Yelp Me Yelp You: Sentiment Analysis and Associated Data in Yelp Reviews

MADELEINE CROUCH, University of Colorado, Boulder
KHOA LE, University of Colorado, Boulder
DALTON MCCLAIN, University of Colorado, Boulder
MICHAEL BARLOW, University of Colorado, Boulder

ACM Reference format:
Madeleine Crouch, Khoa Le, Dalton McClain, and Michael Barlow. 2017. Yelp Me Yelp You: Sentiment Analysis and Associated Data in Yelp Reviews. 1, 1, Article 1 (April 2017), 4 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION/MOTIVATION

Yelp is a website which allows users to create an account and review businesses. Our goal is to analyze data from Yelp to gain further insight for business owners to develop a satisfactory experience for their customers. Yelp has a public dataset available on their website with information about users, businesses, and reviews.

Our first objective is to determine if certain attributes are indicative of business success with respect to their average review rating (see sec. 2.3 for revenue relation). If we can find a way to predict whether or not a business will do well later on in its operation, that information will be useful to businesses looking to improve their model.

Our second objective is to analyze the text of reviews for a sentiment. The sentiment analysis will show the emotions conveyed by the reviewer as being pleased or angry at their experience with the business. Using the results of this analysis we hope to be able to determine if certain times of day, city locations, or other factors contribute to the experience of a customer. This information could also contribute to a business model. As an example, if certain locations are more prone to angry customers it could indicate that the people in that area have high standards, the businesses are poorly run, or just that those customers tend to leave angry reviews.

Our last objective is to find recurring problems in businesses. We will analyze review text to find the specific issues that people are having with businesses. If a business has many reviews which all indicate a similar issue, then that is something a business owner would want to be aware of so that they can eliminate the problem.

2 PRIOR WORK

Since our problem involves predictions surrounding businesses, we sought out similar research to guide our investigation. Thus, we focus our literature review on studies that utilized some form of sentiment analysis, constructed predictive models, and/or conducted meta-analysis of trends in the aforementioned sentiment, rating, or business interest.

2.1 Predicting Business Attention

A previous winner, Hood et al., published a paper on assessing a business?s current rating/review state and future expectations with respect to review numbers on Yelp. The bulk of their work centered around the creation of additional features for the data (e.g. number of similar businesses within 1km, features of subsets of reviews such as average rating and number of unique users). From there they utilized various feature reduction and clustering techniques to build a prediction model. They were able to identify features that offered greater accuracy in prediction of interest in the business (calculated as the predicted number of reviews within 6 months after a target date) as well as a create model for prediction that outperformed a basic linear regression.

2.2 Latent Subtopics

Huang et al. (2013) utilized probabilistic models to discover underlying subtopics within Yelp reviews. This offered a framework to categorize the reviews based on review text. From here, they were able to analyze the prevalence of various star ratings within each category. In this way, they could advise business owners as to the common trends among good and bad reviews that may not be apparent without considering the review text. McAuley and Leskovec built on this idea by combining such latent review topics with hidden rating driving factors, allowing them construct a rating prediction model.

2.3 Reviews and Revenue

In one of the more influential papers on the subject, Michael Luca showed a correlation between Yelp rating and revenue, highlighting the fact that rounding to half stars could significantly impact two similarly-rated restaurants if their averages fell on either side of the divide (e.g. 3.24 and 3.26). He goes on to build a model of market response based on review volume as well as the impact of user expertise. He concluded that such responses are consistent with a Bayesian learning model.

3 PROPOSED WORK AND DATASET INFORMATION

3.1 Proposed Work

The dataset has many optional attributes. There is an attribute called 'attributes' that is a list of optional attributes. The dataset is inconsistent and cannot be relied upon for each business. For instance, one of these attributes is 'Bike Parking'. This attribute

DOI: 10.1145/nnnnnnn.nnnnnnn

is either true, false, or nonexistent. Speciifcally the fact that this attribute might not be available for multiple businesses makes it unreliable.

In addition, the data is not only restaurants but any kind of business. As a result they have another piece of data which is the categories list. For each business this attribute attempts to classify the business. However some categories are extremely broad and the same type of business can be described in different ways. One example is that there is a bakery, but it?s also categorized as food. Many businesses will sell food, but the category of food will not be in their list of categories and instead it will be labelled with more specific categories like Italian food.

The majority of the data cleaning process will be involved in handling missing data. It is possible to ll in these missing values with default values like false but this can be dangerous to make assumptions about what businesses have and don?t have.

The Yelp dataset contains up to one million different attributes about businesses from 11 different cities in 4 countries. The majority of these attributes will be unnecessary so there will be a large effort in data reduction. For instance there are photographs included, but our questions does not require image analysis. There also attributes such as postal codes that are unnecessary because latitude and longitude is more accurate.

3.2 Dataset

The Yelp dataset comes from the Yelp Dataset Challenge. This is an ongoing challenge and this is the ninth round. The challenge offers $5,000 to the winner and the final submission is an academic paper.

We will be considering the following from the Yelp dataset:

- 4.1 million reviews

```
"review_id":(encrypted review id),
"user_id":(encrypted user id),
"business_id":(encrypted business id)"
"stars":(star rating, rounded to half-stars),
"date":(date, formatted: YYYY-MM-DD),
"text":"review text",
"useful":('useful' vote count),
"funny":('funny' vote count),
"cool": ('cool' vote count),
"type": "review"
```

- Information on 144,000 businesses in 11 cities from 4 countries (incl. the US, Germany, Canada, and the U.K. (with 1.1 million business attributes)

```
"business_id":(encrypted business id),
"name":"business name",
"neighborhood": "hood name",
"address":"full address",
"city":"city",
"state":"state"(if applicable),
"postal code":"postal code",
"latitude":latitude,
"longitude":longitude,
"stars":(star rating, rounded to half-stars),
"review_count":(number of reviews),
"is_open":0/1 (closed/open),
"attributes":[(localized attribute tags)],
"categories":[(localized category names)],
"hours":[(hours strings)],
"type": "business"
```

- Over 1M users

```
"user_id":(encrypted user id),
"name":"first name",
"review_count":(number of reviews),
"yelping_since": (date, formatted: YYYY-MM-DD),
"friends":[(encrypted user ids)],
"useful":('useful' vote count sent by user),
"funny":('funny' vote count sent by user),
"cool": ('cool' vote count sent by user),
"fans":"number of fans the user has",
"elite":["an array of years the user was elite"],
"average_stars":floating point average like 4.31,
"compliment_type": (compliment count),
...(same for each compliment type)...
"type":"user"
```

- Over 200,000 user-submitted photographs, 947,000 tips, and aggregated check-in data

4 METHODS AND TOOLS

The tools and methods used for this project will include a database for storing and integrating the data, libraries and modules for cleaning and preprocessing, and various tools for transformation, mining, pattern evaluation, and knowledge presentation. Given the json-formatted data, we've opted for using a NoSQL database like MongoDB or Cassandra. Fortunately, there are a number of options, considering our educational status, that will allow large storage volume as well as low latency read/write access. Hosting in such a manner will require a number of Python libraries like Dora and FTFY ('fixes text for you') intended for data cleaning and preprocessing will be used for these purposes. Yelp has already performed a substantial amount of cleaning on the data, so it is likely that minimal additional cleaning will be necessary, save for replacing missing values as mentioned above.

Numpy and Scipy will be used to transform and mine the data, and other Python modules like Seaborn and Plotly (a Python API for D3) will be used for knowledge visualization and presentation. Finding angry reviews and recurring problems will make heavy use of NLP, so the team will rely on tools like Weka, the Text Analytics API from Microsoft Cognitive Services, and possibly some custom supervised text classifiers trained with the Facebook fastText pretrained word vectors to perform sentiment analysis on the review text.

5 EVALUATION

5.1 Methodology

Evaluating our predictions as to whether a business success will hinge on comparing the star rating (from one to ve stars) and our predicted star rating based on the other attributes in the review, such as day of the week and month, season, or location. This could be done by training an ML algorithm like Kernel Ridge regression

on 80% of the data and testing it on the remaining 20%, and then adjusting parameters to minimize the error between the actual star ratings and our predicted star ratings for the testing data.

Evaluating the sentiment analysis of the review text will likely be simple given that we will be using tools like Weka and Microsoft Sentiment Analysis (see following section) to perform sentiment analysis rather than creating entirely custom algorithms. These tools incorporate helpful metrics to analyze the error of a particular output. Alternatively, we can manually verify the sentiment of a (relatively small) number of reviews by reading them and comparing our human judgements of the sentiment to the output of the tools. For evaluating recurring problems, we will employ a similar approach, except we will group reviews by business and analyze them as a time series to and recurring issues.

For all three of the objective questions, employing association rule frameworks that use support-confidence frameworks will be helpful to and strong associations, while null-invariant measures like the Kulczynski measure or the cosine measure will be helpful to determine how interesting these associations are.

6 PROGRESS

There have been several major changes to the project tools and methods. Initially we planned on using a relational database, but ran into a lot of trouble. The inconsistencies between the user and business data made it too difficult to successfully import it into the database, so instead we decided to import the data straight into Weka. However, the format of the JSON file did not allow a successful import into Weka.

Next we tried to convert the JSON file into CSV format. Once again Weka had issues due to the mismatch in attributes for each business (some had more attributes while others had less).

Instead we opted to use a different tool, RapidMiner. Like Weka, RapidMiner incorporates a extensive machine learning functionality like classification, regression, clustering, and association and frequent itemset mining. RapidMiner is commercial software, but an educational version is available that removes limitations on the free version like only being able to process 10,000 rows at a time. Additionally, RapidMiner is more compatible with the missing data in our dataset and automatically corrects the formatting issues we had when attempting to use Weka. It does not support JSON so we reused the conversion code (found on the team GitHub as convert.py) to convert the data files to CSV.

7 RESULTS

7.1 Businesses

After importing the CSV formatted version of the data we tested some basic visualization of the data. For instance, in Figure 1 we created a histogram of the average star for a business versus the state in which it?s located. Basic graphs like these help refine our question. For example, why does Illinois have significantly lower average stars than others?

The most pervasive issue with our business data is that many entries are missing data, most notably the 'attributes' and address information. Since these values are managed by the account owner, they vary widely across the businesses.

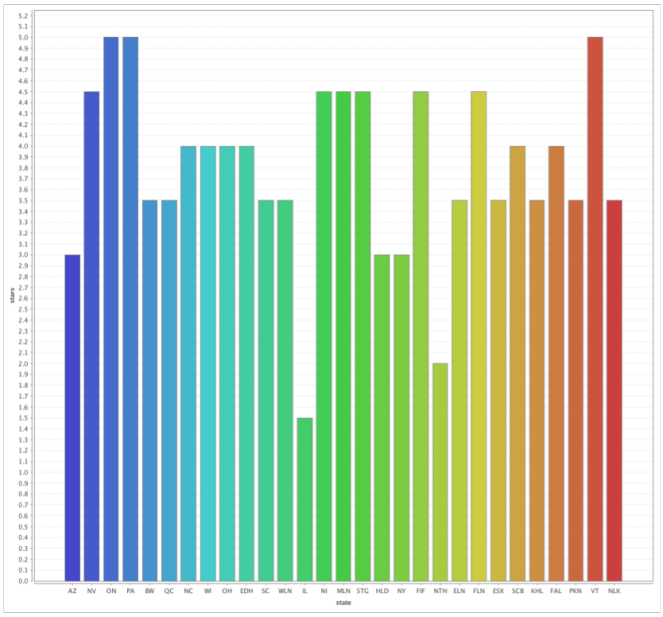
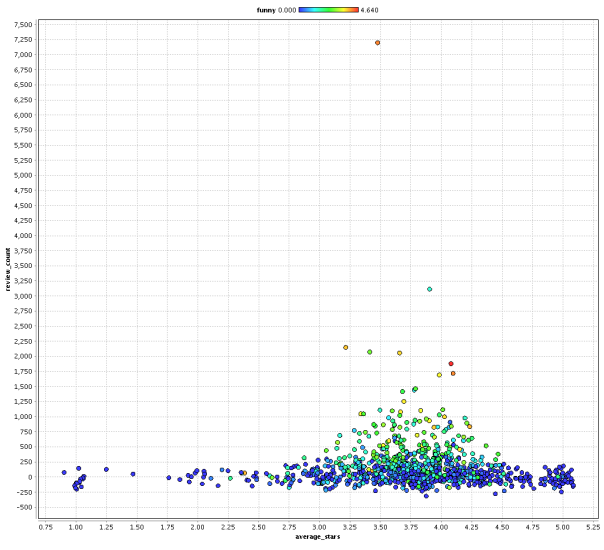


Figure 1

7.2 Users

There are some very interesting features here. It essentially provides a window into users' status, connections, and interaction within the Yelp ecosystem.



8 NEXT STEPS

It is still necessary to reduce dimensionality. We?ve currently been working on the whole dataset and simply ignoring all attributes that fall in the category ?other.? This is cumbersome and it will become increasingly detrimental as we use more advanced techniques like clustering. The most prioritized attribute is the attributes portion. This attribute has the most issues because it is a miscellaneous catch-all where any attribute that didn?t fit with the others is thrown in. The issue is that it totally lacks consistency in the number and types of miscellaneous attributes included for each business. One business may have a Free Wifi attribute while others may not. After looking through the data it was clear that there wasn?t one attribute in this misc. category that was reliably present for all businesses. Although we have moved on to work with the other attributes, this one still needs work to decide if it should be omitted completely or if default values should be set instead.

In addition, we need to combine the businesses and reviews dataset. This becomes extremely useful when examining the sentiment of a review and finding a correlation with the attributes of a business. This helps us build a model to potentially predict if a business receives negative or positive reviews based on its attributes.

As mentioned before, we will be interested in clustering businesses together as to whether they receive positive or negative reviews based on their attributes. This can be difficult with the high number of dimensions of these datasets. We will try CLIQUE and PROCLUS for subspace clustering. PROCLUS will be particularly useful because the subspace for one business to qualify as ?positively reviewed? could be different than the subspace for another business. This allows us to process this data faster, but also consider the complexities of various subspaces.