Using NLP to Find the Most Predictive Words of Hotel Reviews

Tanuj Chawla

Introduction

- Overview of the hospitality industry
- Importance of customer satisfaction and loyalty
- The role of customer reviews in driving business success

Data Source: DataFiniti

- Number of hotels: 1736
- Number of reviews: 10,000
- Rating Scales: 1 - 5, with some in decimals.

Audience

- Hotels, businesses, restaurants

Key Questions

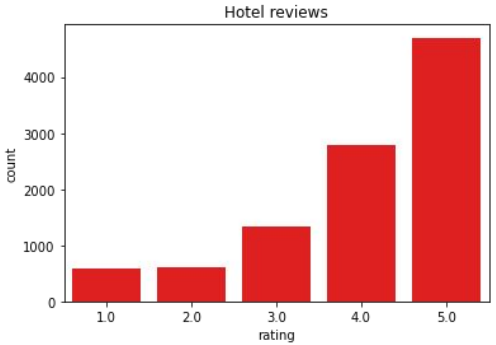
- What are the key factors that customers love/ hate about our hotel
- Where are some areas of improvement that we can identify?
- How can we improve customer satisfaction using this model?

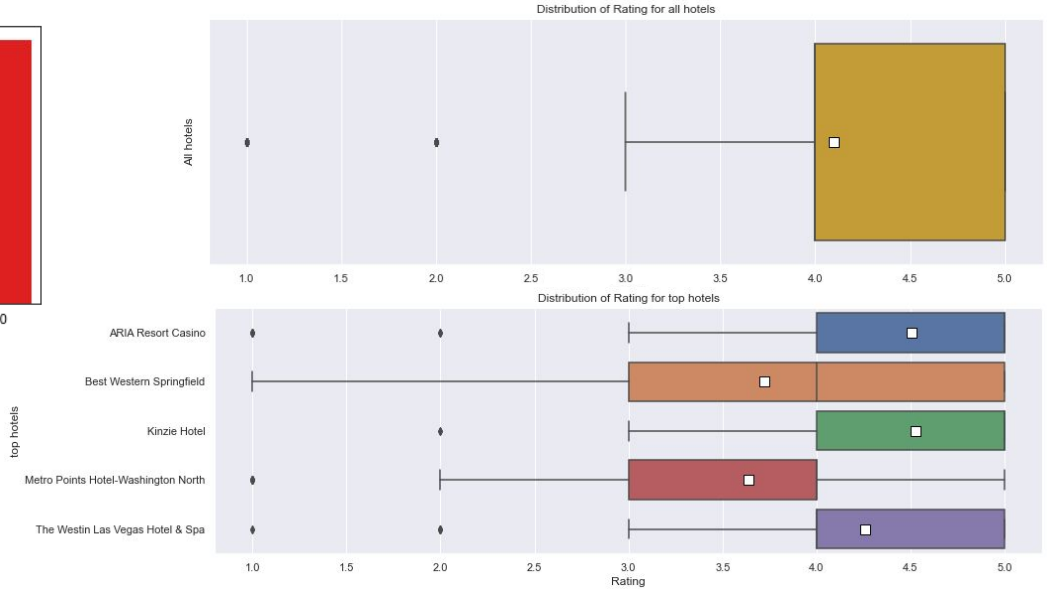
Data Wrangling

1. Imbalance rating scale:
    a. Scale was 1-5 including decimals. Rounded to nearest whole number up and down creating a 5 point scale. Reviews primarily 4 and 5 star. Then created Binary rating scale for further exploration.
2. Text Preprocessing for clean text reviews
    a. Rounded review ratings up and down to the appropriate whole number.
    b. Dropped unnecessary columns and kept review text, rating, and hotel name.
    c. Fixed spelling errors
    d. Broke down text reviews in preparation for sentiment analysis
    e. Added a new feature: word_count & word_count_title
    f. Extracted five most reviewed hotels to examine along with the entire dataset: Metro Points Hotel-Washington North ,The Westin Las Vegas Hotel & Spa, Best Western Springfield , ARIA Resort Casino, Kinzie Hotel
    g. Used Lang Detect to filter only english reviews.
    h. Started with a dataset size of 10,000 rows and 25 columns. After cleaning and removing data, it was reduced to 9998 rows and 3 columns.

Exploratory Data Analysis

Ratings: Scale was 1-5 including decimals. Rounded to nearest whole number up and down creating a 5 point scale. Reviews primarily 4 and 5 star. Then created Binary rating scale for further exploration.

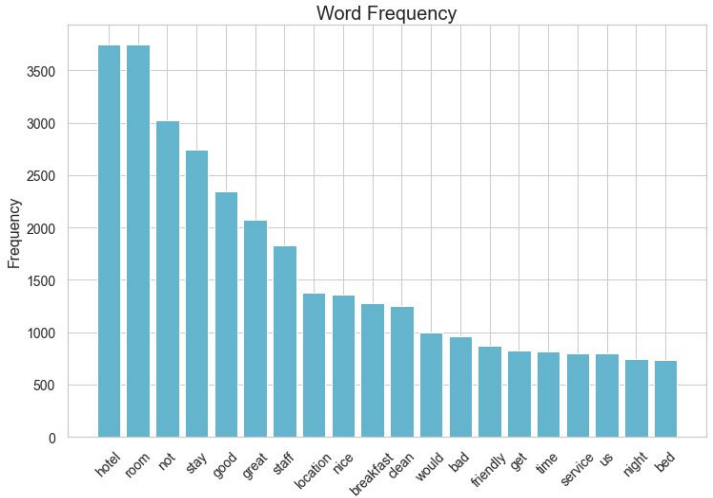




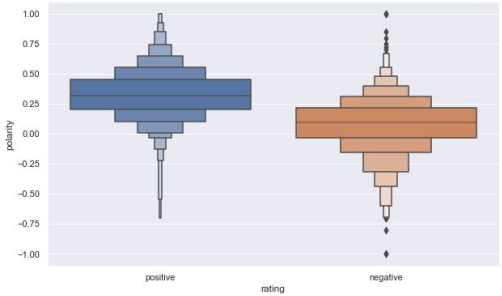
EDA (cont.)

20 Most Frequent Words

20 most abundant words in the preprocessed text corpora. Words like hotel room, great staff, nice breakfast, and others are good features of a hotel. We can also see some negative words as well as some noise.



Reviews



Pre-processing & Training

Vectorizers

1. **CountVectorizer:** uses a "bag-of-words" approach where the text is analyzed by word counts. It counts the occurrence of each token for each review.

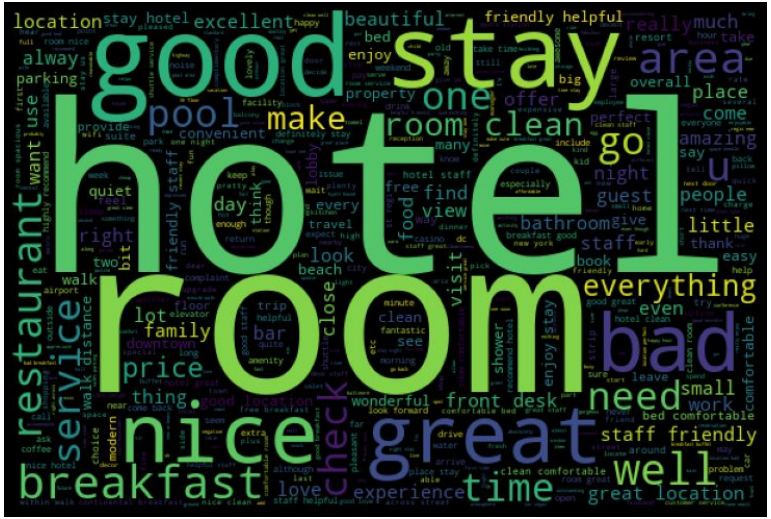
2. **TfidfVectorizer:** uses the term frequency(tf) and the inverse document frequency(idf) to create weighted term frequencies.

Vectorizer	ROC-AUC	Best Parameters
CountVectorizer	0.883	min_df=1, alpha=1, fit_prior=True
TfidfVectorizer	0.835	min_df=1, alpha=1, fit_prior=True
CountVec w/ GridSearch	0.9101	min_df=.001, alpha=1, fit_prior=True
TfidfVec w/ GridSearch	0.9144	min_df=.001, alpha=1, fit_prior=True

Machine Learning (Cont..)

Most Predictive Words - Positive Reviews

Predictive Features: Words	
High Words	**P(high | word)**
definitely stay	0.95
great hotel	0.95
would definitely	0.95
friendly staff	0.95
love	0.94
excellent	0.94
great stay	0.94
wonderful	0.93
enjoy	0.93
amazing	0.93
Low Words	**P(low | word)**
smell	0.34
horrible	0.33
stain	0.33
terrible	0.33
hair	0.33
hot water	0.30
paper	0.29
dirty	0.22
call front desk	0.20
call front	0.20

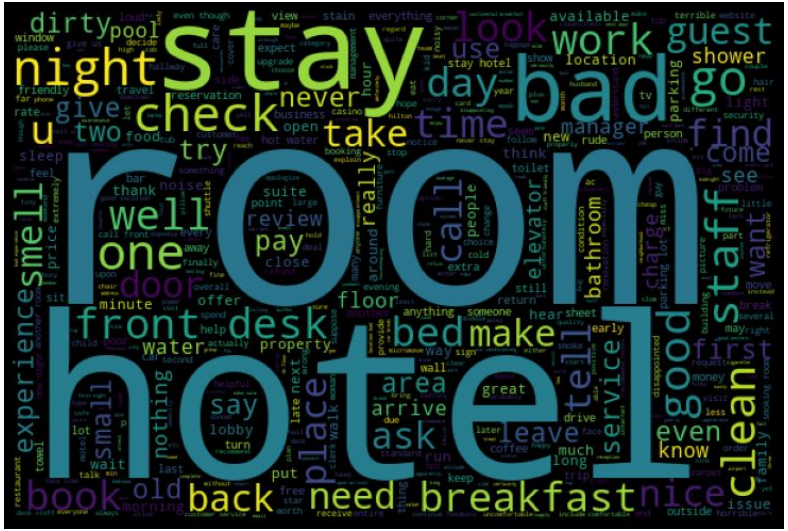


Machine Learning (Cont..)

Most Predictive Words - Negative Reviews

Predictive Features: Words

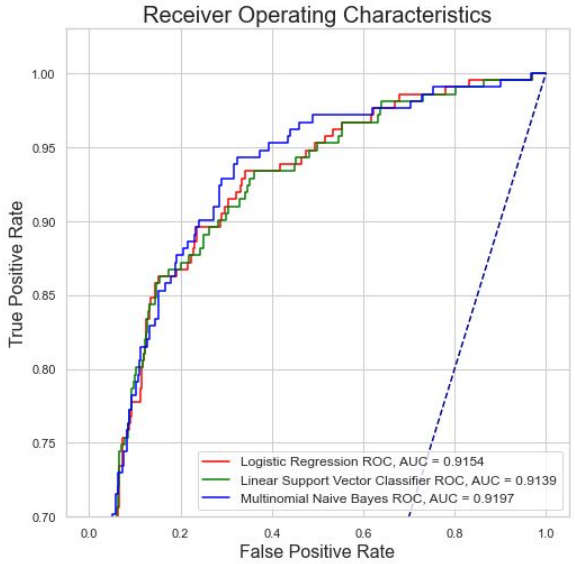
High Words	P(high | word)
call front	0.75
dirty	0.75
bug	0.69
refuse	0.69
musty	0.68
paper	0.66
hot water	0.66
decent hotel	0.66
leak	0.66
outdated	0.64
Low Words	**P(low | word)**
amazing	0.06
wonderful	0.06
great stay	0.06
enjoy	0.06
would definitely	0.05
love	0.05
great hotel	0.05
friendly staff	0.05
excellent	0.05
definitely stay	0.05



Machine Learning (Cont..)

Classifiers

Classifier	ROC-AUC	Best Parameters
MultinomialNB	0.9197	alpha=0.15, fit_prior=True
LogisticRegressionCV	0.9154	C=1, max_iter=1000
Linear SVC	0.9139	C=0.1



Conclusion

The key findings of our project on analyzing hotel reviews using natural language processing (NLP) were as follows:

- Amenities, location, and staff service were among the most highly rated features of a hotel, as identified by customer reviews.
- Cleanliness and value for money were identified as areas for improvement, based on customer feedback.
- NLP techniques, including sentiment analysis, were used to classify reviews as positive, or negative, providing an overview of the overall sentiment of guests.

These findings have important implications for the hospitality industry, as they provide insight into the key factors that influence customer satisfaction and loyalty. By understanding the preferences and needs of their customers, hotels can implement targeted strategies to improve the guest experience and drive business success.