**Overview**

The proliferation of online reviews has become a staple of digital communications and commerce. Product reviews (Amazon), movie reviews (Rotten Tomatoes), and hotel reviews (TripAdvisor) are standard practice for digital sites, with most of the content generated by and commented on by users. The method of User Generated Content (UGC) as a tool for researching, recommending, or buying products and services has also become a central part of social media and online marketing. Twitter, Instagram, and TikTock are predicated on the creation and dissemination of UGC, and a class of content creators and highly compensated online influencers has emerged from the practice.

While UGC is not new—it has been an essential part of the Internet since its inception—how it can be parsed and analyzed continues to have business implications for marketing and sales. However, the sheer volume of online reviews can make conducting an analysis complicated. An analysis of the reviews on TripAdvisor offers hotel companies insights that would be difficult and costly to do using standard marketing analyses such as customer satisfaction surveys. For the hotel and travel industry, the value of identifying favorable or positive reviews is essential to their business success. Thus, we have chosen to conduct sentiment analysis to narrow the scope of the inquiry and identify the most frequent positive and negative words in hotel reviews.

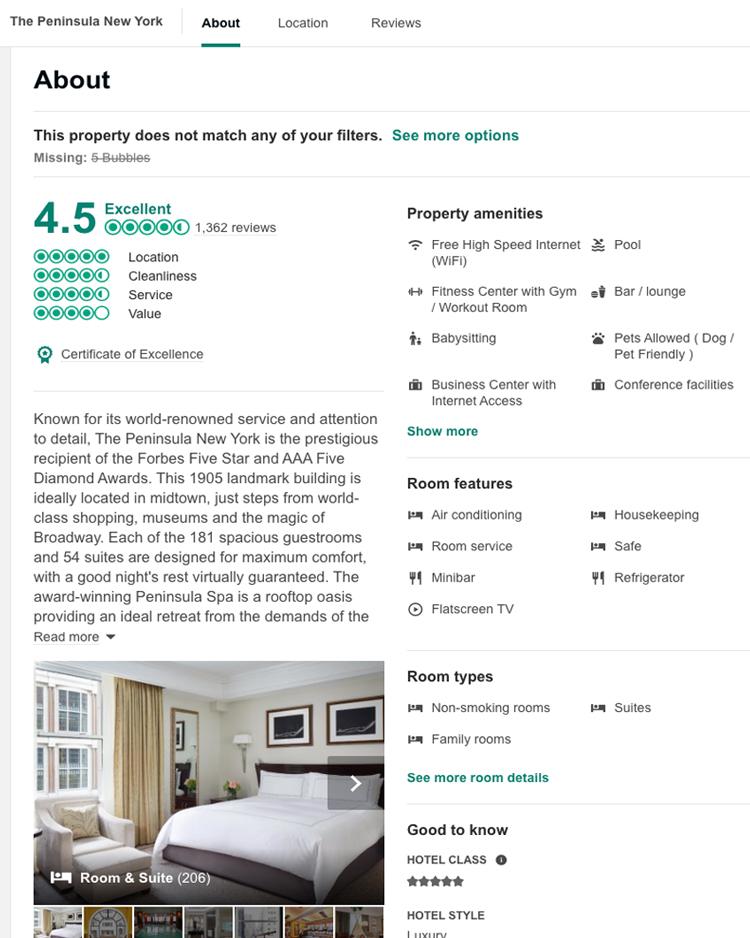
We have chosen to limit the sentiment analysis to TripAdvisor reviews of 10 hotels in New York City. Initially, the study planned to focus on one neighborhood, but the scope broadened to allow a future analysis of location and nearby attractions as variables.

**Background**

TripAdvisor, founded in 2000, is the world's largest travel platform with more than 830 million reviews of 8.6 million properties, airlines, and cruises. There are 507 hotels listed in New York City.

Hotel reviews are based on a scale of 1–5 corresponding to rankings of terrible, poor, average, good, excellent. Users can also rate hotels on attributes including location, cleanliness, service, and value on a scale of 1–5. Review copy can only be submitted by travelers and not by property owners or managers. Travelers are advised to include the following information in their reviews:

* Describe the service you experienced
* Tell us how you liked the room
* Talk about the location & amenities
* Say what you liked best & least

*Figure 1:* Sample TripAdvisor review

**Dataset Description**

To create the dataset, we randomly selected 10 hotels in different neighborhoods in Manhattan and [collected](https://drive.google.com/open?id=1FSuVQDO-3u4i-SukPVe0kSO6WBOE61mV) the review text (13,114) and the ratings.

The reviews were [merged](https://drive.google.com/open?id=1qAw4M6-h-88AZsJUYOMmNXnUwql6X8kz) into one file with the results outlined in the table below:

*Table 1:* 10 hotels to be scraped

|  |  |  |
| --- | --- | --- |
| **Hotel** | **Number of reviews** | **Rating** |
| Arlo NoMad | 1,420 | 4.5 |
| Broadway Plaza Hotel | 1,771 | 4 |
| CitizenM New York Bowery | 637 | 4.5 |
| Hotel Mulberry | 691 | 4.5 |
| Lex Boutique Hotel | 877 | 4.5 |
| Lyric | 769 | 4.5 |
| Peninsula New York | 1,363 | 4.5 |
| Radisson Wall Street | 2,455 | 4 |
| SoHotel | 1,819 | 4 |
| The Gregory Hotel | 1,312 | 4 |
| Total | 13,114 |  |

**Processing/Data Preparation**

The sentiment analysis intends to identify the frequently used positive and negative words in reviews; therefore, we are interested only in those reviews that are rated 1-2 and 4-5. The first step of the [data understanding](https://drive.google.com/open?id=1uOqE7A_YbNpWGpj0qrxxpzr2wPlb-pn4) was to remove ratings of 3 since they are considered average and are neither positive nor negative. We did this by setting a binary target: if rating is > 3 = 1 (positive review); <3 = 0 (negative review).

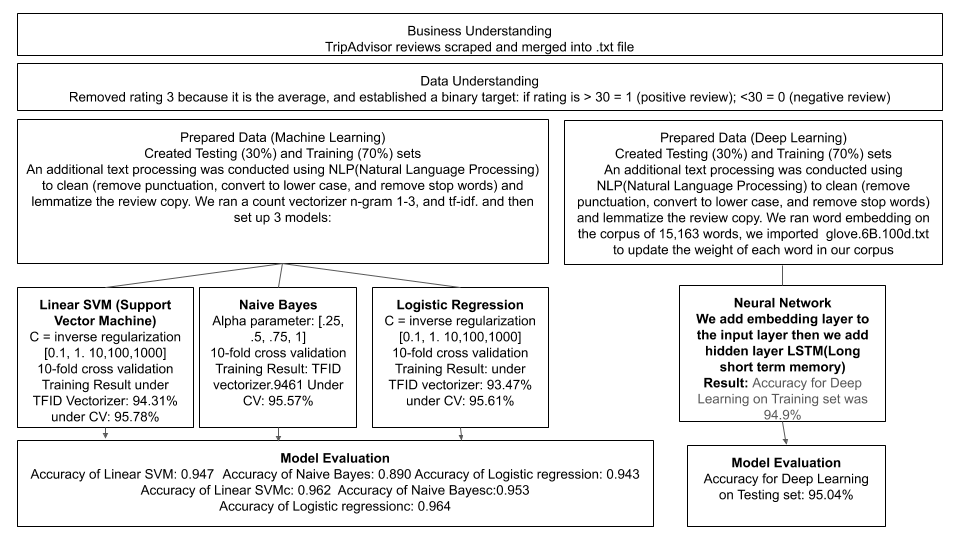
The resulting data set was cleaned using NLTK:

* Normalize case: every token is in lowercase
* Stemming: remove suffixes (like plurals or "ing")
* Stop-word removal
* Punctuation removal

Using a [machine learning process,](https://drive.google.com/open?id=1qAw4M6-h-88AZsJUYOMmNXnUwql6X8kz) a training set was created by dividing the dataset into 70% for testing, and 30% for training and three models were run: Support Vector Machine (SVM), Naive Bayes, Logistic Regression. For each model, we ran a count vectorizer, 10-fold cross-validation, and tf-idf.

Additionally, a [deep learning process](https://drive.google.com/open?id=1-t4gQ7pD_7tuqJTBy7c6LcsRFS8Xv1jp) was run using a neural network. All four models were tested for accuracy of collecting positive and negative words.

The following flow chart maps the step-by-step process of evaluating the accuracy of the data set.

*Figure 2:* Flowchart of data preparation and modeling

**Model Evaluation**

The results of the model testing showed the SVM model to be the most accurate at identifying positive and negative words. We collected the words, and below is a sample of what the modeling revealed.

*Table 2:* Positive and negative words

|  |  |
| --- | --- |
| **POSITIVE** | **NEGATIVE** |
| helpful | worst |
| friendly | rude |
| efficient | dirty |
| **arlo** | terrible |
| **maria** | poor |
| spacious | uncomfortable |
| excellent | horrible |
| accommodating | stay away |
| welcoming | rat |
| exceptional | staff rude |
| excellent hotel | unhelpful |

The results reveal words that are not surprising to see in hotel reviews, both positive and negative (excellent hotel vs. horrible). However, what was telling was the frequency with which proper names (Maria, Arlo) were mentioned in the positive words, whereas that didn't occur in the negative words. Guests tend to remember excellent service and will say it in their reviews.

This can have business implications for human resources and hiring, as well as training and staffing. Similarly, the negative words had insights for staffing but trended toward accommodations and comfort, indicating to management what needs improvement.

**Conclusion and Next Steps**

Although the model was accurate at identifying positive and negative words in the reviews, our study was exploratory and should continue on larger samples.

As mentioned above, the results of the sentiment analysis go beyond identifying good and bad hotels. It highlights opportunities for further study: mapping visitor sentiment against specific rating attributes such as location, cleanliness, service, and value. This can have business implications beyond marketing and can have a real impact on staffing, training programs, and third-party vendors (linens, food, etc.). An analysis such as this offers property managers new tools to examine how their hotel is being perceived and how they can improve on all aspects of their business model.

**References**

Chang, Yung-Chun; Ku, Chih-Hao; Chen, Chun-Hung, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor."

International Journal of Information Management ,Volume 48, October 2019. Pages 263-279 <https://doi.org/10.1016/j.ijinfomgt.2017.11.001>

García-Barriocanal, Elena; Sicilia, Miguel-Angel; and Korfiatis, Nikolaos, "Exploring Hotel Service Quality Experience Indicators In User-Generated Content: A Case Using Tripadvisor Data" (2010). MCIS 2010 Proceedings. Paper 33.

<http://aisel.aisnet.org/mcis2010/33>

O'Connor, Peter, “User-Generated Content and Travel: A Case Study on Tripadvisor.com”