The Language of Food: Analysis of Yelp Reviews

Jessica Hays, Amy Lai, Ben Thompson
W266 Spring 2018

Abstract

Crowd-sourced platforms like Yelp use Likert-based rating scales to assess businesses and services. Although such a medium provides important information to both businesses and consumers, little is understood about differences between the individual rating categories. To address this question, our project utilized text reviews and employed several models to explore the "linguistic markers" of different rating categories. Compared to topics that were extracted from the reviews by Latent Dirichlet Allocation (LDA), we identified keywords that were most strongly associated with the rating categories based on linear models, and explored the semantic meaningfulness and syntactic roles of these words through word embeddings and part-of-speech tagging. We found that logistic regression based on tf-idf scores best captured the tone of reviews, whereas tf-idf scores alone highlighted the content of reviews. Our results emphasize that different modeling approaches highlighted different linguistic characteristics of text.

Introduction

Today, businesses and consumers are widely present on crowd-sourced platforms such as Yelp. Businesses increasingly depend on ratings and reviews to expand business and increase revenue (1), while consumers readily rely on such information to find businesses that meet their needs and share their experiences with other consumers. As a result, these platforms often provide rich information that can greatly inform business or personal decisions.

However, in evaluating businesses and services, crowd-sourced platforms often use Likert-based rating scales that generally tend to be ill-defined, with little to no indication about the meaningfulness or significance of individual rating categories. For example, a number of us have been in situations in which we have used Yelp to find a "good" restaurant. But as we sift through over 20 recommended places (each with various ratings and more than 100 reviews), how do we choose an option? What do the ratings indicate? How does a 2-star rating differ from that of a 4-star? Do these standards apply across different geographic locations, types of businesses, or time periods?

By using different models to analyze text reviews, our project sought to address these questions. We believe such findings can help shed light on the "linguistic markers" of different rating categories, and evaluate the increasingly widely yet largely implicit use of Likert-based rating scales.

Background

Prior research has developed popular topic modeling approaches such as Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) that extract topics from text documents (2,3). With the increasing popularity and influence of crowd-sourced platforms, other studies have further introduced variations of these models to extract features of products from consumer reviews (4,5). Although such models provide useful information about key concepts expressed

through text, they are unsupervised methods that inherently restrict the ability to evaluate the features that are generated.

Moreover, prior research suggests that how language is used may differ due to cultural and social factors. One study found that differences in the language used to advertise potato chips reflected socioeconomic status (6). It found that longer and more complex text was used by higher-end brands to appeal to more educated consumers, whereas shorter and less complex text was used by lower-end brands to appeal to less educated consumers (6). However, little is known about how language used in consumer reviews may vary by rating categories.

Thus, to address these issues, our project used supervised methods to identify keywords that were most strongly associated with rating categories, and further identify keywords by factors such as geographic location and time period. We hypothesized that this approach would highlight different linguistic features of consumer reviews than popular topic modeling techniques.

Methods

We used data provided by the Yelp Dataset Challenge Round 11 (7). The full dataset contains information about users, businesses, business check-ins, and reviews. In order to obtain more interpretable and meaningful results, we focused only on businesses whose category was "restaurant." The final sample included over 49,000 restaurants from various cities including Phoenix, Arizona (USA), Toronto, Ontario (Canada), and Las Vegas, Nevada (USA), and over 800,000 reviews spanning 5 years (2012-2017).

Exploratory data analyses

We performed exploratory data analyses to better understand the text reviews and star ratings. We found that the reviews were written in English, generally adhered to standard grammatical rules, and frequently varied in length ranging from several sentences to several paragraphs. Because many models often regard each unique word as a distinct feature, we preprocessed the reviews by lowercasing characters, removing punctuations, and replacing digits with a single token to reduce the total vocabulary size and increase overall textual consistency.

Furthermore, we found that the ratings were integers on a scale of 1 to 5. The distribution of ratings was skewed toward 4- and 5-star ratings. Because reviews for 1- and 2-star ratings and those for 4- and 5-star ratings were similar, we combined these ratings to increase the number of examples for lower star ratings and increase interpretability of results. The final rating scheme used in this project consisted of rating category: 0 (includes 1-2 star ratings), 1 (includes only 3 star ratings), and 2 (includes 4-5 star ratings).

Identification of keywords

To address the unsupervised nature of many topic modeling techniques, we trained several linear models to identify keywords that were most strongly associated with the rating categories.

First, we trained a Multinomial Naive Bayes model based on the occurrence of words to assess baseline performance. We believed that this was an appropriate baseline model because Naive Bayes is a simple bag-of-words model whose mathematical theory is most similar to that of LDA

(3). Because we expected common words such as "the" to appear frequently, we removed stop words (a combination of default stop words from NLTK and Scikit-learn and other manually identified words) and restricted the vocabulary size to 10,000 in order to obtain more meaningful results. The keywords that were most strongly associated with the rating categories was defined as features that had the largest coefficients in each category.

Second, because common words may still be highly represented in our baseline model even after removal of stop words, we trained a Logistic Regression model based on tf-idf scores to identify more meaningful keywords that have greater discriminatory power among the rating categories. To mitigate potential effects of class imbalance, we oversampled minority labels with replacement. The keywords that were most strongly associated with the rating categories were also defined as features that had the largest coefficients in each category. A similar procedure was used to determine keywords by geographic location (Phoenix, Arizona; Toronto, Ontario; Las Vegas, Nevada), cuisine (coffee, fast food, Mexican), and time period (2012, 2017).

Furthermore, given that tf-idf may be an effective way to create keyword tags for documents (8,9), we used tf-idf scores to tag each review with the keywords that had the greatest score. This approach allowed us to search and organize reviews based on their keyword tags, and identify a different set of keywords associated with the rating categories by counting the number of times each tag was used within a category. We then compared these keywords with those from our previous model to gain further insight about the reviews.

Analysis of keywords

Finally, to understand the semantic characteristics and meaningfulness of the keywords, we used pre-trained GloVe word embeddings and computed cosine similarity to determine words that were similar, dissimilar, and most unique (only unigrams were used due to lack of embeddings for phrases). We defined the most unique word to be that in which either the sum or the max of the cosine similarity (when compared to all other words in the set) was the smallest. We then expanded this exploration to find groups of words that formed meaningful clusters. Because the words were largely similar and did not appear to have an innate hierarchy, we used k-means rather than hierarchical clustering.

We also explored the role that different parts of speech may have played in each rating category. Since individual words may have multiple parts of speech depending on the context, it was difficult to accurately tag words *after* they had emerged from the model without their original context. To address this, we tagged all words in every review and counted the frequency of each part of speech tag for each word. After features were extracted from the model, we then assigned each word with the tag that had the highest occurrence. This allowed us to identify keywords by part of speech tag to draw further insights.

Results

Using a Multinomial Naive Bayes model based on the occurrence of words (accuracy: 79%, F1: 80%), the keywords that were most strongly associated with each rating category are shown in Table 1. Although common words like "really" were still present, our results show high overlap among the rating categories. We found that nouns such as "food", "place", and "service" were

strongly related to all categories. Interestingly, we found that the keywords identified by Naive Bayes were similar to topics that were generated by LDA (results not shown).

For the Logistic Regression model based on TF-IDF scores (accuracy: 80%, F1: 81%), the keywords that were most strongly associated with each rating category are presented in Table 2a. In contrast to our baseline model, the results highlight that a distinctive set of adjectives was strongly associated with each rating category. We found that negative adjectives such as "disgusting", "tasteless", and "awful" were strongly associated with 1-2 star ratings, whereas positive adjectives such as "fantastic", "delicious", and "perfect" were strongly related to 4-5 star ratings. We further examined keywords by geographic location, cuisine, and year (Tables 2b-2d, respectively). To our surprise, we found that the keywords were largely similar across cities, cuisine, and year.

Although improving model performance was not the focus of our project, our error analysis showed that the model had the greatest difficulty classifying the 3 star rating category. We found that words that expressed both negative and positive sentiments such as "little disappointed", "hit miss", "bad great", and "mixed feelings" were strongly correlated with this category. When this category was excluded, we found that the model correctly classified the reviews with an accuracy of roughly 95%. Further error analysis showed that model also had difficulty classifying reviews that were overwhelmingly negative or positive, but received a rating that contradicted that sentiment. An example is shown below.

Actual Class: 2
Predicted Class: 0
Review: *"FAIL! Was there this afternoon at 1pm, my pasta did not come out until 1:50pm. There were only 3 tables in the entire resturant. Took zero ownership, at 1:40pm I asked about the pasta, waiter said it'll be out in 5 min, nope. Comped only my meal, not my friends (you messed up on one order and ruined two experiences). 4 star for quality of food."*

Interestingly, when we used tf-idf scores to create keyword tags for each review, the keywords associated with the rating categories were primarily types of food (Table 3). In contrast to our previous model, these results highlighted the contents of a review (e.g., pizza, sushi, coffee, etc.). They also showed significant overlap between rating categories, as many reviews had similar keyword tags (e.g., "pad", "thai", and "pad thai").

Word embeddings, part-of-speech tagging
Our analysis of word embeddings revealed that many of the top words were similar. Overall, there were generally 2-3 larger clusters of words, with fewer words in the remaining clusters. An example of clusters of keywords for 1-2 star ratings are shown in Table 4. We also found interesting results by identifying the most unique words, with "cockroach" as the most unique word for 1-2 star ratings.

Finally, results for part-of-speech tagging (filtering for only nouns and verbs) are shown in Table 5. Overall, we found that 4-5 star ratings were more generically enthusiastic (e.g., "heaven," "gem", "love", "recommend") whereas 1-2 star ratings were more specific about factors that may

have contributed to their dissatisfaction (e.g., “cafeteria,” “microwave;” “ignored,” “refused”). Interestingly, the content of *poorer* reviews seemed to drive these insights.

Conclusions

This project explored the “linguistic markers” of rating categories used by Yelp. By using different models to analyze text reviews, we found that Naive Bayes-related models captured the content of reviews, whereas logistic regression models based on tf-idf scores reflected the tone of those reviews. We further found that tf-idf scores also described the content of reviews, and that part-of-speech tagging highlighted keywords that may have driven the description of the contents of reviews. These results underscore that different models highlight different types of information. In doing so, our findings illustrate the unique characteristics of reviews and suggest that the platform and language used to express satisfaction (or dissatisfaction) with food may be more innate, intuitive, and similar than we believed. Future work to apply these models on other datasets is needed to determine whether the types of information retrieved are consistent.

Code

Our code repository can be found here.

References

1. Anderson M, Magruder J. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. The Economic Journal. 2012:957-989.

2. Papadimitriou C, Raghavan P, et al. Latent Semantic Indexing: A Probabilistic Analysis. Journal of Computer and System Sciences. 2000:217-235.

3. Blei DM, Ng AY, et al. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003:993-1022.

4. Moghaddam S, Ester M. ILDA: Interdependent LDA Model for Learning Latent Aspects and Their Ratings from Online Product Reviews. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011:665-674.

5. Titov I, McDonald R. Modeling Online Reviews with Multi-Grain Topic Models. Proceedings of the 17th International Conference on World Wide Web. 2008:111-120.

6. Freedman J, Jurafsky D. Authenticity in America: Class Distinctions in Potato Chips Advertising. Gastronomica. 2011:46-54.

7. Yelp. Yelp Dataset Challenge. https://www.yelp.com/dataset/challenge. Accessed on: February 15, 2018.

8. Liu F, Pennell D, et al. Unsupervised Approaches for Automatic Keyword Extraction using Meeting Transcripts. Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics. 2009:620-628.

9. Danesh S, Sumner T, et al. SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction. Lexical and Computational Semantics. 2015:117-126.

Table 1. Keywords by rating category from Multinomial Naive Bayes with Countvectorizer

1-2 star ratings	3 star rating	4-5 star ratings
food	good	food
place	food	good
service	place	place
good	like	great
like	service	service
time	really	like
order	time	time
ordered	great	really
restaurant	ordered	delicious
got	chicken	best

Table 2a. Keywords by rating category from Logistic Regression with Tf-Idf

1-2 star ratings	3 star rating	4-5 star ratings
worst	aok	delicious
zero stars	hit miss	amazing
disgusting	dont wrong	wont disappoint
terrible	little disappointed	excellent
horrible	decent	best
cockroach	wanted love	exceeded
negative stars	bad great	perfect
awful	wouldnt way	great
tasteless	stars	awesome
disappointing	mixed feelings	fantastic

Table 2b. Keywords by geographic location from Logistic Regression with Tf-Idf

Star Ratings	Phoenix, Arizona	Toronto, Ontario	Las Vegas, Nevada
1-2	money	horrible	told
1-2	wo	dry	sad

1-2	cold	disappointed	disgusting
1-2	mediocre	awful	disappointed
1-2	rude	rude	rude
1-2	poor	disappointing	terrible
1-2	awful	asked	poor
1-2	horrible	terrible	bland
1-2	worst	bland	horrible
1-2	terrible	worst	worst
3	probably	pretty good	little
3	average	good	good
3	bit	average	bit
3	little	pretty	meh
3	think	ok	stars
3	okay	decent	average
3	pretty	okay	overall
3	decent	overall	ok
3	good	bit	okay
3	ok	stars	decent
4-5	friendly	friendly	fantastic
4-5	definitely	perfect	favorite
4-5	favorite	definitely	perfect
4-5	awesome	fresh	love
4-5	excellent	excellent	excellent
4-5	best	love	awesome
4-5	love	best	best
4-5	amazing	great	great
4-5	great	amazing	amazing
4-5	delicious	delicious	delicious

Table 2c. Keywords by cuisine from Logistic Regression with Tf-Idf

Star Rating	**Coffee**	**Fast food**	**Mexican**
1-2	asked	asked	awful
1-2	disappointing	cold	bland
1-2	minutes	going	cold
1-2	money	horrible	disappointed
1-2	order	minutes	dry
1-2	poor	ordered	horrible
1-2	rude	rude	rude

1-2	terrible	slow	terrible
1-2	told	terrible	wo
1-2	worst	worst	worst
3	bit	special	standard
3	good	bagel	bit
3	overall	average	nice
3	okay	food good	little
3	pretty good	pretty	good
3	little	good	pretty
3	average	decent	overall
3	ok	standard	okay
3	price	okay	decent
3	pretty	ok	ok
4-5	awesome	fresh	friendly
4-5	perfect	favorite	favorite
4-5	friendly	definitely	definitely
4-5	favorite	friendly	excellent
4-5	definitely	awesome	awesome
4-5	best	amazing	love
4-5	love	best	best
4-5	amazing	love	great
4-5	delicious	delicious	amazing
4-5	great	great	delicious

Table 2d. Keywords by year from Logistic Regression with Tf-Idf

Star Rating	**2012**	**2017**
1-2	bad	awful
1-2	better	bland
1-2	bland	disappointed
1-2	cold	horrible
1-2	dry	money
1-2	horrible	poor
1-2	mediocre	rude
1-2	overpriced	terrible
1-2	terrible	wo
1-2	worst	worst
3	bit	alright
3	decent	average

3	good	bit
3	little	decent
3	ok	fine
3	okay	good
3	pretty	ok
3	pretty good	okay
3	stars	stars
3	typical	tad
4-5	amazing	amazing
4-5	awesome	awesome
4-5	best	best
4-5	delicious	definitely
4-5	excellent	delicious
4-5	fantastic	excellent
4-5	favorite	favorite
4-5	great	great
4-5	love	love
4-5	loved	perfect

Table 3. Keywords by rating category from Tf-Idf

1-2 star ratings	**3 star rating**	**4-5 star ratings**
pizza	pizza	pizza
burger	burger	sushi
sushi	sushi	burger
order	chicken	breakfast
chicken	sandwich	chicken
minutes	buffet	coffee
manager	breakfast	thai
waiter	steak	love
fish	wings	great
table	coffee	sandwich

Table 4. K-means clustering of keywords for 1-2 star ratings

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
unacceptable awful terrible disgusting horrible	bland tasteless flavorless inedible cockroach	never cold zero	disappointing disappointment worst	poor sick	waste	poisoning

Table 5. Key nouns and verbs by rating category from Logistic Regression with Tf-Idf

Key Nouns by Rating Category		
1-2 star ratings	**3 star rating**	**4-5 star ratings**
cockroach	feelings	gem
joke	overall	heaven
cafeteria	copy	bomb
garbage	damper	notch
microwave	inconsistency	blast

Key Verbs by Rating Category		
1-2 star ratings	**3 star rating**	**4-5 star ratings**
left	torn	love
processed	improves	beat
wasting	wowed	recommend
ignored	detracted	delivers
refused	depends	exceeded