# What are people thinking about your business? Analysis of Yelp reviews

Vishal Sharma
Utah State University
vishalusu@gmail.com

Ruchi Chauhan
Utah State University
r.chauhan@aggiemail.usu.edu

## ABSTRACT

The importance of data science and its insights for businesses have long being realized. As the business move more towards customer-centric models, the understanding of customer preferences is important more than ever to provide customised experience. Our project aims to draw insights about the customer behaviour based on their reviews on Yelp and help business owners improve their business.

## Keywords

Yelp; User behavior; Business improvement; Prediction

## 1. INTRODUCTION

Yelp started in 2005 as a platform for users to rate and review local businesses (majorly restaurants). Business owners create their profile and users rate the business from 1-5 stars. A user can write a review, can perform a Geo checkin, write a small tip/advices. Other users can view reviews and if they find information (reviews, tips, etc) useful they can vote on helpful, funny, etc. Currently, yelp has a massive amount of business data, businesses can view their current reviews and manually try to figure out what can be helpful. Reading hunderds of thousands of reviews for a business is not very practical and there is no automated system which can help them in such process. In this project, we focus on above challenge and try to help business owners improve their business by giving them detail insights of the reviews and user behaviors.

## 2. DATASETS

The dataset is downloaded from the yelp dataset challenge, 2016. This massive dataset contains information about local businesses in

- 10 cities across 4 countries.
- 2.7M reviews and 649K tips by 687K users for 86K businesses 566K business attributes, e.g., hours, parking availability, ambience.
- Social network of 687K users for a total of 4.2M social edges.
- 200,000 pictures from the included businesses.

Countries and Cities included: U.K.: Edinburgh, Germany: Karlsruhe, Canada: Montreal and Waterloo, U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison.

## 3. PROJECT SKETCH

We plan to solve the following problems:

- Visualizing word cloud of reviews and tips data to find useful information. What people are talking about business and if they have any concerns about Price, Parking, Wifi, Location, Environment etc.
- Visualizing check-in data to see trends (eating habit, time, etc.) in different cities using tableau or Bokeh or Seaborn in Python.
- Using reviews and check-ins, to predict how many customers a business can have in a day and analyzing it over different cities. This can be achieved using Poisson Distribution. In probability theory, a Poisson process is a stochastic process that counts the number of events in a given time interval cite[wiki].
- Finding experts on Yelp, using user's Reviews, Friends, Votes, Compliments and Fans data. We can also try finding the communities of most influential users and analyze for any correlation.
- Finding negative reviews using sentiment analysis and then performing Latent Dirichlet Allocation (LDA)[cite] to find topics/words in the reviews which can be useful to the business to improve.

Since this is a classroom project we are not sure if we will be able to solve all the problems listed above. We will try our best to get most of it done.

## 4. TOOLKIT

We plan to use Python and its libraries for developing our project. Few listed below:

- Cleaning and Processing : NLTK, Pandas, Numpy, MongoDb and Hadoop
- Modelling : Scikit-Learn and Weka
- Visualization : R, MatplotLib, Tableau and Seaborn

To process such large data, most of our processing will be on either Multithreaded or Hadoop environment.

## 5. CONCLUSION

We will try to find interesting facts and figures from yelp dataset. The insights gained from the dataset can help develop a model that can be used by small businesses at micro level to make changes as necessary. The end product will improve user experience and quality of business.