# CS584: TripAdvisor Reviews Analysis and Prediction of Attribute values

**Krina Uday Shah[1], Varun Reddy Doddipalli[1]**
[1]Stevens Institute of Technology
kshah119@stevens.edu, vdoddipa@stevens.edu

**ABSTRACT**

Sentiment analysis is essential for detecting and understanding customer emotions. Companies use these to understand the feelings of their customers and thus to improve customer services. In our project we aim to collect user reviews from Tripadvisor website. Tripadvisor offers online hotel bookings for transportation, accommodation, travel experiences and restaurants. We target those customers who have booked their hotel in NYC and extracted their reviews for our analysis. From the model trained, we try to obtain the estimated values for the rating attributes from the user text reviews which in turn would help the business partner of the hotel realize the quality of service that the hotel provides and which specific areas can be improved.

## 1 Introduction

Analysis of user comments and reviews can help businesses in understanding how their customers are feeling about their products and services, which in turn provides deep insights into major stakeholders in the business on how to improve specific areas of products and services.

Tripadvisor is a travel company that assists its customers in finding the best rates for their hotel stay as well as booking tickets for their trip. One of the services it offers is their comprehensive hotel booking suite which enables its users to not only view hotels based on location, cost, cleanliness, and various other factors but also review the stay of other travelers at those hotels. The review left by the user might include more details about services which they might be happy or unhappy about, however the overall numerical rating does not provide any information regarding the details of those services.

Our project aims at bridging the gap between these text-based reviews using Sentimental Analysis as well as identifying certain other categories from popular words used in the review text which users have left for specific hotels. These new categories not only help the users narrow down their search for their perfect stay, but also helps the businesses to ascertain which services need to be improved to increase customer satisfaction and bring in more business into their respective hotels.

## 2 Background

We reviewed the work detailed by Hsiu-Yuan Tsao and Ming-Yi Chen in the article "The asymmetric effect of review valence on numerical rating"[1] and the work detailed by Bin Fang and others in the article "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics"[2], where the authors have conducted a sentiment analysis via text mining, using self-developed computer programs to retrieve a data set from the Tripadvisor website. This study finds there is an asymmetric relationship between review valence or the verbal review text and numerical rating.

Therefore, the overall rating that is provided to a hotel is not a reliable measure of services offered by a specific hotel branch or customer satisfaction. The authors mention that assumption verbal review text is symmetrically related to the numerical rating might be a false one, since brand image is a significant factor that customers consider while writing these reviews on Tripadvisor. Similarly, other factors or services offered by a specific hotel might not be considered while providing their independent overall rating to the hotel. The authors further conclude that marketers

could adopt sentiment analysis via text mining of online reviews as a valid measure or predictor of consumer satisfaction or numerical ratings. Strong brands should direct more attention to negative reviews, because in such reviews the negative impact transcends the positive. In contrast, weak brands should aim to exploit as many positive reviews as possible to minimize the impact of any negative reviews.

We noted that part the "Brand Image" of the Hotel is just one of the factors that might affect the Review valence and overall rating. Other factors would include the services offered by the specific Hotel Branch, such as the quality of food and dining services, gym and fitness services, staff politeness, etc. All these keywords can be identified, and a sentiment analysis would provide us with more insights as to whether the customer reviewing the hotel had a positive or negative experience on these specific factors. This might in turn help us to bridge the gap between the review valence and the overall rating provided by Tripadvisor.

## 3 Data

In order to extract data from Tripadvisor, we will be implementing web scraping using Selenium. The bracket for number of hotels to scan is restricted to 6 hotels, and for each hotel we will be scraping 20 review pages where each page constitutes of 10 different reviews. The fields that we are extracting from the reviews are name of the hotel, overall ratings, number of reviews, username of reviewer, review date, no. of contributions, no. of votes review received, reviewer's overall ratings, review title, review text, date of stay, individual category ratings (review, value, rooms, location, clean, service, sleep).

## 4 Methodology and Evaluation

We will be designing a multi-class sentiment analyzer similar to the work done for on app reviews[3]. Here we will predict the values of the individual category ratings such as review, value, rooms, location, clean, service, sleep based on the reviews provided by the users. To be able to achieve this, we will use a model as depicted in the picture below.

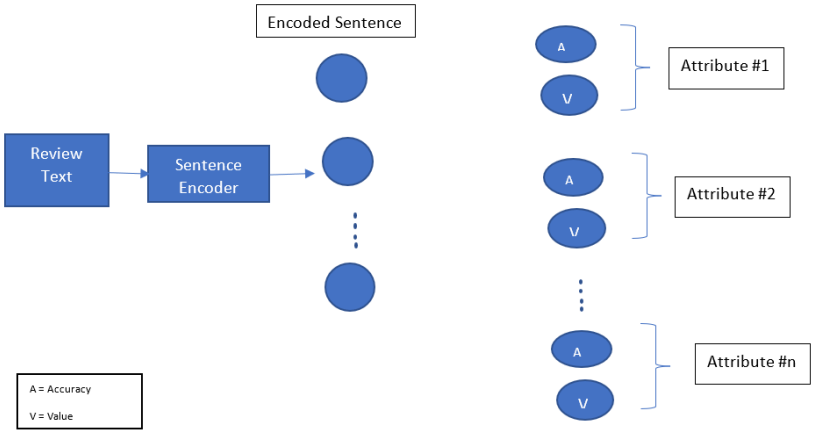


Figure 1: The rough design of the architecture

Here as shown in the picture, the model consists of encoder which would include a variant of transformer[4] model, the sentence would be encoded. Using the encoded review, we would predict the accuracy of the rating and value of the rating for each attribute respectively. The reason for predicting the accuracy of the rating is that because we cannot expect the text review to contain details about all the attributes such as some of user might have only given sleep rating and location rating while other attributes like cleanliness rating might not be given and thus, we will make predictions only for those mentioned.

Based on the works of Nikolaos Korfiatisa and others in the paper[5], we will be comparing the model predictions to the reviews who have the helpful votes and evaluate the prediction based on that.

## References

[1] Hsiu-Yuan Tsao, Ming-Yi Chen, Hao-Chiang Lin, and Yu-Chun Ma. The asymmetric effect of review valence on numerical rating: A viewpoint from a sentiment analysis of users of tripadvisor. *Online Information Review*, 43, 09 2018. doi: 10.1108/OIR-11-2017-0307.

[2] Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*, 52:498–506, 2016. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2015.07.018. URL https://www.sciencedirect.com/science/article/pii/S0261517715001715.

[3] Xian Fan, Xiaoge Li, Feihong Du, Xin Li, and Mian Wei. Apply word vectors for sentiment analysis of app reviews. pages 1062–1066, 11 2016. doi: 10.1109/ICSAI.2016.7811108.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

[5] Nikolaos Korfiatis, Elena García-Bariocanal, and Salvador Sánchez-Alonso. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217, 2012. ISSN 1567-4223. doi: https://doi.org/10.1016/j.elerap.2011.10.003. URL https://www.sciencedirect.com/science/article/pii/S1567422311000639.