Using Star Ratings as Trend-Aware Performance Metrics

Caroline Chang, Kenta Takatsu

October 22, 2017

1 Research Context

Past research on the Yelp dataset has focused on extracting latent subtopics from review text using Latent Dirochlet Allocation (LDA) (Huang et al. 2014). Modified LDA algorithms have also been devised to better integrate semantic analysis into generating such subtopics (Linshi et al. 2014). However, what these past studies have not yet explored is a systematic way of assessing business performance – as indicated through star ratings —- as well as the prediction accuracy of past predicted star ratings for different businesses. It would be useful to have a fine-tuned classifier that gives an estimate of the average star rating for particular businesses over a given time interval.

Across the entire Yelp dataset, there are thousands of categories into which a review could potentially fall and not all reviews in any given category are necessarily uniform in "data-richness." In order to properly develop a classifier that analyzes star rating trends over time, we need to be aware of two inherent limitations of this dataset: 1. lack of abundance of reviews for a business and 2. lack of time spread among existing reviews. To train a prediction classifier as accurately as possible, we need to extract subtopics from a large enough category with a variety of reviews per time interval bin. Research in machine learning has shown that using different classifiers in recommendation systems can lead to varying results, which prompts us to choose a classifier carefully.

Generally, linear classifiers are not the most flexible for various datasets, but training and testing a linear regression model on the Yelp dataset would help us establish a baseline for how much more accurate our classifier could be when predicting star ratings. Support vector machines (SVM) and decision tree classifiers are presumably more effective when fitting various data. (Joachims et al. 2017, Jadhav et al. 2016) Developing this "trend-awareness" system with a classifier of decent accuracy will only help the interactions between users and businesses; it will allow users to make more informed choices when selecting restaurants, convenience stores, hotels, etc. and give various industries incentive to continually improve their business models.

2 Research Objective

Our main objective is to develop a stream-lined star rating prediction system that can ultimately give users an idea whether particular businesses are experiencing positive and negative trends. A fine-tuned classifier for a given category would allow us to predict star ratings and provide performance trends over time for businesses. Through this process, we would also be able to determine an appropriate time interval that acts as a good gauge for most accurately predicting the next star rating. Subsequently, determining trending topics for a particular category (while accounting for exponential increase in Yelp users) could also possibly shed some insight -— and perhaps even substantiate drastic fluctuations in the generated business performance trends.

3 Research Road Map

3.1 Experiment Design

Our experimental design is broken up into several stages: 1) pre-processing, 2) testing-and-training classifiers, and 3) trend analysis. In the pre-processing stage, we will use LDA to extract latent

topics from the review text pertaining to businesses from a given "data-rich" category (i.e. "Pizza" or "Chinese restaurants"). Then, for each review in that category, we will generate a k-dimensional embedding that essentially returns a vector containing the frequency distributions of the determined k topics from LDA. For instance, a business that has n number of reviews will have n number of k-dimensional vectors. This k-dimensional embedding matrix will subsequently be used to train and test classifiers with topics as features and star ratings as the response variable. In the second stage, we will start with testing and training a linear regression model on the chosen category's data. We will train a linear regression model by first selecting the largest business with the most reviews (which are spread over a decent amount of time) in a given category and initially train the model on roughly 75% of the reviews. Then, we test the model on the other 25% of the star rating data. After training the model on a 75%-25% basis, in this stage, we will try to detect an optimal "n review" bin size (i.e. how many reviews does the model have to be trained on to most accurately predict the next star ratings). We will compare how the predicted star rating is from the actual star ratings for the remaining 25% of reviews for that business. For all the other restaurants in this category, we will test our model to see how accurate it is by star-rating comparisons. We will then explore other models -— namely, linear SVM and decision tree classifiers -— in order to find a more robust model than linear regression. Once we have found the most optimal classifier, we will apply the model (with minor tweaks) to a different category that has enough of an overlap features (or topics) for validation and assess its potential generalization. Based on this model, we will develop a predictive trend analysis to correlate the star rating to time progression. Using the curve fit that was generated by the model, we can then have a good understanding of the business performance and user approval (through star ratings). After determining business performance on categories, we will then determine topics that gain popularity over time -— features that could influence drastic business performance changes -— while accounting for the exponential increase of number of Yelp users.

3.2 Research Timeline

- October 22nd:
  Developing the LDA algorithm, generating k-index embeddings for reviews of a particular business, and beginning testing-training classifiers (linear regression)
- November 5th:
  Finishing linear regression model, exploring potentially more effective classifiers (i.e. SVM, decision trees), and determining optimal classifier
- November 19th:
  trend analysis for businesses in a particular category (could be multiple) given the most optimal classifier, upward-trend, increasing popularity topics
- December 1st : Deadline
  We will focus on writing in the last 2 weeks.

3.3 Resources

- Pandas, gensim, nltk, numpy, scikitlearn libraries
- Business and review json files from the Yelp dataset
- SVM software for experimentation purposes (possibly)
- CDS server for possible category-merging to increase the data-richness of certain categories (might not be done for this project)

References

[1] Huang, J., Rogers, S., Joo, E. *Improving Restaurants by Extracting Subtopics from Yelp Reviews*. et al. 2014.

[2] Jadhav, S.D. *Efficient Recommendation System Using Decision Tree Classifier and Collaborative Filtering*. et al. 2016.

[3] Joachims, T. *Ranking SVM for Learning from Partial-Information Feedback.* et al. 2016

[4] Linshi, J. *Personalizing Yelp Star Ratings: a Semantic Modeling Approach.* et al. 2014