AUTOMATIC BUSINESS ATTRIBUTE LABELING FROM YELP REVIEWS: A MACHINE LEARNING APPLICATION

- by Michael Iannelli and Ephraim Rosenfeld
- Courant Institute of Mathematical Sciences – New York University
WHAT IS YELP?
• “Yelp combines traditional business listings in a directory like Yellow Pages with social elements. Customers can leave feedback on their experiences with that business which … informs future customers of what they might expect and … keeps standards high, or forces an improvement of those standards to prevent negative feedback.”

What is the Yelp Dataset Challenge?
- Since 2009, Yelp has published samples of its business and reviewer data so that students can compete in research projects based off these datasets
- Data consists of business profiles, reviewer account information, and business reviews for: 8 cities in the U.S., 2 in Canada, 1 in the U.K., and 1 in Germany

What is the purpose of our analytic?
- To predict business profile attributes using review data

Why is this valuable?
- Manual curation of business account information can be time-consuming, error-prone, and inaccurate
- Leveraging crowd-generated data can provide objective insight into a business or venue

1. Source: https://www.techjunkie.com/what-is-yelp/
HOW DO WE PERFORM THE PREDICTIONS?

1. REVIEW DATA AS TERM FEATURE VECTORS
   - Reviews are represented as term feature vectors
   - TF-IDF scores can be used to give a weight to terms in the vector
   - Binary classifiers suitable for sparse, noisy, high dimensional vectors (Naïve-Bayes and Linear SVC)
   - Advantage: “open box” where top-performing features can be analyzed

2. WORD2VEC
   - Reviews are represented as summaries of distributed word or document representations
   - Binary classifiers suitable for dense, lower-dimensional vectors, e.g. SVC, logistical regression, and gradient-boosted machine (GBM)
   - Advantage: greater accuracy by utilizing context of words
LOOKING AT THE DATA

The Yelp Challenge consists of the following JSON files:

1. Businesses: name, id, and profile attributes
   
   ```json
   {"business_id":"eYzm1jUK0GsI2KOLTt2PbQ", "name":"Bâton Rouge Steakhouse & Bar", "neighborhood":"Downtown Core", "address":"218 Yonge Street", "city":"Toronto", "state":"ON", "stars":3.0, "review_count":107, "is_open":1, "attributes":{"Alcohol: full_bar","Ambience: {'romantic':False, 'intimate':False, 'classy':False, 'hipster':False, 'touristy':False, 'trendy':False, 'upscale':False, 'casual':True}, "BikeParking":False, "BusinessAcceptsCreditCards":True, "Caters":False,"GoodForKids":True}}
   ```

2. Business tips: short, informational posts that Yelpers provide to each other and to potential customers:
   “Cash-only! Their Lyonnaise potatoes are very well seasoned.”

3. Business Reviews: lengthier than tips and usually contain a narrative of a customer’s experience:
   “Like any other Zoe's, this location has great sandwiches, salads, kabobs, and more … I like the location -- Birkdale is super convenient -- but parking can sometimes be a challenge because of the popularity of this shopping center. Things are tight inside the restaurant, too, but they're very kid-friendly and don't mind if you have a stroller with you. Friendly staff, great food, great location.”
DATA PREPARATION AND MODELING

1. Review JSON files are aggregated by business to make one large concatenated review

2. Business Profile information is joined to the review data to form a single JSON record

3. An equal number of positive and negative labels are selected to create balanced datasets

4. Review data is turned into a word feature vectors and the business attribute acts as the class label

5. Feature Selection
   5a. Top-performing features are extracted for further analysis

6. Labelled datasets are fed to a classifier

7. Cross validation is used to assess the results
RESULTS

• Accuracy improved when using larger datasets

• Accuracy differed based upon attribute, with classification of more obscure attributes performing more poorly

• TF-IDF term-vector modeling performed as well, or better than, classification using Word2Vec feature vectors (SVM performed the best)

• Chi-square feature selection did not substantially improve performance
By imposing a minimum-document frequency threshold on uncommon terms, the feature size “levels-off” as the sample-size increases.

Increasing the review-count-per-business threshold increased accuracy.

Performance with different business-review thresholds.
## ANALYSIS OF TOP-PERFORMING FEATURES

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Positively-correlated SVC Features</th>
<th>Negatively-correlated SVC Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepts Credit Card</td>
<td>card, pricey, online, ordered, hotel</td>
<td>cash, atm, debit, plastic, cards, cart</td>
</tr>
<tr>
<td>Dogs Allowed</td>
<td>patio, outside, pet, dog</td>
<td>marriot, lobster, lounge, salon</td>
</tr>
<tr>
<td>Restaurant Delivers</td>
<td>delivery, delivers, phone, deliveries, ordered, (“pizza” w/ chi2 and tf-idf)</td>
<td>smoking, casino, register, cost, seated</td>
</tr>
<tr>
<td>Good for Kids</td>
<td>kids, family, families, friendly, daughter, slushies</td>
<td>bar, reservation, crowd, hip, downtown, soju, drunk, trendy, casino, cocktail, dj</td>
</tr>
<tr>
<td>Wheelchair Accessible</td>
<td>mall, elevator, hotel, plaza</td>
<td>stairs, upstairs</td>
</tr>
</tbody>
</table>
## What Top-Performing Features Tell Us About Location and Culture

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Positively-correlated SVC Features</th>
<th>Negatively-correlated SVC Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepts Credit Card</td>
<td></td>
<td>Pittsburgh, PA</td>
</tr>
<tr>
<td>Dogs Allowed</td>
<td>Scottsdale, AZ and Stuttgart, Germany</td>
<td>Las Vegas, NV</td>
</tr>
<tr>
<td>Wheelchair Accessible</td>
<td>Scottsdale, AZ</td>
<td>Toronto, ON and Montreal, QC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Top TF-IDF Terms</th>
<th>Bottom TF-IDF Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurant Delivers</td>
<td>pizza, chinese, rice, sushi, chicken, lunch</td>
<td>hefeweizen (a kind of beer), abendessen (dinner), essens (food)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Top Chi-Squared Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dogs Allowed</td>
<td>sushi, rice, thai, korean, noodles, Japanese, pho, ramen</td>
</tr>
</tbody>
</table>
FUTURE WORK

• Improve performance
  • Incorporate n-grams into our modeling
  • Calibrate the inverse-document frequency (IDF) to give less weight to ubiquitous terms
  • Use a higher confidence threshold or review-count-per-business threshold to classify fewer businesses with greater

• Top-performing feature analysis
  • Assess the polarity of chi-squared correlated features

• Use vector arithmetic for business attribute research:
  • Example: I like all qualities of Business-X except for one, e.g. “Smoking Prohibited”:
  • Vector arithmetic: [Business-X] + [Smoking Permitted] = [Business-Y where smoking is permitted]