*Lavanya Kwatra*

*BrainStation Data Science, Summer 2020*

*YELP REVIEW ANALYSIS*

A Classification of Sentiment through Review Analysis

**EXECUTIVE SUMMARY**

This report explores the social perception of restaurants, using the Yelp Reviews Dataset, with a granular focus on classification of the sentiment conveyed by the text data in the review. This can be used to identify what attributes users consider important, from the words that are used frequently to indicate positive or negative sentiments. Further, based on user history & user interaction in the form of “votes,” “friends,” or “fans,” and other social features, we can employ a social network analysis to find user clusters and implement the optimal segmentation & targeting strategy to increase customer lifetime value for Yelp. In this context, employing hierarchical membership levels with promotional incentives to curate “high quality” user-generated content is preeminent. Redesigning the user experience may assist with better data collection by increasing user interactivity and traffic. These inputs are significant for a model to accurately capture the semantics of social perception through text & user reactions.

**INTRODUCTION**

Founded in October 2004, Yelp is an online platform that publishes crowd-sourced reviews for businesses, along with the recent addition of Yelp Reservations, which is an online booking service. Headquartered in San Francisco, California, Yelp develops, hosts and markets their website ([www.yelp.com](http://www.yelp.com)) as well as the Yelp mobile app. Yelp prides itself on their motto of “connecting people with great local businesses,” thereby, creating a marketplace within their user community to study the social perception of the businesses they advertise. (Yelp, Inc., n.d.)

As a result of this marketplace model, in a way, businesses advertise to new customers through the user-generated reviews on Yelp and rely heavily on consumer engagement, such as “useful votes” on the review another user wrote about their establishment. Yelp generates insights from their online community of users, who write reviews & rate businesses on a scale of 1 to 5 while interacting with other users & the posted content in a social networking format.

A traditional consumer centric approach would focus on tracking user activity levels, such as:

1. the total number of reviews written by the user,
2. whether the user leaves "tips" or suggestions for the restaurant online,
3. whether the user checks into restaurants when they dine there,
4. the amount of time the user spends on the platform,
5. the user’s social interaction with other users through “votes” on reviews
6. whether the user actively finds friends in the user community,
7. whether the user follows top Yelper's or similar users,

The most popular reason people use Yelp is when they are searching for a new restaurant, based on how other users found it. This is captured in the user “votes” on reviews posted by other users in the Yelp community. Based on a study by Saeideh Bakhshi, Partha Kanuparthy and David A. Shamma at Yahoo Labs, a deeper understanding of the social perception of these "votes" as user feedback can help design better recommendation and social networks. Using the same review data from 2013, they found that active and older members tend to engage more on the platform, by writing more reviews and reacting to other reviews. Reviews with a higher word count are perceived as more "useful" overall by the community, while those who give higher ratings are perceived as "cool" and review with a negative tone are more likely to be voted "funny" by other users. Additionally, they found that new businesses can generate a "review momentum" by receiving more "useful" votes on initial reviews by users, which helps establish the business. (Bakhshi, Kanuparthy, & Shamma, 2013)

**DATA ANALYSIS & MODEL EVALUATION**

Yelp’s website has a well-documented dataset of 6 .json files: business, user, review, tips, check-ins and photos. Based on which data is relevant to study the social perception of restaurants in the Yelp user community, review, business and user data was selected. The three files are almost 11GB in size, containing over 8 million reviews for 200,000+ businesses across North America. The three chosen files were read in using Pandas, followed by merging the corresponding business & user attributes to the review text & rating. Next, a representative sample of approximately 500,000 rows was exported as a .csv file to the local directory.

Next, this sample was uploaded to a S3 bucket on AWS to use a Sagemaker Notebook Instance for quicker computation. Initial cleaning steps such as filtering for only restaurants that are still open, followed by dropping columns that are incomplete or irrelevant, followed by string cleaning & feature engineering a column to represent user age. Finally, a sample of 10,000 rows with user, review and business attributes was exported as a .csv file to preprocess, train & fit to our classification model on the local computer. (Note: I have contacted AWS Support to troubleshoot & increase my storage so I can allow my model to learn from more data, as my instance didn’t allow me to perform the required preprocessing steps.)

Notebook 3

Notebook 4

**CONCLUSIONS & NEXT STEPS**

Using a sentiment analysis model to understand the user's opinion of the business from review text & the corresponding rating can be used for topic modelling & identifying key business attributes predicting these ratings. However, while frequently occurring words in reviews may convey a popular trend in the community, a classification of sentiments as positive and negative does not completely capture the overall social perception of the business. Exploring our data, we realize most of the user generated votes are nulls, i.e., many reviews that have no “useful”, “funny,” or “cool” votes at all. Further, the votes are unevenly distributed across our dataset. These findings suggest that Yelp could reconsider user interface design & metrics to improve data collection.

Yelp’s product model is driven by user activity: to generate content, provide data for business insights and as social feedback for a user recommendation system. In this context, it is crucial to filter content that is considered "high quality" from both a business & consumer standpoint. Apart from fake & real review detection, understanding the semantics embedded in text as well as peer evaluation in Yelp's online community is important to drive up the profit reaped by users, businesses as well as Yelp. For instance, our model must be able to classify the tone of the review based on the text. Perhaps, reviews with lower star ratings received more “funny” votes because they had a hint of sarcasm, which some users found amusing.

User generated content can be improved by incentivizing reviews written by old and experienced users, subsequently increasing other users’ confidence in the content. (Bakhshi, Kanuparthy, & Shamma, 2013) This lays the foundation to introduce member “levels”, as seen in several e-commerce channels, to segregate active & experienced Yelper's from new Yelper's. Meanwhile, Yelp can help businesses "start a new conversation between the voter and the reviewer" around a particular topic & thereby, create a meaningful social network. This ties into the E-WOM (electronic-word-of-mouth) marketing concept of using "influence" from popular active & older members to set a trend in the community regarding the social perception of the business. This is Yelp’s value proposition for businesses, by generating an online “buzz” while simultaneously tracking consumer reactions & market trends.

Further, incentives for writing reviews in a social networking model will provide Yelp with a gold mine of information on their users, with several opportunities for expansion, such as a social networking platform for food bloggers & food lovers. To boost popularity among users, Yelp can curate their browsing feed with collaborative content from users’ social media platforms. This has an added benefit of augmenting the images and check-in data collected by Yelp from their users. User interface design features such as graphic call-to-action “vote” reactions have also been shown to increase consumer engagement.

Therefore, an accurate classification of both semantics and sentiments conveyed in text review data using transfer learning language processing models such as Fast AI, along with hierarchical cluster analysis using DBSCAN, followed by classification of positive and negative attributes for each user cluster develops a more robust recommendation system. Subsequently, Yelp’s customer lifetime value & benefit to all stakeholders can be maximized with a well-trained classifier for different clusters, based on their segmentation and targeting decisions. Yelp can realize their goal of matching users to great local businesses, by leveraging user insights in an optimal way.

**BIBLIOGRAPHY**

Bakhshi, S., Kanuparthy, P., & Shamma, D. A. (2013). *Understanding Online Reviews: Funny, Useful or Cool?* Retrieved from Yahoo Labs Inc.: https://kanuparthy.files.wordpress.com/2016/05/cscwf630.pdf

Yelp, Inc. (n.d.). *About: Yelp*. Retrieved from Yelp: https://www.yelp.ca/about