Ryan French, Shiwen Wang, Yuan Yao, Rachna Shetty

IST 718

Professor Willard Williamson

9-7-2019

**Yelp Review Classification – Project Abstract**

Through the use of the Yelp Dataset available on Kaggle.com, our group hopes to evaluate the reviews present in the data in order to create an algorithm capable of predicting highly positive customer reviews (defined as a four or higher out of five). Through understanding the most instrumental features of this model, we will enable businesses to better implement measures to boost their reviews as they see fit.

Originally uploaded as part of the Yelp Dataset Challenge, an endeavor for Yelp to allow students to perform analysis and research with their data, the dataset consists of 5,200,000 user reviews, information on 174,000 businesses, and was collected from 11 metropolitan areas. This information is housed within seven CSV files which collectively are about 4 GB in size.

In addition to our goal of building an effective classifier for reviews using Machine Learning algorithms, we also hope to answer a variety of other questions such as:

* Which users tend to provide the most ratings? Are their ratings generally skewed in one direction? Do top contributors tend to travel or stay in one location?
* How much does average business review score deviate across each of the areas that data was collected?
* Which categories of businesses (restaurants, stores, nightlife, etc) tend to receive the highest average scores?

**Dataset link:** <https://www.kaggle.com/yelp-dataset/yelp-dataset>