Tripadvisor

TripAdvisor Hotel Reviews

Exploratory Data Analysis
And
Hotel Rating Prediction

Table of Content

Data consists of 20K Hotel Reviews ranged from 1-5 stars

- Overview
- Load the dataset
- Data Visualization
- Data Preprocessing
- Build Machine Learning Models
- Oversample the data
- Build Machine Learning Models after oversampling
- Compare different models
- Pick the model with highest Accuracy Score

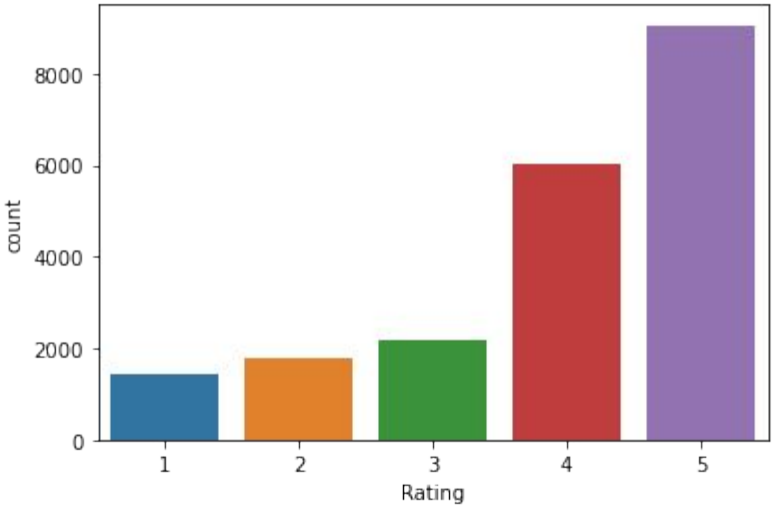
Data Preprocessing

The first step in building machine learning model is to preprocess the texts, which includes:

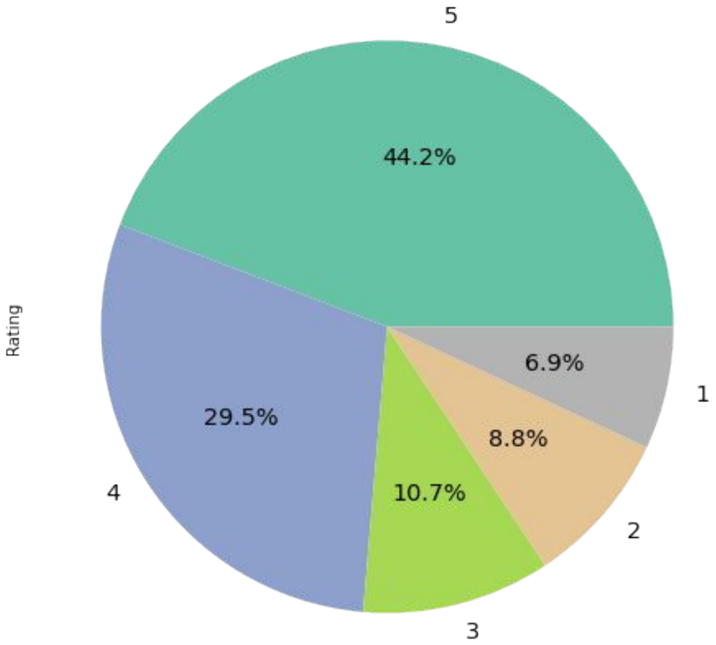
- Data Cleaning
- Transform all the words in lowercase
- Remove Punctuation
- Remove Stop Words
- Convert words into meaningful root
- Convert raw data to matrix of TF_IDF

Data Visualizations

Each category count: We can see from both charts that the number of 5* Feedback are the highest and 1* Feedback are the lowest

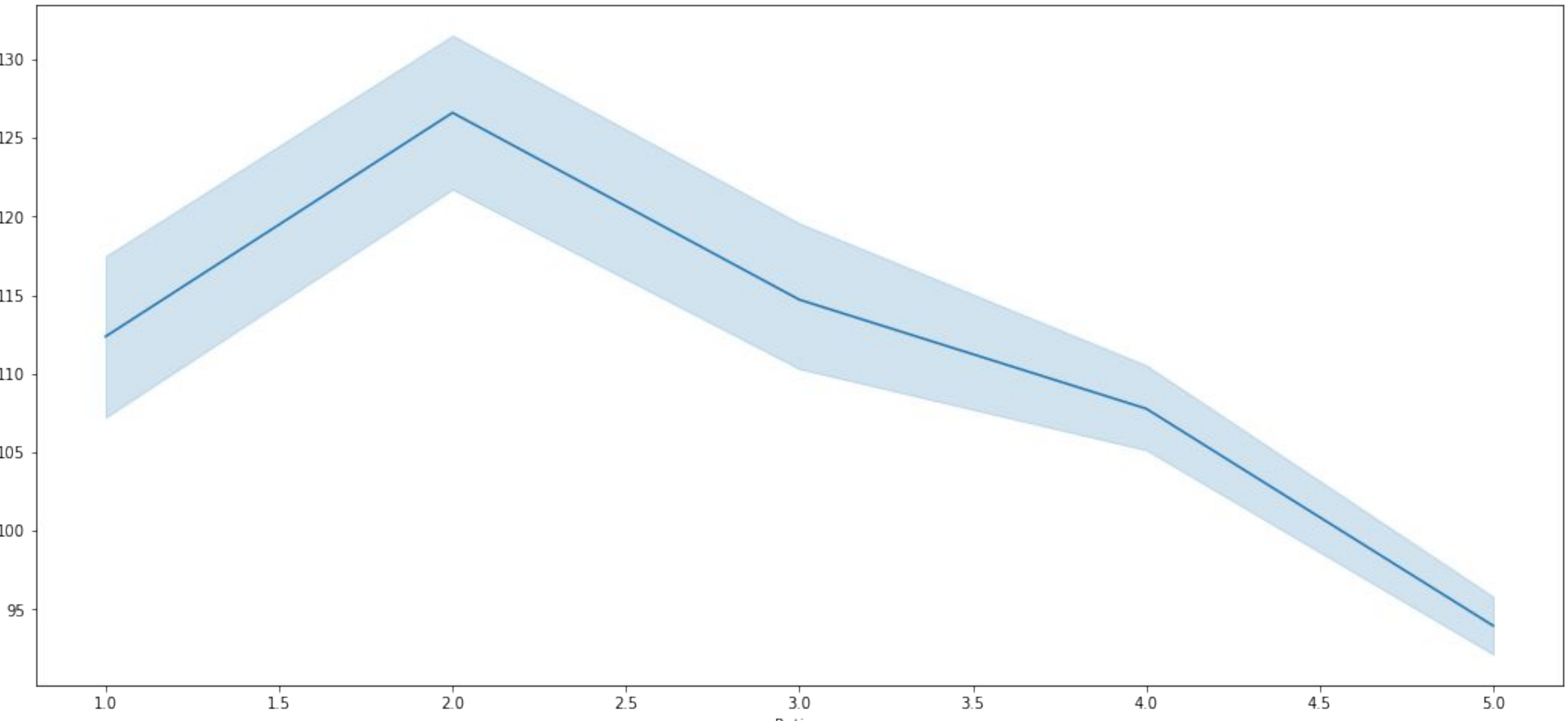


Percentage of Feedback per Category



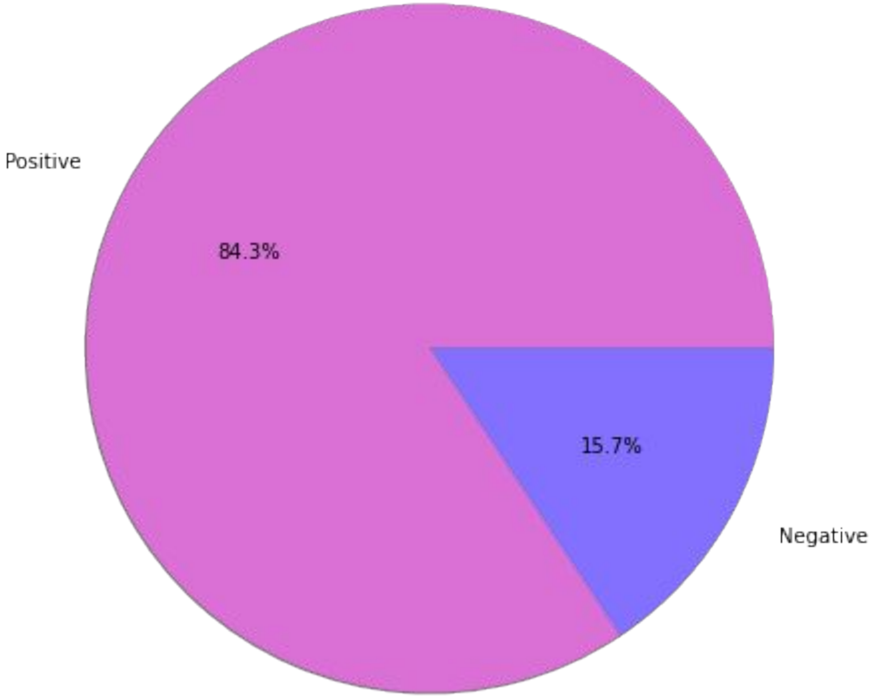
Line Plot of the word-count Across Rating Category

We can see that number of words used in 2 Star reviews are highest amongst all and it decreases when the star rating goes higher



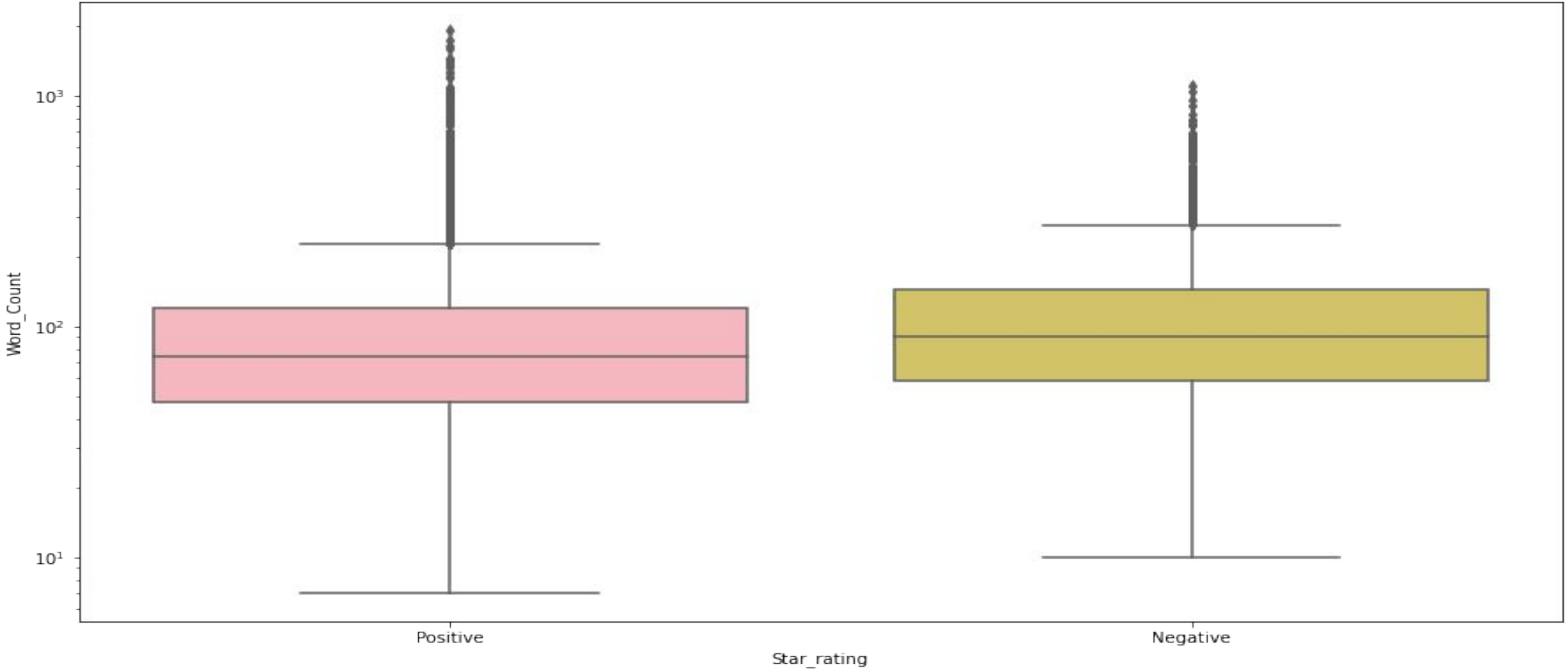
Total Number of Positive vs. Negative Reviews

We can see that our data is highly imbalanced (data is not evenly distributed within classes), which in many cases will cause poor performance of Machine Learning model



Average Word Count of Positive vs. Negative Feedbacks

The plot shows that Negative Reviews on average have higher Word Count than the Positive ones



Building Machine Learning Model

I used four different Machine Learning models for predicting the reviews, since the data was highly imbalanced, I first trained the model with original imbalanced data and used SMOT (oversampling Method to make data balanced) to see if there will be any changes in the performance of our models. It can be seen that two of the models show improvement and two stayed unchanged.

Model Performance Comparison Before and After Oversampling

Machine Learning Algorithm	Accuracy Score	Accuracy Score (After Oversampling)	Change
Logistic Regression	95%	95%	Unchanged
Decision Tree Classifier	85%	89%	4+ ↑
Random Forest Classifier	87%	94%	7+ ↑
XGBoost	91%	91%	Unchanged

Conclusion

Our best performing model is Logistic Regression with 95% accuracy, the second is XGBoost with 91% accuracy rate. We can see that oversampling did not help these two models perform better, however accuracy rate for the the Decision Tree Classifier and Random Forest model improved few points. The Random Forest model is second best with accuracy rate of 94% (after resampling).