CS 412: Intro to Machine Learning
Final Report

How Yelpy are Yelpers?

By

The CS412 Yelp Team
Hongwei Jin
Krutarth Joshi
Aayush Kataria
Natawut Monaikul
Ashwin Sattiraju
Zhan Shi
Dan Zhao

Contents

1 Introduction

As of 2015, over 89 million people visit Yelp online per month [10], whether it be to leave a review for a business or look up local restaurants to try. With so many users, there is naturally a massive amount of data being generated each day, so when Yelp Inc. released a dataset containing just about everything collected from their website over the past several years, many machine learning tasks immediately presented themselves. So then, given this vast dataset, what kinds of predictions/inferences can we make?

1.1 Background

Yelp is a website on which users can leave and read reviews about businesses, as well as find general information about those businesses, including hours of operation, menu, location, etc. Businesses on Yelp range from restaurants to salons to department stores, even to universities. Yelp also serves as a social media site, as reviewers can create profiles that display their location, favorite places, past reviews, and so on. Yelp's recently-released dataset includes all of this information and more. The main focus of this project is based on the fact that users can leave a star rating (a whole number between 1 and 5) for a business.

1.2 Problem

As with many domains in which one can leave reviews and ratings for something, a common task is to predict how positively or negatively that something will be rated by its audience (as in [7] for movies, [8] for YouTube videos, and [6] for books). Thus, Yelp lends itself to the machine learning task that we would like to address: can we predict the star rating a reviewer will give a business? Making accurate predictions about how a user will rate a business has applications in detecting spam (fake users leaving reviews that are inconsistent with their star ratings), more accurately aggregating user reviews to determine an overall star rating for a business, and recommending other businesses to Yelp users.

2 Methods

Given the large dataset released by Yelp (retrieved from [11]), we must first determine what subset of the data we will focus on to make our predictions. We explored two directions based on the knowledge source we used to approach the machine learning task: a review-based branch and a profile-based branch.

2.1 Review-Based Branch

When a user rates a business, they can also leave a textual review of the business discussing what they did and did not enjoy about the business to justify their star rating. In the Yelp dataset, each review is listed with the user and the business, the text in the review, the star rating given by the user, and the date of the review. In addition to a full review, a user can also leave a "tip" about the business – a short, simple piece of information a user wants to impart on other users to summarize their thoughts or feelings about a business. These tips can be considered as part of the textual review. In the review-based branch, we claim that all of this information is sufficient to predict the star rating a user will give a business.

2.1.1 Features

We extract several features from users' reviews:

- **Sentiment score.** Given a textual review, we would like to score the positivity or negativity of the text. We calculate this score by summing the positivity of negativity ratings of each individual word. There are many ways to determine these word ratings. One method, as described and employed in [4], is to use a semantic network such as WordNet [5], which stores adjectives as bipolar clusters – a word such as *fast* would have an entry with its synonyms (such as *quick*, *swift*, *prompt*, etc.) and its opposing word (*slow*), along with its synonyms. A set of positive words can then be obtained by starting with a "seed" set of positive words and searching all of the neighbors, where all synonyms would also be considered positive words, while all opposing clusters would be considered negative words. This has also led to the development of SentiWordNet [3], a resource for obtaining positivity and negativity ratings of words.

  However, in our system, we would like to minimize the use of external resources. Therefore, to obtain ratings for words strictly using only our training set, we first calculate relative frequencies of each word in five-star reviews and one-star reviews in the training set. The idea behind this is that positive words will appear frequently in the most positive reviews, while negative words will appear frequently in the most negative reviews. Clearly, many words inherently occur frequently, regardless of its positivity or negativity, so we also remove the 500 most frequent words across all reviews in the training set to filter out common words. The relative frequencies of the remaining words then provide "scores" for each word, where positive words receive a score equal to their relative frequencies, and negative words receive a score equal to the negative value of their relative frequencies. A list of some of the words extracted using this method are given in Table 1. Using these word lists, the sentiment score of a review is simply the sum of the weights over all words in the text that are found in the lists. Not only did this method provide a way to score words, it also exposed some common aspects of businesses that people tend to focus on, for positive or for negative reasons, using words that would otherwise have no connotation of polarity in a general context. For example, *dry* was found to be a commonly used word in negative reviews; however, in its most general sense, *dry* arguably does not necessarily have a positive or negative connotation, but in the context of, say, food, it is generally a negative word.

Positive Words		**Negative Words**	
Word	**Score**	**Word**	**Score**
thank	0.00145	worst	-0.00378
perfectly	0.00134	horrible	-0.00343
easy	0.00132	rude	-0.00313
yummy	0.00129	terrible	-0.00278
reasonable	0.00129	waited	-0.00254
professional	0.00127	poor	-0.00188
attentive	0.00116	problem	-0.00181
glad	0.00112	bill	-0.00176
wow	0.00112	leave	-0.00166
recommended	0.00105	dry	-0.00143

Table 1: Sample positive and negative words and scores

- **Punctuation count.** When people become emotionally excited, whether it be in a positive or a negative manner, their "online language" tends to reflect their emotions [9]. One of the features of this language is the amount of punctuation used. For example, "Very good!!!!" and "Very bad!!!!" contain several exclamation points and reflect a strong, emphasized, emotional statement. Furthermore, punctuation is commonly used to draw emoticons, which can again reflect the polarity of the text.

- **Cap word count.** Another feature presented in [9] in conjunction with excessive punctuation is the use of upper-case letters. For example, "I LOVE it" and "I HATE it" use a completely capitalized word to emphasize the word, indicating an expression of strong emotion.

- **Emphasized word count.** The final character-based feature we consider (again, as proposed in [9]) is the number of words which contain the intentional repetition of letters in a word, e.g., *mmmmmm* and *ewwwwwww*.

- **Word count.** A general feature extracted from a piece of text is the number of words in the text. We do not posit a specific relationship between the number of words and the polarity of a review, but we include the feature in hopes that there might be.

- **Adjective count.** The number of adjectives used in a piece of text can also suggest the polarity/neutrality of it [1]. Therefore, we extract as a feature the number of adjectives used in a review with the help of a part-of-speech tagger (provided by Python's `nltk` package).
- **Adverb count.** It is also shown in [1] that the number of adverbs used in a piece of text reflects the writer's sentiment, so using the same method, we also consider this feature.
- **Weekday/Weekend.** It has been found in a study analyzing product reviews in 2014 that weekday purchases receive more positive feedback than weekend purchases [2]. In light of this, we also use the date of a review to determine whether the review was left on a weekday (Monday-Thursday) or a weekend (Friday-Sunday) and use this as our final feature.

2.1.2 Classifiers

We use a variety of classifiers with these features to predict the star rating of a review (with the help of Python's `scikit` package): Decision Tree, Random Forest (with 10 decision trees), AdaBoost (with decision trees as weak learners), 5-Nearest Neighbors, 10-Nearest Neighbors, and 20-Nearest Neighbors. Each classifier predicts one of five labels (the five whole-valued star ratings) for each review in the test set.

2.2 Profile-Based Branch

2.2.1 Features

2.2.2 Classifiers

2.2.3 Recommender System

3 Evaluation

We evaluated the classifiers from each branch separately (and the recommender system in the profile-based branch).

3.1 Review-Based Branch

Due to the large size of the entire dataset (over 1.5 million reviews), we systematically divided the dataset in our evaluation. Moreover, because a part-of-speech tagger significantly increases the running time of extracting all of our features, we measured our classifiers' performances when they contained and did not contain the adjective and adverb count features separately. We first randomly chose 100,000 reviews from the full dataset, from which we randomly selected 70,000 to be the training set and 30,000 to be the test set. The features not including the adjective and adverb counts were then extracted from the training set, and then the classifiers were trained on these features and evaluated on the test set. The process of randomly choosing and splitting 100,000 reviews and training and evaluating the classifiers was repeated 9 more times, and the results were averaged over the 10 trials.

We then randomly chose 10,000 reviews from the full dataset, from which we randomly selected 7,000 to be the training set and 3,000 to be the test set. The full set of features was then extracted from this smaller training set, and the classifiers were again trained and evaluated. As in the larger dataset, we performed a total of 10 trials.

3.2 Profile-Based Branch

3.2.1 Classifier Evaluation

3.2.2 Recommender System Evaluation

4 Results

These are the results.

4.1 Review-Based Branch

4.2 Profile-Based Branch

5 Conclusion

This is the conclusion.

5.1 Summary

5.2 Lessons Learned

5.3 Future Work

References

[1] Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., and Subrahmanian, V. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *International Conference on Weblogs and Social Media - ICWSM* (2007).

[2] Digital Strategy Consulting Ltd. Online review habits: Weekday purchases get better feedback. `http://www.netimperative.com/2015/03/online-review-habits-weekday-purchases-get-better-feedback/`, 2015.

[3] Esuli, A., and Sebastiani, F. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)* (Genova, Italy, 2006), pp. 417–424.

[4] Hu, M., and Liu, B. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)* (Seattle, Washington, USA, 2004).

[5] Miller, G. A. WordNet: A lexical database for English. *Communications of the ACM 38*, 11 (1995), 39–41.

[6] Mooney, R. J., and Roy, L. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries* (2000), pp. 195–204.

[7] Oghina, A., Breuss, M., Tsagkias, M., and de Rijke, M. Predicting imdb movie ratings using social media. In *Advances in Information Retrieval*, vol. 7224. Springer Berlin Heidelberg, 2012, pp. 503–507.

[8] Siersdorfer, S., Chelaru, S., Nejdl, W., and Pedro, J. S. How useful are your comments?: analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th international conference on World wide web* (2010), pp. 891–900.

[9] Yassine, M., and Hajj, H. A framework for emotion mining from text in online social networks. In *2010 IEEE International Conference on Data Mining Workshops* (Genova, Italy, 2010), pp. 1136–1142.

[10] Yelp Inc. An Introduction to Yelp Metrics. `http://www.yelp.com/factsheet`, September 2015.

[11] Yelp Inc. Yelp Dataset Challenge. `http://www.yelp.com/dataset_challenge`, 2015.