Analyzing the Yelp dataset

Michael Backs and Daniel Baron

December 20, 2019

Abstract

Yelp is a web platform to collect and present user ratings and suggestions for restaurants and other locations such as bars, clubs and various other places. Yelp releases a dump of their dataset once a year to allow students to analyze it. This paper is about analyzing the Yelp dataset to find funny reviews and classying them into four funniness categories.

Introduction

Text classification is applied in a number of fields and situations. Jurafsky and Martin (2019) give marketing or politics as an example for fields that wish to classify a text or product as positive or negative. They also say that text classification is often used to identify spam emails or find the language a given text is written in (ebd) While sentiment analysis and spam detection tasks use binary classification, we will pursue a multiple classification approach in our project. Our aim is to predict a 'funniness' category for a given review realeased on the web platform Yelp. These reviews have a number of funny votes ranging from zero to roughly 20. Assuming that the number of funny votes reflects how funny a text is, we will train a model to classify a given text in terms of funniness.

Iacob and Deniz (2019) used the same dataset to inspect a similar aspect in their Kaggle Notebook. They are splitting the reviews into positive and negative ones and train a decision tree to figure out if a review is rather positive or negative. By leaving out 3-star-reviews, they don't use any neutral reviews: "It would make sense to associate 4- and 5-star reviews with a positive sentiment and 1- and 2-star reviews with a negative sentiment. 3-star reviews would be neutral, but for simplicity purposes, we will only attempt to predict the positive and negative sentiment, and we will revisit neutral later." Iacob and Deniz (2019)

The dataset used in the following sections is taken from Kaggle and comprises roughly 5,2 million user reviews released on Yelp Inc. and Crawford (2019). For training and testing an excerp from the data is used.

Dataset

The semi structured dataset consists of five JSON files including reviewed businesses, a record of dates and locations when reviewers were at a business, user reviews, a list tips given by users and informations abour users who wrote reviews.

We limit our first analysis to reviews given for all Illinois-based locations present in the Yelp Dataset, which results in 42.316 reviews for 1.930 businesses. To improve classification training results further data can be taken into account by other states. Stop words have been removed using Zipf's Law.

The dataset is already pre-classified by the Yelp user database, so individuals had the chance to vote for a review to be funny. 7.763 out of the extracted 42.316 reviews were rated funny at least once, which leaves 34.553 reviews up for classification.

It is undoubtedly clear that the reviews were not rated by experts but by ordinary users. Thus the ratings lack any formal or tracable criteria according to which they were rated funny.

The reduced dataset was created by preprocessing the businesses file consisting of 192.609 businesses and feeding latitude and longitude coordinates into the reverse geocoder python library written by Thampi (2016) to get a more accurate state

relation since we found the adresses and states not to match up in every case.

As Iacob and Deniz (2019) already pointed out in their notebook, the idea of building a CART model (Classification and Regression Trees) seems convincing and would probably be a good idea to follow up with. After removing stop words, the funny reviews should have some words in common that we then could search in the other reviews and see if they are funny as well. Rhyming words at close distance or colloquial expressions could hint at an amusing review. Moreover, the use of imperative sentences as a means of directly adressing the reader could further contribute to creating entertainment. Additionally, neologisms which are not typing errors might be of interest.

Methodology

As a baseline we will use a counting vector to represent each review's features to train a naive Bayes classifier. Alternatively an even easier feature representation can be employed which is extracting negative and positive statements. 90 percent of the preprocessed data is going to serve as training data while the remaining 10 percent are reserved to test the model's performance.

The precticted labels are funniness categories one to four: The first category means that a review is not funny while the second category indicates that it is slightly funny. The third and fourth categories imply that a review in question is quite funny or extemely funny respectively.

Our subjectively labeled dataset might pose a problem. Nevertheless, we assume that zero funny votes mean that a review was not deemed funny by the majority of the Yelp userbase. Surely, a zero rated review could still be funny because users just might not have cared to rate it funny or it might be the case that too few users read the review in question.

In order to improve the result that is given by the baseline classifyer model, we will apply word embeddings which take into account distribution and context of words in contrast to counting vectors' bag-of-words representation. The letter only work with a word's frequency. Word2vec might also be tested because short and dense vectors yield better results in NLP tasks according to Jurafsky and Martin (2019) . Moreover, we will draw upon more data to further improve results.

Several reviews rated funny contain rhymes, word plays, colloquial language or idioms. This observation might provide more robust features. We will look at bigrams or trigrams checking for rhyming words by comparing a word with a list of each word's possible rhyming words.

References

Iacob, Suzana, and Frederico Deniz. 2019. "Sentiment Analysis of the Yelp Reviews Data." 2019. https://www.kaggle.com/suzanaiacob/sentiment-analysis-of-the-yelp-reviews-data.

Inc., Yelp, and Chris Crawford. 2019. "Yelp Dataset." 2019. https://www.kaggle.com/yelp-dataset/yelp-dataset.

Jurafsky, Dan, and James H. Martin. 2019. *Speech and Language Processing*. 3rd edition draft. https://web.stanford.edu/~jurafsky/slp3/.

Thampi, Ajay. 2016. "Reverse-Geocoder." 2016. https://github.com/thampiman/reverse-geocoder.