Text Sentiment Analysis Prediction Using Disneyland Reviews

David Berberena

Bellevue University

DSC 680 Applied Data Science

Amirfarrokh Iranitalab

October 27, 2024

Text Sentiment Analysis Prediction Using Disneyland Reviews

Topic

To better align my study of data science with my company of choice to work for and showcase the skillset I have learned using data involving the company, I will perform a text sentiment analysis study on reviews concerning the guest experience at three of the six Disneyland parks (Disneyland California, Disneyland Paris, and Disneyland Hong Kong as outlined in the forthcoming data) under the Walt Disney Company. With Disney being a market leader in the entertainment industry and theme park scene, analyzing the guest sentiment of their time at Disney parks across the world can help the Company understand and identify guest wants, concerns, and areas of opportunity to increase the guest experience and ultimately retain the maximum number of guests for as long as possible.

Business Problem

The Walt Disney Company has grown exponentially from its humble start as an animation studio to what may very well be the most common household name in the United States regarding the entertainment industry. The Company comprises various parts that create its economic fingerprint: Parks and Resorts, Media Networks, Studio Entertainment, and Consumer Products. The Parks and Resorts sector of the Company has been facing recent customer retention issues due to many factors and business changes, so understanding the guests' sentiments regarding their experience at Disneyland parks across the globe better positions the Company to be proactive rather than reactive to guest commentary. Review sentiment can typically be classified as either positive, neutral, or negative, and the automation

of such classification to drive decision-making geared towards the improvement of the guest experience will help the Walt Disney Company abate the recently waning attendance at their theme parks. Predicting customer sentiment through their reviews will also strengthen the relationship between the Company and the guest.

Datasets

After much searching, I found a dataset on Kaggle that contains over 42,000 TripAdvisor reviews from visitors of the Disney parks in California, Paris, and Hong Kong. The elements of the dataset are as follows:

1. Review_ID: the unique ID given to each review for anonymity and data privacy
2. Rating: the reviewer's rating of their Disneyland experience ranging from 1 (unsatisfied) to 5 (satisfied)
3. Year_Month: the month and year when the reviewer visited the theme park
4. Reviewer_Location: the country of origin of the visitor
5. Review_Text: the review/comment made by the visitor regarding their Disneyland trip
6. Disneyland_Branch: the location of the Disneyland park the reviewer visited

Looking at the dataset variables, there are a few things that can be done to achieve the intended results while also allowing many insights to be seen from the data by way of graphical visualizations and model performance metrics that will ultimately benefit the Walt Disney Company and guest satisfaction. The key variable is review sentiment, which will be derived from the Rating variable present in the dataset.

Methods

Initial analysis of the business problem and the data prompts exploratory data analysis of the dataset. I plan to clean and preprocess the dataset's Review_Text column using various techniques, including removing stopwords, applying port stemming, removing special characters, and converting the text to vectors through Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. I will also convert the Rating variable into the outcome variable of review sentiment by way of setting each rating equal to a sentiment (4-5 is positive, 3 is neutral, 1-2 is negative). Visualizations relating to the review sentiment for each Disneyland park would provide region-specific insight, allowing underlying trends to be seen. Other visualizations such as word clouds and charts outlining the number of reviews per sentiment category would help the Walt Disney Company see what areas of the guests' park experience should be focused on. To craft a predictive model in the effort of correctly classifying review sentiment, the data would then be split at an 80/20 ratio and be subject to multiple classification machine learning models such as Naïve Bayes, Logistic Regression, Random Forest, and Gradient Boosting. Whether all of these models will be used depends on the accuracy by which the previously crafted models predict the review sentiment. To make sure the models are being crafted with the best performance in mind, a grid search with cross-validation hyperparameters will be implemented. The performance metrics for such a classification problem will be the accuracy statistic, the F1 score, the precision score, and the recall score. Along with these values will be a confusion matrix that will visually explain the model's ability to predict true positives and true negatives against any false positives or negatives.

Ethical Considerations

Ethics are always a concern when dealing with data; this research is no different. An apparent ethical concern is that I currently work for the Company, thereby asserting some form of bias to be recognized and set aside to be an impartial viewer of the data to report results that are not obscured by my direct affiliation with the Company. Adding to this is the privacy of the users involved in the reviews featured within the dataset. While their identity is to be protected and/or anonymized for data protection purposes, it seems that only some measures have been taken, and more could be done to prioritize the privacy of the review creators. While the variable outlining the reviewer's location could provide definitive insight as to the sentiment of reviewers in a specific region of the world, the reviewers could potentially be reidentified by combining the location, the review contents, and the year/month it was published.

Another data ethics discussion relating to this study is the fact that these reviews came from TripAdvisor, which brings into question the website's data usage policies and whether the reviewers had been notified that their reviews could be used for educational research like this. TripAdvisor is one of the most trusted hospitality/tourism websites out there, and for them to not tell their users that any reviews posted could be used for market research would be a big concern. Checking TripAdvisor's policies as they relate to the scraping and usage of their data for research purposes would mitigate that concern and allow the study to progress towards a level mimicking real-world application.

Challenges/Issues

What I believe may be a challenge is explaining to an audience the process by which sentiment analysis works and how it benefits the Walt Disney Company and ultimately the guest. After a certain point in data science, the jargon becomes muddled when elaborating to those who are part of the general public and can be hard to properly convey without losing the audience's attention. Regarding the data, I believe the most daunting challenge is seeing how the predictive models will handle the conversion of the Rating variable scaled 1 through 5 to the review sentiment variable with positive, neutral, or negative classes. I believe the problem may lie with the neutral classification being converted from the 3 rating, as the strength of words lying within these seemingly neutral reviews may have the reviews' sentiment leaning one way or the other, decreasing the predictive accuracy of a predictive model. This incorrect classification can potentially stem from an underlying bias in the data as well, maybe through the reviewers' locations or the Disneyland park about which the review has been created. These issues and challenges will be taken into consideration when undertaking the research for the sake of the Walt Disney Company better understanding their guests' preferences.

References

In addition to the Kaggle dataset found which holds the data I am to work with in this project, I have also located an article providing insight as to how the Walt Disney Company would benefit from utilizing text sentiment analysis. The article links are below:

1. https://www.kaggle.com/datasets/arushchillar/disneyland-reviews
2. https://determ.com/blog/top-5-benefits-of-sentiment-analysis-for-businesses/