

Milestone 1: Project Proposal

Felipe Rodriguez

Bellevue University

DSC 680 Applied Data Science

Professor Amirfarrokh Iranitalab

May 12, 2024

Topic

Sentiment Analysis of Boston Airbnb Reviews

Business Problem

The goal of this project is to analyze the sentiment of Airbnb review in Boston to understand costumer's opinions towards Airbnb listings. By analyzing this, we aim to identify the positive and negative aspects of a listing. This project will allow Airbnb hosts to understand their reviews so they can make enhancements and improve their services.

Datasets

The data is being obtained from Kaggle and is sourced from Airbnb. The data contains the following information: listing_id, id, date, reviewer_id, reviewer_name, and comments. The data will allow for sentiment analysis based on the features provided. The main feature that will be used for modeling and predictions is the field comment. The data can also be broken down by year as well.

Methods

The methods used in this project will include various visualizations to display the sentiment of the listings. Some of these views can be broken down by year or month. Prior to creating the model, the text will need to be preprocessed. This will includes removing stop words and creating sentiment scores for each review. After this, a model can be created. Using Naive Bayes, a model will be created to predict the sentiment of new reviews. To evaluate the model's

performance, by a classification report with precision, recall, and F1-score for each sentiment class will be created.

Ethical Considerations

Some ethical considerations with this project are to protect privacy. This consideration comes primarily from the use of reviews by consumers. The reviews should be anonymized to protect consumer privacy. When conducting sentiment analysis, another thing to consider is to ensure that the model does not discriminate against certain groups of people or listings.

Challenges/Issues

There are potential challenges that can occur in the initial phases of data cleansing and when conducting the model. The data needs to be preprocessed and there could be some issues when removing stop words, formatting, and determining sentiment. When creating the model, there needs to be different type of reviews and sentiment to ensure model accuracy.

References

The data is being source from Airbnb and contains reviews of listings in Boston. Using this set, a model can be created to forecast the weather. One thing to note is that the data contains historic reviews from 2009 to 2016. By splitting this data into testing and training sets, we can confirm the accuracy of the model. The model can be confirmed with accuracy using F1 score as well.