Literature Review

Li Liu

April 2019

Identifying unsatisfied customer needs is a crucial part in firms' marketing research. Surveys, interviews, focus groups are examples of traditional methods to understand the customers. In the digital age, customers interact directly with firms and other customers through channels such as e-commerce websites and review forums.

This research project focuses on understanding the factors that lead to customers' dissatisfaction towards the firms. More specifically, we investigate the customers experience and business attributes that would predict restaurants' bad reputation. Topics and sentiments extracted from the reviews are good measures of the subjective and diverse customers experience. Business attributes are characters of the restaurants that are the same for all customers, such as location, food categories, and decoration. Bad reputation is operationalized by low average rating score. We want to develop the algorithms and methods that could enable firms to detect customers' unfulfilled needs and make data-driven reactions.

Yelp (2019) provides a large dataset to students who are interested in mining and analyzing the undiscovered insights from the text reviews, business data, user metadata, and uploaded photos. Many research communities have explored this dataset to conduct empirical research and develop machine learning algorithms. We surveyed some literature that provide the foundation for our research. Broadly, these papers could be clustered into three groups:

1. Methodologies in NLP/Text mining
2. Text data in business/marketing
3. Yelp data for business research

1. Methodologies in NLP/Text mining

Sentiment analysis and topic modeling are two common methods for review data. Since the sentiments are highly correlated with the rating scores, topic modeling is the main method we apply for finding the underlying topics in reviews. Blei et al. (2003) purposed the latent Dirichlet allocation (LDA) which is a three-level hierarchical Bayesian model to represent each document as the probabilistic mixture of a set of topics. In our case, each review text might be talking about topics such as “food”, “service”, “price”. This information is encoded as new dummy features of whether the review is mainly talking certain topic or not. Then we aggregate the features together as new variables. For example, a restaurant might have 70% of the reviews talking about the “price” and average rating of 2, which suggests the pricing might be too high.

Büschken and Allenby (2016) extend the LDA model by restricting each sentence, instead of each word, is only assigned with one latent topic. Their new model achieves improved prediction performance in hotel and restaurant reviews. We plan to incorporate this extension on top of the traditional LDA model.

2. Text data in business/marketing

The advances in NLP and text mining, combined with traditional data and models, have many applications in the business and marketing research. Gentzkow et al. (2019) describe the opportunity of “text as data” by extracting information from documents to serve as new input variables. Product reviews help researchers reveal what drives customers’ decisions (Gentzkow et al., 2019, p. 2).

Balducci and Marinova (2018) provides a overview of using unstructured data (text, voice, image, video, etc) in marketing research. Text reviews, as one common form of user-generated content, is valuable source for capturing marketing insights and creating values for both businesses and customers (Balducci and Marinova, 2018, p.580). They also recognize the research gap in applying unsupervised learning and convolutional neural network to extract information (Balducci and Marinova, 2018, p.582).

User-generated content analyzed with machine learning methods could be a even better alternative to identify customer needs, compared with interviews and focus groups (Timoshenko and Hauser, 2019). Their results show promising potential for us to identify the unsatisfied customers needs after their dining experience from the ML approach.

After briefly surveying the relevant papers mentioned in Gentzkow et al. (2019) and Balducci and Marinova (2018), we want to contribute to the literature by combining unsupervised learning (topic modeling) and supervised learning (predictive models) to improve the pricing strategies and better manage relationships with customers.

3. Yelp data for business research

Researchers have used Yelp data to measure economic and business activities. Luca (2011) estimated the effect of online reviews on restaurant demand by regression discontinuity method. Besides from the Yelp review data, he aggregated it with the restaurants revenue data from the Washington State Department of Revenue (Luca, 2011, p.8). Thus, he operationalized reputation by rating stars and consumer demand by reported revenue. This suggests adding external data source would be useful. One example is Glaeser et al. (2017) as they demonstrate the government survey (they use County Business Patterns (CBP) offered by Census Bureau) could be greatly complemented by the new economic activities metric predicted from Yelp data. For our project, we are uncertain about the availabilities of the revenue data. However, the microdata by the Current Population Survey offers a wide range of information on neighborhoods' demographics. This might help us infer consumers' possible demand and willingness to pay. For instance, a mediocre Asian restaurant might receive good ratings in the suburban areas as people don't have many options for exotic food. By contrast, diners in the downtown area of Chicago have higher expectation of food quality.

Given the unobservable heterogeneity of reviewers, aggregating the rating scores by taking the arithmetic average has potential issues for missing the underlying changes in product quality and prior knowledge (ex. learning from previous reviews and scores). Dai et al. (2012) develop a adjusted weighted average algorithm for aggregating the individual scores by accounting for the reviewers' characteristics and review histories. We plan to incorporate this algorithm by recalculating the restaurants' rating scores.

Reference

Balducci, B., & Marinova, D. (2018). Unstructured Data in Marketing. *Journal of the Academy of Marketing Science*, *46*, 557–590. https://doi.org/(2018) 46:557–590 https://doi.org/10.1007/s11747-018-0581-x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Büschken, J., & Allenby, G. M. (2016). Sentence-Based Text Analysis for Customer Reviews. *Marketing Science*, *35*(6), 953–975.

Dai, W., Jin, G. Z., Lee, J., & Luca, M. (2012). Aggregation of Consumer Ratings: An Application to Yelp.com. *NBER Working Paper*, *No. 18567*.

Gentzkow, M., Kelly, B. T., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *forthcoming*.

Glaeser, E. L., Kim, H., & Luca. (2017). Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity. *Harvard Business School NOM Unit Working Paper*, *No. 18-022*.

Luca, M. (2011). Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School Working Paper*, *No. 12-016*.

Timoshenko, A., & Hauser, J. R. (2019). Identifying Customer Needs from user-Generated Content. *Marketing Science*, *38*(1), 1–20. https://doi.org/10.1287/mksc.2018.1123

Yelp Dataset. (2019). Retrieved April 27, 2019, from https://www.yelp.com/dataset/challenge