CSE 6242: Data & Visual Analytics
Fall 2018
Project Proposal - “Yelp me out here!”
Team #36 – The Plug-ins
October 15, 2018

Problem Statement:

We aim to develop a resourceful business platform that would enable existing business owners and entrepreneurs to establish and flourish their businesses at a healthier and faster pace. Our group seeks to eliminate every merchant’s inconvenience caused by inefficiently aggregating and analyzing information regarding customer needs, patterns and satisfaction quotient, dynamic logistic requirements, new outlet location etc. Merchants currently lack the tools and resources to effectively drive their business.

Current Practices:

Impact of Local Experts on review trends

An expert review influences consumer demand for experiential goods and is positively correlated, Reinstein et. al.[1].

The Yelp Dataset also yields information about local experts. Jindal[2] says that the frequency with which a reviewer reviews business establishments can help identify if the user is an expert in the particular area.

Yilin He[3] focuses on identifying influential users by finding relationships between users who reviewed at a later time, than the influencer.

Business Impact of Reviews

Rating increase in Yelp is usually accompanied by increase in Revenue. Michael Luca[4]

Customers react better to well-presented information. Michael Luca[4]

Frequently reviewed restaurants appear to be clustered close to each other and mostly are located in popular locations. Their rating system is dependent on each other as well. Peter Hajas[5]

An empirical way to model a review filter to overcome fake reviews, also model the nature of the business and the conditions under which the businesses are more likely to fake reviews. Michael Luca[6]

Understand how yelp dataset can be put to practical use while setting up new businesses[7]

Using yelp reviews to improve restaurants by understanding what features to add to the business[8]

Measuring User's influence in the Yelp recommender system[9]

Sentiment Analysis

Sentiment analysis is the field of study that analyzes people's opinions, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, and their attributes [10].

There are three levels of sentiment analysis referred to by Collomb et al. [12], Liu [10] and Medhat et al. [11], as document-level, sentence-level and aspect-level[13].

The machine learning method uses supervised learning techniques to determine sentiment by training a known dataset. Lexicon-based sentiment analysis involves calculating the sentiment polarity of a piece of text. Statistical methods of sentiment analysis were also described by Collomb et al. [12] where these methods represent each review as a combination of aspects and ratings. Review ratings can be predicted just from the review text [14].

Visualization

Visualisations are an integral part of our project. There have been past works which talks about effective visualisations. Lee et. al.[15], and Vliegen et al.[16] both indicate the red flags in a visualisation that makes it difficult to extract information out from it. As mentioned in Wijk[17], effective visualisations should enable the users to generate useful and actionable insights.

Classification on Textual features

This paper evaluated various supervised learning models and concluded the superiority of using Logistic Regression using unigrams and bigrams based on experimental evidence and provided the platform to estimate a rating based on textual review. Nabiha Asghar[18]

Innovation:
Most developmental efforts in the direction towards business analysis are consumer-centric and focus on ensuring that consumers have a variety of information available to choose the right merchant. Unfortunately, merchants do not have access to organized and logged consumer preferences and patterns across various geographical locations to enhance their business strategy. We seek to fulfill this void by our application. We have incorporated a variety of features ranging from, but not limited to, status and potential of growth of any business assessed geographically and consumer preference wise, peak hour analysis to manage logistics better, correlation of secondary factors such as location, ambience, valet parking, free Wi-Fi, cuisine preference etc with the performance of business and impact on business rating by sentimental analysis of a review.

We will be utilizing Yelp dataset to produce worthy, expressive analysis of current trends and patterns that could enable merchants to factor in variables in decision making that they would have otherwise ignored and also improve quality of their services by looking at out intuitive visualizations rather than digging through multiple review websites. Even an entrepreneur could employ this platform to make fundamental decisions such as finalizing location on the basis of analyzing geographical distribution of preferred locations for similar businesses.

Stakeholders:

Entrepreneurs, small and medium size business holders and even large scale corporations would benefit immensely from out application. We expect a cost cutting in business investment and provide merchants a concise list of dos and don'ts based on user feedback.

Impact:

An application that could enable a merchant to know immediately if customer was dissatisfied, to anticipate the staffing needs in advance and serve accordingly based on peak hour analysis has great benefits. Even entrepreneurs can use the data to decide where to start their business, what features to include and ground truth data can be used to measure how the profit was affected by the move.

Costs/Risks/Payoffs:
The dataset is publicly available hence no cost there. The primary cost will be in the time spent during development.

Since the features taken into consideration can be correlated with one another more tightly than anticipated, focusing on one-to-one relationships can lead to oversimplification and incorrect results. The platform must be seen as a recommendation model and not as source of truth. However, the application has huge potential to provide key insights on matters that may get overlooked occasionally and justifies this whole endeavor.

Plan of Action:

Participants	**Week 1-2**	**Week 3-6 (midterm check)**	**Week 7-8 (finalpoint check)**	**Week 9-10**
	Data manipulation and DB & project setup	**Developing Metrics**	**Metrics cont. and UI/UX**	**Deployment and Testing**
Aastha and Anirudh	Clean data	Peak hour analysis metric	Impact of peer and expert review metric	Testing
Archit and Havish	Design Database schema and insert data	Geographical distribution of business metric	Sentimental analysis metric	Debugging and polishing
Manu and Sangharsh	Environment setup	Correlation to secondary factors metric	Implementing design details	Gathering feedback and improvising

References:

[1] The Influence Of Expert Reviews On Consumer Demand For Experience Goods: A Case Study Of Movie Critics
[2] Finding Local Experts From Yelp Dataset
[3] Identifying Influential Users in Social Network with Review Data
[4] Reviews, Reputation, and Revenue: The Case of Yelp.com, Michael Luca
[5] Analysis of Yelp Reviews, Peter Hajas
[6] Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud, Michael Luca
[7] Restaurant Setup Business Analysis Using Yelp Dataset
[8] Improving Restaurants by Extracting Subtopics from Yelp Reviews
[9] Measuring user's influence in the Yelp recommender system
[10] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
[11] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4):1093-1113, 2014.
[12] A. Collomb, C. Costea, D. Joyeux, O. Hasan, and L. Brunie. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. Technical Report RR-LIRIS-2014-002, LIRIS UMR 5205 CNRS/INSA de Lyon/Universite Claude Bernard Lyon 1/Universite Lumiere Lyon 2/Ecole Centrale de Lyon, Mar 2014.
[13] C. Rohrdantz, M. C. Hao, U. Dayal, L-E. Haug, and D. A. Keim. *Feature-Based Visual Sentiment Analysis of Text Document Streams. ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Visual Interfaces for Text Analysis*, 3(2):26:1–26:25, Feb 2012.
[14] Yelp Dataset Challenge: Review Rating Prediction
[15] : Sukwon Lee ; Sung-Hee Kim ; Ya-Hsin Hung ; Heidi Lam ; Youn-Ah Kang ; Ji Soo Yi - How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking
[16] : Roel Vliegen ; Jarke J. van Wijk ; Erik-Jan van der Linden - Visualizing Business Data with Generalised Treemaps
[17] : J.J. van Wijk - The value of Visualization
[18] Yelp Dataset Challenge: Review Rating Prediction, Nabiha Asghar.