DSC495 Capstone Proposal - Yelp Review Analysis

Jake Mavrides 11/2/2023

What is your data source and how will you access it? (Q3)

I will be using the Yelp reviews dataset for my capstone project. The original dataset off of Yelp's website was 8+ gigabytes, so I found a subsetted version of the original dataset on Kaggle which I will use. The data is subsetted from Yelp's website directly, where they publish reviews on their website to be used by others. The dataset has the following columns:

* Column 1 - Unique Business ID
* Column 2 - Date of Review
* Column 3 - Review ID
* Column 4 - Stars given by the user (1-5)
* Column 5 - Review given by the user
* Column 6 - Type of text entered - Review
* Column 7 - Unique User ID
* Column 8 - Cool column: The number of cool votes the review received
* Column 9 - Useful column: The number of useful votes the review received
* Column 10 - Funny Column: The number of funny votes the review received

Link: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset. (Note: after further investigation this particular link is also multiple files of a few gigabytes each, so I might use a different subset, but it will be taken from the same dataset nonetheless).

What is the problem you are attempting to solve, why is it useful? (Q1&2)

As stated, there are millions and millions of Yelp reviews publicly available. According to Yelp's website, the archive currently has around 7 million reviews!

Many reviewers will leave either 1 star or 5 star reviews, and 2, 3, and 4 star reviews are much less common. Additionally, how can we tell the difference between a 1 star and 2 star rating? What about a 4 star and 5 star rating? We will use sentiment analysis and similar NLP methods to answer these questions and gain less-subjective insight into someone's "actual experience". From here we could create a 'true rating' for businesses, which would remove lots of the biases. We could also even do things like finding out which business types have the most or least biased reviews.

For example, say you have someone who tends to be more strict giving out good ratings. That person's review may say "best Italian food I've ever had", but they may only give 4 stars because they don't believe in giving perfect scores. The **true** rating would be estimated by our machine learning algorithm as a 5 out of 5 stars, wherein the **actual** rating given was 4 out of 5 stars.

On the other hand, someone may have had a poor experience with a business, but tends to be more forgiving and nicer with their reviews, so their **actual** rating is more likely to be a 5, even when the **true** rating (determined by our sentiment analysis algorithm) should (in theory) be lower.

What techniques will you use? (Q4)

Data cleaning will include removing stopwords, setting characters to lowercase, trimming whitespace, etc.

Some fairly basic data visualizations will be used when it helps explanations. Word jumble maps may also be used to help see the more common words in high vs low rated reviews, etc.

For the final modeling, there are have a number of options regarding NLP strategies. I will use a number of techniques including SVM and neural networks. We will compare the different models and pick whichever scores the best on our data. Towards the end of the capstone, I will create formal write ups of results and conclusions. If there seems to be a valid business aspect of the project I will expand on how a business may implement my results.

Challenges (Q5)

I think I will learn some newer concepts I have seen online used for sentiment analysis, such as Naive-Bayes or random forest, so this will be my biggest challenge. Another challenge I could see coming up would be cleaning the data efficiently without losing important words, or trying to fairly judge the accuracy of different models.