Classification and Sentiment Analysis on Yelp Dataset

# Introduction

As one of the most popular platforms which aim to connect people with local businesses, Yelp highly relies on it review and rating system. Yelp’s dataset provides business with both text data (reviews) and non-text data (other attributes such as address, hours, ambience, etc.). As a business management team, how should they allocate the budget and time to the data analysis to get the most benefit? Should they focus on only text data (review) analysis, only non-text data analysis or both?

This project is trying to help business answer these questions through classification and sentiment analysis on the Yelp Dataset. The main techniques include sentiment analysis (n-grams) and classification (Naïve Bayes, Random Forest and Multinational Regression). Python and related packages (e.g. NLTK) will be used for solving the problem.

# Literature Review

1) Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews (by Boya Yu, Jiaxu Zhou, Yi Zhang, Yunong Cao)

This article introduced a machine learning based method to help identify main features of different types of restaurants. By analysing the customer reviews through word frequency and polarity, the author identified the significance of words in terms of the corresponding type of restaurants. Also, it is concluded that customers tend to valuate service over other aspects when they rate a restaurant.

2) Business Reviews Classification Using Sentiment Analysis (by Andreea Salinca)

This article tries to provide a better method to automatically classify customers’ sentiments towards a business by using only the text review from Yelp. Rating is classified as either positive or negative. The author performed experiments on various combinations of several feature extraction methods including unigrams and removing stop words, and identified the classifier with highest classification accuracy.

3) Prediction of Rating Based on Review Text of Yelp Reviews (by Sasank Channapragada)

The author tries to identify a way to predicate business rating only based on text reviews. Three approaches including 1000 common words, 1000 common words with stemming and 1000 common adjective words are used to train the data and test the prediction accuracy. Same methods are used to see the result of further predicting if the rating is considered as funny, useful or cool. Results show that only text review itself is good at predicting the rating but may not be enough to further predict how other people see the review.

4) Sentiment Analysis of Yelp Reviews: a Comparison of Techniques and Models (by Siqi Liu)

The article shows the comparison results of using different natural language processing (NLP) techniques on sentiment prediction models. It shows that some NLP techniques like bi-grams, normalization have positive effect on model performance. And simpler models like Logistic Regression are more effective at predicting sentiment than other more complex models.

5) Yelp Review Rating Prediction: Sentiment Analysis and the Neighbourhood-based Recommender (by Rui Qiao)

This research covered two approaches to predict review ratings: sentiment analysis and user-specific recommendation. NLP methods are used to perform a binary sentiment analysis on the text review to help predict positive and negative rating towards a restaurant. The author argues that as similar users have similar patterns in rating, it is possible to build a recommendation system based on reviews from same community’s users. And this helps to recommend business to new users from the same community.

6) Yelp Review Rating Prediction: Machine Learning and Deep Learning Models (by Zefang Liu)

This article focuses its analysis on comparing different machine learning and transformer-based models in rating prediction from text reviews. Besides classification analysis, the author applies models like BERT which created for natural language understanding to rating prediction. The result shows that transformer-based models may generate high accuracy on text review analysis and the prediction based on it.

7) Sentiment Analysis of Yelp’s Ratings Based on Tex Reviews (by Yun Xu, Xinhui Wu, Qinxia Wang)

The author of the paper tries to compare the performance of algorithms on predicting rating based on text review. A public dictionary of words is used as a basic classification features. Then stop words and stemming are added to see the impact on the performance. Results show that adding stop words and stemming can improve the algorithm performance.

# Dataset

The dataset used in this project is the Yelp Open Dataset which can be downloaded from <http://www.yelp.com/dataset>. The dataset has five JSON files including over 8 millions reviews and 209,393 businesses.

For the purpose of the project, only the business and review dataset will be merged and used. Business dataset contains business data including location data, attributes, and categories, while review dataset contains full review text data including the user\_id that wrote the review and the business\_id the review is written for. Business\_id will be used as the key to merge two datasets. Also, as the dataset is large, this project will use “Restaurants” business data filtered by “categories” of the business dataset.

# Approach

## Step 1: Data Preparation

1. Download Yelp Dataset from <http://www.yelp.com/dataset>.
2. On Python, install necessary packages including pandas, tqdm and import pandas, tarfile, tqdm, json, csv, numpy, time for future use.
3. Extract Business and Review JSON files from the downloaded tarfile.
4. Convert review dataset to data frame and save as csv file.
5. Convert business dataset to data frame and save as csv file.
6. For business dataset, split column ‘attributes’ into separate columns and clean the dataset
7. Merge two dataset into one by using business\_id as the key.
8. Use source code to filter sample dataset based on business “categories” column.

## Step 2: Analysis only based on text review data

1. Focus analysis only on the text review content. Apply necessary NLP techniques such as bi-grams, removing stop words, normalization
2. Identify a list of words from the text review as classifier
3. Split dataset for training and testing
4. Apply Naïve Bayes, Support Vector Machines and Multinational Regression by using the classifier we identified
5. Compare the performance with F1 score and confusion matrix

## Step 3: Analysis only based on business attributes

1. Focus analysis only on the business attributes and use it as classifier
2. Split dataset for training and testing
3. Apply Naïve Bayes, Support Vector Machines and Multinational Regression with the classifier
4. Compare the performance with F1 score and confusion matrix

## Step 4: Analysis based on both text review and business attributes

1. Combine step 2 and step 3 by using both list of words and business attributes as classifier
2. Split dataset for training and testing
3. Apply Naïve Bayes, Support Vector Machines and Multinational Regression with the classifier
4. Compare the performance with F1 score and confusion matrix

## Step 5: Comparison of three approaches and final conclusion

1. Compare performance of the results from step 2, 3 and 4
2. Identify the difference and impact of different approaches
3. Provide final conclusion and answer the project question: which approach business should use to get a better understanding of their business
4. Identify possible future improvement