 17,870 Airbnb original text reviews with positive or negative labels) that can be shared and used by global scholars and practitioners to study online reputation systems and sharing-economy platforms. For example, the rapidly growing sharing-economy platforms, such as Airbnb, Uber, Lyft, and Taskrabbit can utilize the large labeled corpus (the text review knowledge base) generated during this research to train and tune their machine learning models either to improve their current online reputation systems or establish a brand new platform trust mechanism.

Last, drawing from the SSE theory and review extremity (RE) literature, we used both statistical and econometric analysis methods and obtained key insights regarding the complex effects of review length on text review reputation score, that is, an inverted-U-shaped correlation between review length and text review score, a significant improvement over the previously revealed negative relationship between review length and online reputation [25, 68]. This finding has both business and managerial implications for all online platform markets. For example, because a long review suggests a negative review according to our results, online consumers may skip or skim the very long reviews and just focus on reading short reviews, thereby making speedy online purchase decisions. Further, because our study suggests that a very short review is either an extremely positive review or an extremely negative review, online consumers need to read them very carefully to differentiate highly regarded products from poorly rated one, and then make the right online purchase decisions. Moreover, because a recent consumer survey reveals that most online consumers only read about ten text reviews before making a purchase decision and most long text reviews are often ignored by online consumers, it should be very beneficial for online consumers if an online platform management summaries or highlights critical issues in the long text reviews. This has been proposed and demonstrated using the text analytics and machine learning approaches in this research, leading to more informed and rapid online purchase decision making.

The superiority of the text review reputation score over others, such as star rating, needs to be further studied theoretically. Besides the effect of review length on text review score, the impacts of product type, review age (date), reviewer attribute, and online agent behavior on the online reputation and platform economic performance, such as price, revenue, profit, and social welfare, can be further studied. Moreover, the application of the large Airbnb labeled corpus created and transfer learning of Airbnb reputation systems (text review score) are likely to be widely researched and implemented for other online platform reputation systems.

7. CONCLUSIONS
The current online review systems, including that of Airbnb, are often biased and have nearly 95% positive J-shaped distributions. During online e-commerce activities between buyers and sellers, these biased reputation systems cause what we call adverse selection and moral hazard in economics; thus, they severely undermine the market success of online platforms. While various online trust models, such as feedback-based, statistics-based, fuzzy-logic, game-theoretic, and analytics have been surveyed to tackle the online trust and reputation problems, this paper presents a novel systematic text analytics and machine learning approach to analyze more than three million user-generated text reviews on the Airbnb platform. Our first step is to use the advanced NLP text analytics techniques and unsupervised lexicon-based learning to build a new and large Airbnb review sentiment corpus. Next, leading supervised machine learning algorithms, including MNB, LR, and SVM are trained by the corpus knowledge base to construct the innovative platform text review scores. Last, the text review scoring model is applied to predict new platform reputation scores through a transfer learning approach.

The newly generated text review score has been evaluated from various perspectives. Among the AI model evaluation metrics, SVM score has been chosen as the text review score for transfer learning to predict new online reputation scores; this is because the SVM algorithm demonstrated its superior performance in solving the unstructured text review problems in terms of model accuracy, precision, recall, and F1-score. From the descriptive statistical analytics, the relationship between review length and text review scores has been investigated. Further, an econometrics IV regression analytics has been employed to examine the complex interplay between review length and text review score; it has been discovered that an inverted-U-shaped relationship between review length and text review score (reputation) exists-an important new finding, compared with the previous studies. Furthermore, it has important implications for both online platform management and reputation systems research in the future.

8. DISCLAIMERS
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. And the author wishes to thank Dr. Bonwoo Koo and Dr. Keith Hipel for their guidance and advice during his PhD research at University of Waterloo, Canada.

REFERENCES


© 2021 by the Xiangming Samuel Li. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).
AUTHOR’S BIOGRAPHY
Dr. Xiangming Samuel Li is a business professor at University Canada West and a PhD candidate in Management Sciences, University of Waterloo, Canada. He has received over 23 years of intensive management experiences in global ICT multinationals such as Nortel, Nokia, Motorola, Saveje, BTI, Anhub, and Linaro; and he is also actively engaging in academic research and teaching roles at University Canada West, University of Waterloo, University of Toronto, Hangzhou Dianzi University, and Zhejiang University of Technology. He holds an MBA degree from University of Toronto, a Master's degree of System Design Engineering from University of Waterloo, and a Bachelor' degree in Computer Sciences.