Final Project Report

Sentiment analysis of hotel reviews

Instructor:
Stephen Kunath

Reported by:
Changhao Ying
Zixuan Huang
Ying Wang
Chang Che

December 10, 2020

Introduction

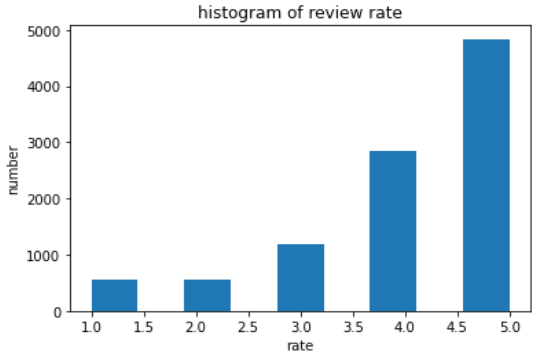
Sentiment analysis is part of the Natural Language Processing (NLP) technique that consists of extracting emotions related to some raw texts. The goal of this project is to show how sentiment analysis can be performed using Python. We will use the hotel reviews dataset to experiment with sentiment scoring and other natural language processing techniques. One of the most recognized research issues in the service industry is predicting customer behavior or understanding customer intent, which is particularly important for the hospitality and tourism industry[2][4].In the early stage, Sentimental analysis mostly aimed to English content[3], now it could apply to multiple languages[1][5].

Description of dataset

The data was derived from https://www.kaggle.com/datafiniti/hotel-reviews. This dataset is a list of about 2,000 hotels and 10,000 reviews from Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more. Each observation consists of one customer review for one hotel. Each customer review is composed of textual feedback of the customer's experience at the hotel and an overall rating[4][5]. In our project, we will mainly use the review text, review title, and rate and eliminate other unrelated features.

Data preprocessing and Exploratory Data Analysis

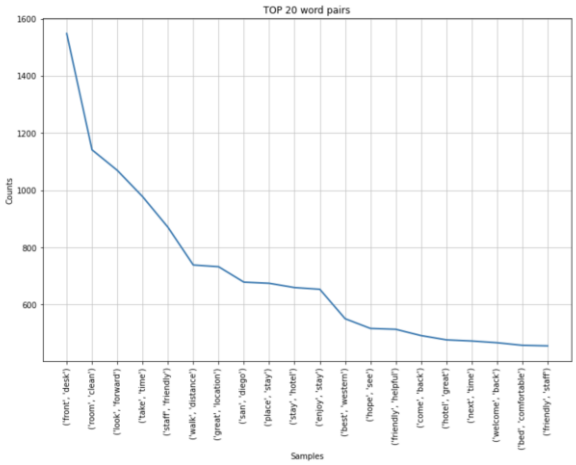
First, we did some exploratory data analysis and had an overview of the relationship between reviews and corresponding rates. We merge review text and review titles into one column called 'review' and print the best 10 reviews with the highest score and the worst 10 reviews with the lowest score. We also plot (Graph 1)a histogram to see the distribution of rates.



Graph 1. Distribution of review rates

Graph 2. Word cloud

Second, we cleaned the data, including lowering text, tokenizing text, removing punctuation, stop words, numbers, empty tokens, letters, and lemmatizing text. Graph 1 is the result of the top 20 most frequently tokenized pair words and their frequency.Graph 3 describes the word cloud after tokenization of text and title.



Graph 3. Top 20 token pairs

Third, we used SentimentIntensityAnalyzer from NLTK library to analyze the sentiment in the review texts and generated four columns which are neg, neu, pos and compound. We also generated two new features which are the number of characters and number of words in each review.

Finally, we add the TF-IDF (Term Frequency - Inverse Document Frequency) values for every word and every document. TF computes the classic number of times the word appears in the text IDF computes the relative importance of this word which depends on how many texts the word can be found. We add TF-IDF columns for every word that appears in at least 10 different texts to filter some of them and reduce the size of the final output. Table 1 shows the final data we used for modeling.

Table 1. Final data

	review	rate	clean	neg	neu	pos	compound	...	word_yourself	word_yr	word_yummy	word_zaza	word_zephyr	word_zero	word_zoo
0	This hotel was nice and quiet. Did not know, t...	3	hotel nice quiet know train track near train p...	0.000	0.851	0.149	0.7906	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	We stayed in the king suite with the separatio...	4	stay king suite separation bedroom live space ...	0.070	0.812	0.118	0.4157	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Parking was horrible, somebody ran into my ren...	3	park horrible somebody run rental car stay get...	0.063	0.880	0.057	-0.0772	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Modeling

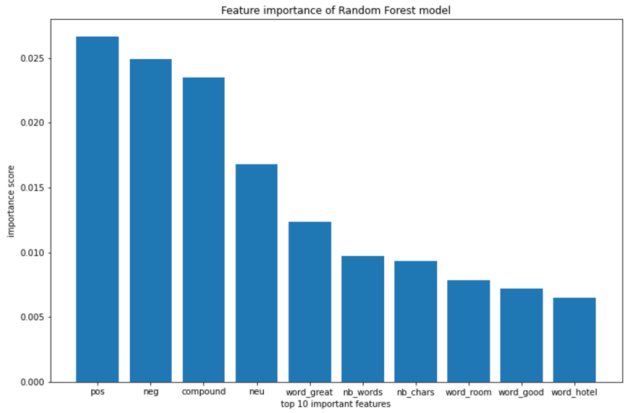
After generating these new features, we built several machine learning and deep learning models to classify the review texts and predict the review rates. This is a classification problem, the target is the rate (from 1 to 5) that each customer provided before. Based on the classification report and confusion matrix, the models that we built for estimation includes: decision tree, random forest, adaboost,LDA,QDA, and MLP.The accuracy of methods is summarized in the table 2.

Table 2. Summary of accuracy of different methods

Method	Decision tree	random forest	adaboost	LDA	QDA	MLP
Accuracy	0.463	0.562	0.531	0.518	0.424	0.458

Summary

According to the results, we can see the Random forest has the highest accuracy which is 0.562. We plot the top 10 important features for the random forest model.



Graph 4. Top 10 important features of Random Forest model

Reference

[1] Demir, Durmaz. “Sentiment Analysis for Hotel Attributes from Online Reviews.” *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020. 1–4. Web.
[2] Farisi, Sibaroni. “Sentiment Analysis on Hotel Reviews Using Multinomial Naïve Bayes Classifier.” *Journal of physics. Conference series* 1192 (2019): 12024–. Web.
[3] Geetha, Singha. “Relationship Between Customer Sentiment and Online Customer Ratings for Hotels - An Empirical Analysis.” *Tourism management (1982)* 61 (2017): 43–54. Web.
[4] Park, Kang. “Understanding Customers’ Hotel Revisiting Behaviour: a Sentiment Analysis of Online Feedback Reviews.” *Current issues in tourism* 23.5 (2018): 605–611. Web.
[5] Sodanil, Maleerat. “Multi-Language Sentiment Analysis for Hotel Reviews.” *MATEC web of conferences* 75 (2016): 3002–. Web.