Effect of Latent Topics in Online User Reviews and Tips on Restaurant Star Rating

Aditya Kammardi Sathyanarayan, Sahil Sanjay Gandhi

Introduction

Over the years, the popularity of online reviews has grown significantly. Reviews help people decide which products or services they should buy, where they should travel and even where they should eat. Over time reviews have slowly become an integral part of business on the internet. Numerous studies have been carried out to help determine the exact impact that both positive and negative reviews have on the popularity, number of sales, product awareness, conversion rate and profitability of businesses selling products or services. An extra half-star rating causes a restaurant to sell out 19 percentage points [1]. However, with owing to the sheer size of data generated online, no business can manually process the information to assimilate the customer's wants. We propose to use Natural Language Programming techniques to study such review texts available online, restaurant reviews in particular, and provide meaningful insights to business owners. We intend to study the impact of the user reviews and user tips for each restaurant business by extracting the hidden topics over all textual data. A natural progression of such a study is to predict the star rating based on the topics discovered. The main idea is to truly understand users' textual reviews and tips and how it impacts their star rating in Yelp.

Related Work

Text streams are ubiquitous in knowledge discovery and data processing workflows. Topic Modelling is a popular machine learning technique for identifying the structure in data. The most well-known topic model, Latent Dirichlet Allocation (LDA), is common for unsupervised text corpus modeling [2]. And application of online LDA to extract hidden topics on the Yelp datasets resulted in some useful insights such as what the users care most in the reviews and pinpoint the areas of interest on specific restaurants [3].

Dataset

The research will be performed on the Yelp Dataset [4]. Yelp releases their data as part of a challenge every year where students from various institutions participate and conduct research on the data. This dataset includes business, review, tips, user, and check-in data in the form of separate JSON objects. A business object includes information about the type of business, location, rating, categories, and business name, as well as contains a unique id. A review object has a rating, review text, and is associated with a specific business id and user id. The tip object is similar to that of the review object, with the difference being that tips are shorter reviews and the text resembles a bulleted action point in contrast to a verbose review. We intend to use the data from user, business, reviews, tips and check-in objects. The total size of the dataset is 4.8 GB with the reviews object being the largest.

Methodology

Our proposed approach is as follows:

1. Extract latent topics using Topic Modelling and analyze distribution of topics per cluster (i.e. restaurant/geographic region) in the feature/dimension space
2. Assign per-topic rating per review and tip by either a simple average or a weighted average computation of the number of stars with the added context of the positive and negative weights of the neighbor words [3]
3. Predict the overall restaurant rating based on the extracted features by using Supervised Machine Learning techniques (Will also be used to evaluate extracted topics)
4. Analyze quantitatively, the effect of each latent topic on the overall star rating of the establishment. This will also be used to assert the "Volume vs Effect" for each topic, meaning we can find the topics that don't affect the overall star rating despite being in a lot of review text

There are many algorithms that allow us to perform the steps outlined above. A few that we would start with are Latent Dirichlet Allocation (LDA), Online LDA, Labelled LDA and Topics over Time for the Topic Modelling. On the Supervised Machine Learning side of things, we would have to consider the feature space and its dimensionality to pick the best model to run. However, a few models we could work with are Regression, Support Vector Machines or K-nearest Neighbors.

Evaluation

As evident from our approach, a significant portion of the approach is unsupervised which is difficult to quantitatively evaluate. To overcome this shortcoming, we plan to perform standard evaluation metrics like the F-measure, ROC-AUC and Accuracy for downstream supervised machine learning tasks. In addition to this, we will run multiple Topic Modelling algorithms and compare the results of them along with a few metrics to qualitatively explain the intermediate clusters generated.

References

[1] M. Anderson and J. Magruder. "Learning from the Crowd." The Economic Journal. 5 October,2011 Link - https://are.berkeley.edu/~jmagruder/Anderson%20and%20Magruder.pdf

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[3] James Huang, Stephanie Rogers, Eunkwang Joo. "Improving Restaurants by Extracting Subtopics from Yelp Reviews" University of California, Berkeley. Link - https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_ImprovingRestaurants.pdf

[4] Link - https://www.yelp.com/dataset/challenge