Information Technology & Tourism manuscript No.
(will be inserted by the editor)

Differences in Chinese and Western tourists faced with Japanese hospitality: A natural language processing approach

REDACTED

Received: date / Accepted: date

Abstract *Omotenashi*, the Japanese spirit of hospitality and service, is known and studied worldwide for its excellence. In recent years, the number of international tourists, particularly Chinese tourists, to Japan has been steadily increasing. However, most studies on tourists' behavior are biased toward the Western world. Previous research has shown that different cultural backgrounds give rise to different expectations and arguably, different satisfaction factors. To address the knowledge gap in this area, a cross-cultural study on the differences between Chinese and Western cultures in the high-standard Japanese hospitality environment is merited. Will the top-grade hospitality of Japan influence both populations equally, or will their cultural differences set them apart? Will they be satisfied with the soft attributes such as service or be more concerned with hard attributes such as location and facilities? We examine these questions and the differences in each population's satisfaction and dissatisfaction factors in different price ranges. Taking advantage of the facile data gathering capability enabled by the Web 2.0 technology, we applied Shannon's entropy metric to extract these factors automatically and then used them in a support vector machine to classify a more extensive data set. We then used dependency parsing and part-of-speech tagging to extract the nouns that were tied to "praising" adjectives. We found that Chinese tourists are less concerned with hospitality and more with room quality than are Western tourists, whereas, the latter are delighted more by the staff behavior. We also found that Chinese tourists are concerned with the lack of a Chinese-friendly environment, and Western customers are unsatisfied with dirty rooms or the smell of cigarettes.

Keywords Sentiment Analysis · Hotels and Lodging · Text Mining · Chinese · English · Satisfaction and Dissatisfaction Factors

REDACTED

1 Introduction

Japan's hospitality has been known historically to be of the highest quality. *Omotenashi*, which describes the spirit of Japanese hospitality, is celebrated worldwide. With roots in Japanese history and tea ceremony, Japanese hospitality is globally famous (Al-alsheikh and Sato 2015). Therefore, it would stand to reason that tourists visiting Japan would have this hospitality as their first and foremost satisfaction factor. However, it is known that customers from different countries and cultures have different expectations (Engel et al. 1990). Thus, it could be theorized that their satisfaction factors should be different. How will different cultures react and perceive hotels and their hospitality in this context?

In the last couple of decades, the economy of Japan has been affected by an increase in inbound international tourism (Jones et al. 2009). A year-on-year growth rate of 19.3% was observed in 2017, with 28,691,073 inbound tourists (Japan National Tourism Organization 2019). The tourist population was mostly Asian (86.14%), and approximately one-fourth of the total (25.63%) was from China. Western countries, including English-speaking countries in Europe, accounted for 11.4% of the total, with 7.23% being countries where English is the official or the de facto national language. The effect of Chinese tourists on international economies is increasing. Consequently, the number of studies on this phenomenon has been increasing (Sun et al. 2017). The tourist market is gradually becoming diverse because of multicultural tourist populations. Diversity in customers' cultural backgrounds means that their expectations when staying at a hotel will be varied. Therefore, hotel management must cater to these expectations to increase customer satisfaction, maintain a good reputation, and generate positive word-of-mouth.

However, many tourist-behavior analyses have been performed involving Western subjects. As such, a gap in knowledge existed until recent decades. In studies involving Asian populations in the analysis, Chinese-tourist behaviors have been evaluated most commonly (e.g. Liu et al. 2019; Chang et al. 2010; Dongyang et al. 2015). Few studies reporting comparisons between Asian and Western tourists' behaviors (e.g. Choi and Chu 2000) are typically survey- or interview-based, using small samples. These studies, although valid, can have limitations. This gap in research creates a need for cross-cultural studies for the increasing Asian and Western tourist populations. It could be said that Westerners account for a smaller portion of the tourist population compared to Asians. However, according to Choi and Chu (2000), Westerners are known as "long-haul" customers, spending more than 45% of their budget on hotels. In comparison, their Asian counterparts only spend 25% of their budget on hotels. Therefore, it is essential to study Asian and Western tourist populations, their differences, and the contrast with the existing literature results.

Owing to the advent of Web 2.0 and customer review websites, researchers realized the benefits of online reviews for research, sales (Ye et al. 2009; Basuroy et al. 2003), customer consideration (Vermeulen and Seegers 2009) and perception of services and products (Browning et al. 2013), among other ef-

fects of online interactions between customers (e.g. Xiang and Gretzel 2010; Ren and Hong 2019). Consequently, information collected online is being used in tourism research for data mining analysis, such as opinion mining (e.g. Hu et al. 2017), predicting hotel demand from online traffic (Yang et al. 2014), recommender systems (e.g. Loh et al. 2003), and more. Data mining, machine learning, and big data technologies can increase the number of manageable samples in a study from hundreds to hundreds of thousands. These technologies can not only help confirm existing theories but also lead to finding new patterns and to knowledge discovery (Fayyad et al. 1996).

In this study, we attempt to evaluate the satisfaction factors of two essential tourist populations that are culturally different from Japan: Chinese and Western tourists. We take advantage of the availability of enormous amounts of online reviews of Japanese hotels by both Mainland Chinese tourists posting on *Ctrip* and Western, English-speaking tourists posting on *TripAdvisor*. Based on these data, we can confirm existing theories regarding the differences in tourists' behavior and discover factors that could have been overlooked in the past. We use machine learning to automatically classify sentences in the online reviews as positive or negative opinions on the hotel. We then perform a statistical extraction of the topics that most concern the customers of each population.

2 Research objective

The objective of this study is to determine the difference in factors influencing satisfaction and dissatisfaction between Chinese and English-speaking tourists in the context of high-grade hospitality of Japanese hotels using text-mining techniques. We aim to determine customer groups' satisfaction and dissatisfaction factors across several price ranges. We use machine learning to classify the sentiment in texts and natural language processing to study commonly used word pairings. More importantly, we also intend to measure how hard and soft attributes influence customer groups' satisfaction and dissatisfaction. We define hard attributes as relating to physical aspects and environmental aspects, such as the hotel's facilities, location, infrastructure, and surrounding real estate. In contrast, soft attributes are the hotel's non-physical attributes related to services, staff, or management.

We use large scale data from online hotel reviews in Chinese and English to study their differences in a statistical manner. In the past, survey-based studies have provided a theoretical background for a few, specific tourist populations of a single culture or traveling with a single purpose. The limited scope of these studies often leads to difficulties in observing cultural and language differences in a single study.

We attempt to clarify the difference in satisfaction and dissatisfaction factors between different cultures. These factors can become the focal point for improving the tourism and service industries and increasing customer satis-

faction. Satisfied customers will then write more positive online reviews that will, in turn, increase sales and attract new customers.

3 Theoretical background and hypothesis development

3.1 Japanese hospitality and service: *Omotenashi*

The spirit of Japanese hospitality, or *Omotenashi*, has roots in the country's history, and to this day, it is regarded as the highest standard (Ikeda 2013; Al-alsheikh and Sato 2015). There is a famous phrase in customer service in Japan: *okyaku-sama wa kami-sama desu*, meaning "The customer is god." Some scholars say that *omotenashi* originated from the old Japanese art of the tea ceremony in the 16th century. However, other scholars found that it originates even earlier, in the form of formal banquets in the 7th century (Aishima et al. 2015). The practice of high standards in hospitality has survived throughout the years. Presently, it permeates all business practices in Japan, from the cheapest convenience stores to the most expensive ones. Manners, service, and respect towards the customer are taught to workers in their training. High standards are always followed to not fall behind in the competition. In Japanese businesses, including hotels, staff members are trained to speak in *sonkeigo*, or "respectful language," one of the most formal of the Japanese formality syntaxes. They are also trained to bow differently depending on the situation, where a light bow could be used to say "Please, allow me to guide you." Deep bows are used to apologize for any inconvenience the customer could have faced, followed by a very respectful apology. Although the word *omotenashi* can be translated directly as "hospitality," it includes both the concepts of hospitality and service (Kuboyama 2020). This hospitality culture permeates every type of business with customer interaction in Japan. A simple convenience shop could express all of these hospitality and service standards, which are not exclusive to hotels.

It stands to reason that this cultural aspect of hospitality would be a positive factor in influencing customer satisfaction. However, in many cases, other factors such as proximity to a convenience store, transport availability, or room quality might be more critical to a customer. In this study, we cannot directly determine whether a hotel is practicing the cultural standards of *omotenashi*. Instead, we consider it as a cultural factor that influences all businesses in Japan. We then observe the customers' evaluations regarding service and hospitality factors in general and compare to other places and business practices in the world. In summary, we consider the influence of the cultural aspect of *omotenashi* while analyzing the evaluations on service and hospitality factors that are universal to all hotels in any country.

Therefore, we pose the following research question:

Research Question 1a: *To what degree are Chinese and Western tourists satisfied with Japanese hospitality factors such as staff behavior or service?*

However, Japanese hospitality is based on the Japanese culture. Different cultures interacting with it could provide a different evaluation of it. Some might be impressed by it, whereas some might consider other factors more important to their stay in a hotel. This point leads us to a derivative of the aforementioned research question:

Research Question 1b: *Do Western and Chinese tourists have a different evaluation of Japanese hospitality factors such as staff behavior or service?*

3.2 Customer satisfaction and dissatisfaction towards individual factors during hotel stay

Customer satisfaction in tourism has been analyzed since decades past, Hunt (1975) having defined customer satisfaction as the realization or overcoming of expectations towards the service. Oliver (1981) defined it as an emotional response to the provided services in retail and other contexts, and Oh and Parks (1996) reviewed the psychological processes of customer satisfaction for the hospitality industry. It is generally agreed upon that satisfaction and dissatisfaction stem from the individual expectations of the customer. As such, Engel et al. (1990) states that each customer's background, therefore, influences satisfaction and dissatisfaction. Previous studies on the dimensions of culture that influence differences in expectations have been performed in the past Donthu and Yoo (1998). Western and Chinese customers can then have very different satisfaction and dissatisfaction factors since they have different backgrounds and cultures. These varying backgrounds will lead to varying expectations of the hotel services, the experiences they want to have while staying at a hotel, and the level of comfort that they will have. These expectations will be there from the moment that they choose the hotel throughout their stay. In turn, these different expectations will determine the distinct factors of satisfaction and dissatisfaction for each kind of customer and the order in which they prioritize them.

Because of their different origins, expectations, and cultures, it stands to reason Chinese and Western tourists could have completely different factors to one another. Therefore, it could be that some factors do not appear in the other reviews at all. For example, between different cultures, it can be that a single word can express some concept that would take more words in the other language. Therefore, we must measure their differences or similarities at their common ground as well.

However, in this study, we study not overall customer satisfaction but the satisfaction and dissatisfaction that stem from individual-specific expectations, be they conscious or unconscious. For example, suppose a customer has a conscious expectation of a comfortable bed and a wide shower, and it is realized during their visit. In that case, they will be satisfied with this matter. However, suppose that same customer with a conscious expectation of a comfortable bed experienced loud noises at night. In that case, they can be dissatisfied with a different aspect, regardless of the satisfaction towards the bed. Then, the

same customer might have packed their toiletries, thinking that the amenities might not include those. They can then be pleasantly surprised with good quality amenities and toiletries, satisfying an unconscious expectation. This definition of satisfaction does not allow us to examine overall customer satisfaction. However, it will allow us to examine the factors that a hotel can revise individually and how a population perceives them as a whole. In our study, we consider the definitions in Hunt (1975) that satisfaction is a realization of an expectation, and we posit that customers can have different expectations towards different service aspects. Therefore, in our study, we define satisfaction as the emotional response to the realization or overcoming of conscious or unconscious expectations towards an individual aspect or factor of a service. On the other hand, dissatisfaction is the emotional response to the lack of a realization or under-performance of these conscious or unconscious expectations towards specific service aspects.

Studies on customer satisfaction (e.g. Truong and King 2009; Romão et al. 2014; Wu and Liang 2009) commonly use the Likert scale (Likert 1932) (e.g. 1 to 5 scale from strongly dissatisfied to strongly satisfied) to perform statistical analysis of which factors relate most to satisfaction on the same dimension as dissatisfaction (e.g. Chan et al. 2015; Choi and Chu 2000). The Likert scale's use leads to correlation analyses where one factor can lead to satisfaction, implying that the lack of it can lead to dissatisfaction. However, a binary distinction (satisfied or dissatisfied) could allow us to analyze the factors that correlate to satisfaction and explore factors that are solely linked to dissatisfaction. There are fewer examples of this approach, but studies have done this in the past (e.g. Zhou et al. 2014). This method can indeed decrease the extent to which we can analyze degrees of satisfaction or dissatisfaction. However, it has the benefit that it can be applied to a large sample of text data via automatic sentiment detection techniques using artificial intelligence.

Previous research has also focused more on soft attributes, with little focus on hard attributes, such as location or infrastructure, mostly focusing only on facilities (e.g. Shanka and Taylor 2004; Choi and Chu 2001). However, hard factors, which are uncontrollable by the hotel staff, can play a part in the customers' choice behavior and satisfaction. Examples of these factors include the hotel's surroundings, location, language immersion of the country as a whole, or touristic destinations, and the hotel's integration with tours available nearby, among other factors.

This leads to another couple of research questions:

Research Question 2a: *To what degree does satisfaction and dissatisfaction stem from hard and soft attributes of the hotel?*

Research Question 2b: *How differently do Chinese and Western customers perceive hard and soft attributes of the hotel?*

The resulting proportions of hard attributes to soft attributes for each population could measure how much the improvement of management in the hotel can increase future satisfaction in customers.

3.3 Chinese and Western tourist behavior

In the past, social science and tourism studies focused extensively on Western tourist behavior in other countries. Recently, however, with the rise of Chinese outbound tourism, both academic researchers and businesses have decided to study Chinese tourist behavior. Explaining this increase, Sun et al. (2017) analyzed the number of studies related to Chinese tourists from 2001 to 2012 and found a steady increase until 2007, followed by the rapid growth of Chinese tourism studies. This increase in Chinese tourist behavior research resulted in several studies focusing on only the behavior of this subset of tourists. To this day, studies and analyses specifically comparing Asian and Western tourists are scarce, and even fewer are the number of studies explicitly comparing Chinese and Western tourists. One example is a study by Choi and Chu (2000), who found that Western tourists visiting Hong Kong are satisfied more with room quality, while Asians are satisfied with the value for money. Another study by Bauer et al. (1993) found that Westerners prefer hotel health facilities. At the same time, Asian tourists were more inclined to enjoy the Karaoke facilities of hotels. Both groups tend to have high expectations for the overall facilities. Another study done by Kim and Lee (2000) found that American tourists were found to be individualistic and motivated by novelty. In contrast, Japanese tourists were collectivist and motivated by prestige and family, with an escape from routine and increased knowledge as a common motivator.

One thing to note with the above Asian vs. Western analyses is that they were performed before 2000 and that they are not Chinese specific but study Asian people in general. Meanwhile, the current Chinese economy boom is increasing the influx of tourists of this nation. The resulting increase in marketing and the creation of guided tours for Chinese tourists could have created a difference in tourists' perceptions and expectations. In turn, if we follow the definition of satisfaction in Hunt (1975), then change in expectations could have influenced their satisfaction factors when traveling. Another note is that these studies were performed with questionnaires in places where it would be easy to locate tourists, i.e., airports. However, our study of online reviews takes the data that the hotel customers uploaded themselves. This data makes the analysis unique in exploring their behavior compared with Western tourists via factors that are not considered in most other studies. Furthermore, our study is unique in observing the customers in the specific environment of high-level hospitality in Japan.

More recent studies have surfaced as well. A cross-country study (Francesco and Roberta 2019) using posts from U.S.A. citizens, Italians, and Chinese tourists, determined using a text link analysis that customers from different countries indeed have a different perception and emphasis of a few predefined hotel attributes. According to their results, U.S.A. customers perceive cleanliness and quietness most positively. In contrast, Chinese customers perceive budget and restaurant above other attributes. Another couple of studies (Jia 2020; Huang 2017) analyze differences between Chinese and U.S. tourists using

text mining techniques and more massive datasets, although in a restaurant context.

These last three studies focus on the U.S.A. culture, whereas our study focuses on the Western culture. Another difference with our study is that of the context of the study. The first study (Francesco and Roberta 2019) was done within the context of tourists from three countries staying in hotels across the world. The second study chose restaurant reviews from the U.S.A. and Chinese tourists eating in three countries in Europe. The third study analyzed restaurants in Beijing.

On the other hand, our study focuses on Western culture, instead of a single Western country, and Chinese culture clashing with the hospitality environment in Japan, specifically. Japan's importance in this analysis comes from the unique environment of high-grade hospitality that the country presents. In this environment, do customers hold their satisfaction to this hospitality regardless of their culture, or are other factors more relevant to the customers? Our study measures this at a large scale across different hotels in Japan.

Other studies have gone further and studied people from many countries in their samples and performed a more universal and holistic (not cross-culture) analysis. Choi and Chu (2001) analyzed hotel guest satisfaction determinants in Hong Kong with surveys in English, Chinese and Japanese translations, with people from many countries in their sample. Choi and Chu (2001) found that staff service quality, room quality, and value for money were the top satisfaction determinants. As another example, Uzama (2012) produced a typology for foreigners coming to Japan for tourism, without making distinctions for their culture, but their motivation in traveling in Japan. In another study, Zhou et al. (2014) analyzed hotel satisfaction using English and Mandarin online reviews from guests staying in Hangzhou, China coming from many countries. The general satisfaction score was noticed to be different among those countries. However, a more in-depth cross-cultural analysis of the satisfaction factors was not performed. As a result of their research, Zhou et al. (2014) thus found that customers are universally satisfied by welcome extras, dining environments, and special food services.

Regarding Western tourist behavior, a few examples can tell us what to expect when analyzing our data. Kozak (2002) found that British and German tourists' satisfaction determinants while visiting Spain and Turkey were hygiene and cleanliness, hospitality, the availability of facilities and activities, and accommodation services. Shanka and Taylor (2004) found that English-speaking tourists in Perth, Australia were most satisfied with staff friendliness, the efficiency of check-in and check-out, restaurant and bar facilities, and lobby ambiance.

Regarding outbound Chinese tourists, academic studies about Chinese tourists have increased (Sun et al. 2017). Different researchers have found that Chinese tourist populations have several specific attributes. According to Ryan and Mo (2001) and their study of Chinese tourists in New Zealand, Chinese tourists prefer nature, cleanliness, and scenery in contrast to experiences and activities. Dongyang et al. (2015) studied Chinese tourists in the Kansai region

of Japan and found that Chinese tourists are satisfied mostly with exploring the food culture of their destination, cleanliness, and staff. Studying Chinese tourists in Vietnam, Truong and King (2009) found that Chinese tourists are highly concerned with value for money. According to Liu et al. (2019), Chinese tourists tend to have harsher criticism compared with other international tourists. Moreover, as stated by Gao et al. (2017), who analyzed different generations of Chinese tourists and their connection to nature while traveling, Chinese tourists prefer nature overall. However, the younger generations seem to do so less than their older counterparts.

Although the studies focusing only on Chinese or Western tourists have a narrow view, their theoretical contributions are valuable. We can see that depending on the study and the design of questionnaires and the destinations, the results can vary greatly. Not only that, but while there seems to be some overlap in most studies, some factors are completely ignored in one study but not in the other. Since our study uses data mining, each factor's definition is left for hotel customers to decide en masse via their reviews. This means that the factors will be selected through statistical methods alone, instead of being defined by the questionnaire. Our method allows us to find factors that we would not have contemplated. It also avoids enforcing a factor on the mind of study subjects by presenting them with a question that they did not think of by themselves. This large variety of opinions in a well-sized sample, added to the automatic findings of statistical text analysis methods, gives our study an advantage compared to others with smaller samples. This study could also help us analyze the satisfaction and dissatisfaction factors cross-culturally and compare them with the existing literature.

Undoubtedly previous literature has examples of other cross-culture studies of tourist behavior and may serve to further highlight our study and its merits. A contrast is shown in Table **??**. This table shows that older studies were conducted with surveys and had a different study topic. These are changes in demand (Bauer et al. 1993), tourist motivation (Kim and Lee 2000), and closer to our study, satisfaction levels (Choi and Chu 2000). However, our study topic is not the levels of satisfaction but the factors that drive it and dissatisfaction, which is overlooked in most studies. Newer studies with larger samples and similar methodologies have emerged, although two of these study restaurants instead of hotels (Jia 2020; Huang 2017). One important difference is the geographical focus of their studies. While Francesco and Roberta (2019) , Jia (2020) and Huang (2017) have a multi-national focus, we instead focus on Japan. The focus on Japan is important because of its top rank in hospitality across all types of businesses. This raises the question: in such an environment, will the customers be universally satisfied with this factor, or will they have differing views within their cultures? Our study brings light to the changes, or lack thereof, in different touristic environments where an attribute can be considered excellent. The number of samples in other text-mining studies is also smaller than ours in comparison. Apart from that, every study has a different text mining method.

3.4 Data mining, machine learning, knowledge discovery and sentiment analysis

In the current world, data is presented to us in larger and larger quantities. Today's data sizes were commonly only seen in very specialized large laboratories with supercomputers a couple of decades ago. However, they are now standard for market and managerial studies, independent university students, and any scientist connecting to the Internet. Such quantities of data are available to study now more than ever. Nevertheless, it would be impossible for researchers to parse all of this data by themselves. As Fayyad et al. (1996) summarizes, data by itself is unusable until it goes through a process of selection, preprocessing, transformation, mining, and evaluation. Only then can it be established as knowledge. With the tools available to us in the era of information science, algorithms can be used to detect patterns that would take researchers too long to recognize. These patterns can, later on, be evaluated to generate knowledge. This process is called Knowledge Discovery in Databases.

Now, there are, of course, many sources of numerical data to be explored. However, perhaps what is most available and interesting to managerial purposes is the resource of customers' opinions in text form. Since the introduction of Web 2.0, an unprecedented quantity of valuable information is posted to the Internet at a staggering speed. Text mining has then been proposed more than a decade ago to utilize this data (e.g. Rajman and Besançon 1998; Nahm and Mooney 2002). Using Natural Language Processing, one can parse language in a way that translates to numbers so that a computer can analyze it. Since then, text mining techniques have improved over the years. This has been used in the field of hospitality as well for many purposes, including satisfaction analysis from reviews (e.g Berezina et al. 2016; Xu and Li 2016; Xiang et al. 2015; Hargreaves 2015; Balbi et al. 2018), social media's influence on travelers (e.g. Xiang and Gretzel 2010), review summarization (e.g. Hu et al. 2017), perceived value of reviews (e.g Fang et al. 2016), and even predicting hotel demand using web traffic data (e.g Yang et al. 2014).

More than only analyzing patterns within the text, researchers have found how to determine the sentiment behind a statement based on speech patterns, statistical patterns, and other methodologies. This method is called sentiment analysis or opinion mining. A precursor of this method was attempted decades ago (Stone et al. 1966). With sentiment analysis, one could use patterns in the text to determine whether a sentence was being said with a positive opinion, a critical opinion. This methodology could even determine other ranges of emotions, depending on the thoroughness of the algorithm. Examples of sentiment analysis include ranking products through online reviews (e.g Liu et al. 2017; Zhang et al. 2011), predicting political poll results through opinions in Twitter (O'Connor et al. 2010), and so on. In the hospitality field, it has been used to classify reviewers' opinions of hotels in online reviews (e.g. Kim et al. 2017; Al-Smadi et al. 2018).

The algorithm used for sentiment analysis in our study is called a support vector machine, a form of supervised machine learning used for binary classifi-

cation. This means a sample of labeled training data is given to the algorithm to detect patterns in the data and use those patterns to establish a method for classifying other unlabeled data automatically. Machine learning is a general term used for algorithms that, when given data, will automatically use that data to "learn" from its patterns and apply them for improving upon a task. Learning machines can be supervised, as in our study, where the algorithm has manually labeled data to know the correct task result template. Machine learning can also be unsupervised, where without any pre-labeled data. In this latter case, the machine will analyze the structure and patterns of the data and perform a task based on its conclusions. Our study calls for a supervised machine since text analysis can be intricate. Many patterns might occur, but we are only interested in satisfaction and dissatisfaction labels. Consequently, we teach the machine through previously labeled text samples.

Machine learning and data mining are two fields with a significant overlap since they can use each other's methods to achieve the task at hand. Machine learning methods focus on predicting new data based on known properties and patterns of the given data. Data mining, on the other hand, is discovering new information and new properties of the data. Our machine learning approach will learn the sentiment patterns of our sample texts showing satisfaction and dissatisfaction and using these to label the rest of the data. We are not exploring new patterns in the sentiment data. However, we are using sentiment predictions for knowledge discovery in our database. Thus, our study is a data mining experiment based on machine learning.

Because the methodology for finding patterns in the data is automatic and statistical, it is both reliable and unpredictable. Reliable in that the algorithm will find a pattern by its nature. Unpredictable in that since it has no intervention from the researchers in making questionnaires, it can result in anything that the researchers could not expect. These qualities determine why, similar to actual mining, data mining is mostly exploratory. One can never be sure that one will find a specific something. However, we can make predictions and estimates about finding knowledge and what kind of knowledge we can uncover. The exploration of large opinion datasets with these methods is essential. The reason being that we can discover knowledge that could otherwise be missed by observing a localized sample rather than taking a holistic view of every user's opinion. In other words, a machine algorithm can find the needles in a haystack that we did not know were there by examining small bundles of hay at a time.

In this study, we predict that several things might occur. Our data show satisfaction and dissatisfaction factors that are universal and strictly cultural factors. However, we expect that both of these options will present themselves. We can also assert that we could arrive at very similar results to previous literature if they are correct in their findings. However, we are using a database of several orders of magnitude larger. We can also expect to discover patterns that researchers previously had not noticed because of the lack of questionnaire design and users' freedom to record their pleasures and grievances.

4 Data Analysis

4.1 Frequent keywords in differently priced hotels

To understand Chinese tourists and English-speaking tourists' satisfaction and dissatisfaction factors when lodging in Japan, we study both the frequency of the words they use. Following that, to know the relevance of a keyword as a preference for each group, we observed each entropy-based keyword's frequencies in our complete data set and in each price range. The frequency of the keywords in the database shows the level of priority it has for customers.

We observed the top 10 words with the highest frequencies for keywords linked by entropy to satisfaction and dissatisfaction in emotionally positive and negative statements to study. The keywords are the quantitative rank of the needs of Chinese and English-speaking customers. We show the top 10 positive keywords for each price range comparing English and Chinese in Table **??**. For the negative keywords, we show the results in Table **??**.

We can observe that the most used keywords for most price ranges in the same language are similar, with a few changes in priority for the keywords involved. For example, in Chinese, we can see that the customers praise cleanliness first in cheaper hotels, whereas the size of the room or bed is praised more in hotels of higher class. Another example is that in negative English reviews, complaints about price appear only after 10,000 yen hotels. After this, it climbs in importance following the increase in the hotel's price.

4.2 Frequently used adjectives and their pairs

Some keywords in these lists are adjectives, such as the word "大 (big)" mentioned before. To understand those, we performed the dependency parsing, and part of speech tagging explained in section **??**. While many of these connections, we only considered the top 4 used keyword connections per adjective per price range. We show the most used Chinese adjectives in positive keywords in Table **??**, and for negative Chinese adjective keywords in Table **??**. Similarly, for English adjectives used in positive sentences we show the most common examples in Table **??**, and for adjectives used in negative sentences in Table **??**.

4.3 Determining hard and soft attribute usage

To further understand the differences in satisfaction and dissatisfaction in Chinese and Western customers of Japanese hotels, we classified these factors as either hard or soft attributes of a hotel. We define hard attributes as matters regarding the hotel's physical or environmental aspects, such as facilities, location, or infrastructure. Some of these aspects would be impractical for the hotel to change, such as its surroundings and location. Others can be

expensive to change, such as matters requiring construction costs, which are possible but would require significant infrastructure investment. On the other hand, soft attributes are the non-physical attributes of the hotel service and staff behavior that are practical to change through management. For example, the hotel's services or the cleanliness of the rooms are soft attributes. For our purposes, amenities, clean or good quality bed sheets or curtains, and other physical attributes that are part of the service and not of the hotel's physical structure are considered soft attributes. Thus, we can observe the top 10 satisfaction and dissatisfaction keywords and determine whether they are soft or hard attributes.

We manually labeled each language's top keywords into either hard or soft by considering how the word would be used when writing a review. If the word described unchangeable physical factors by the staff or management, we consider them hard. If the word implied an issue that could be solved or managed by the hotel staff or management, we consider it soft. For adjectives, we looked at the top four adjective and noun pairings used in the entire dataset and counted the percentage of usage in each context. If it was not clear from the word or the pairing alone, we declared it undefined. Then, we added the counts of these words in each category. A single word with no pairing is always deemed 100% in the category it corresponds to. We add the partial percentages for each category when an adjective includes various contexts. The interpretation of these keywords is shown in the Tables **??** and **??**. We can see the summarized results for the hard and soft percentages of positive and negative Chinese keywords in Figure **??**. For the English keywords, see Figure **??**.