# Capstone Proposal

## Ankur Kothari

**Domain Background:** Due to advancement in technology and cheap availability of the same to almost everyone, the dependency of people on internet is very high, and for every suggestion or problem, we are highly used to refer to internet. This varies from asking the solution of the problem, to check the routes, to check the timings of all sort of things happening in the world and to check whether a place is good or not. YELP is similar site, where we can give and check the reviews and ratings of all types of things ranging from hotels, restaurants, Salons, doctors, food, entertainment and other type of night life.

If someone is new to a city or a country itself (like me), then to get an idea about all above things is very important. But there are incidents when ratings and reviews don't go along, and this may cause ambiguity. Hence to resolve this issue, there should be an algorithm which can read a review (since it is not possible for a person to go through thousands of reviews) and can determine the ratings based on all those reviews, and my motive of doing this project is to solve the above issue. The papers referred for above work are referenced as:

https://webcache.googleusercontent.com/search?q=cache:JA_716wjevcJ:https://bcourses.berkeley.edu/files/65096735/download%3Fdownload_frd%3D1%26verifier%3D85rcoGv9spBYODYklBn63hEXhgqb3tRZfkGBUiFO+&cd=9&hl=en&ct=clnk&gl=us (Prediction of useful reviews on yelp dataset)

http://www.ics.uci.edu/~vpsaini/ (Classifying Yelp reviews into relevant categories**)**

**https://arxiv.org/abs/1605.05362 (**Yelp Dataset Challenge: Review Rating Prediction)

**https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Wa'el_Farhan.pdf** (Predicting yelp restaurant reviews)

Link for source data: https://www.kaggle.com/c/yelp-reviews/data

**Problem Statement:** When the data of yelp for restaurants was checked, there were around 5.2 million reviews and related ratings. For a person, its impossible to go through all the reviews, and the ratings could be highly misguiding, for example a customer can give rating of 2/5 because food was excellent and ambience was perfect but there was no WIFI and discount, while other can also give 2/5 because the restaurant had high speed WIFI with some discounts, but food was terrible. Both these ratings are not relevant, and hence to solve this issue, we can use machine learning which will go through the text, extract important features relevant for the particular domain (food for this case) and averages the ratings based on above.

The problem is related to ratings for the restaurant, which is a discrete value with labels from 1-5, hence this is multiclass classification. The input (features) for the algorithms will be words which will be extracted from the review and output will be ratings from 1-5. Since there is a huge data available, hence machine learning will perfectly help to solve this problem. The data is updated everyday, and the same algorithm can be used for the latest ratings. There are many sites with ratings and reviews like Amazon, Zomato, Grub Hub, TripAdvisor etc, and the same algorithm with minor changes can be used for all such reviews systems.

**Datasets and Inputs:** The Yelp review dataset to be used for this project was downloaded from Kaggle (Yelp Business Rating Predictions - https://www.kaggle.com/c/yelp-reviews/data). The dataset was downloaded as a CSV file, which was then read with the help of Pandas. When the shape of the data was seen, it was as

(5261668, 9), means there were almost 5.2 million rows, each row referring to one review and there were 9 columns.

The columns are as:

**review_id** **user_id** **business_id** **stars** **date** **text** **useful** **funny** **cool**

From the above columns, the input feature which will be used for predictions are words extracted from the text section, while the output labels will be ratings from 1-5. This contains the review for the restaurant, for example: **Super simple place but amazing nonetheless. It....** There are already few works as mentioned in Domain background section (references). The review is directly related to the rating, for example if the review has words like good, nice, excellent, then this means higher ratings, while is review has words like bad, horrible etc then the ratings will be low.

**Solution Statement:** To solve the problem of ratings, we can extract features from the reviews, the features are the words and their occurrence, and this features then can be used to predict labels or the ratings from 1-5. We have a huge data, and this much data is enough for thoroughly understanding the feature and ratings relationship and will work with very high accuracy.

**Benchmark Model:** The accuracy in Yelp Dataset Challenge: Review Rating Prediction, by Nabiha Asghar was 64% with Logistic Regression. The benchmark follows accuracy score (I could not find any model using F1 score), but for my model, I will be evaluating using F1 score. My motive for this project will be to try my best to get highest possible score and to make the model more accurate. The solution for my data will also have 80:20 train test ratio as taken by this benchmark model.

**Evaluation Metrics:** The F1 score will be used for the evaluation of the model. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in y_true. Below is the table representing the number of reviews per ratings from 1-5. It can be seen that the data is highly skewed with higher number of reviews towards higher ratings.

| | |
|---|---|
| **1** | 731363.0 |
| **2** | 438161.0 |
| **3** | 615481.0 |
| **4** | 1223316.0 |
| **5** | 2253347.0 |

The F1 score works as:

F1 = 2*precision*recall/(precision + recall)

| | p' (Predicted) | n' (Predicted) |
|---|---|---|
| p (Actual) | True Positive | False Negative |
| n (Actual) | False Positive | True Negative |

It is the hormonic mean for precision and recall where precision is number of correct positive results divided by the number of all positive results returned by the classifier (true positive / true positive+ false positive)

while recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive, true positive/ true positive + false negative).

**Project Design:** The project will be divided into following stages:

- Pre-processing
- Feature extraction
- Supervised learning
- Performance metrics and evaluation

**Pre-processing:** A review has many excess words which are not actually relevant for the rating, will have punctuation marks and the words can be a mixture of both small and capital letters. So, with the help of vectorizing the dataset, and using stop words, all the punctuation marks, irrelevant stop words like the, in, and etc will be removed. This will also change the whole data to small caps.

**Feature extraction:** First unigram method (bag of words) will be used, where each word is considered as one feature. Here a dictionary will be made where the frequency of each word will be paired with that word. This will be followed by TFIDF, where the word will more frequency will be given less weightage and vice versa. Then I will be using unigram and bigram both, where bigram will be in the word pair, for example recommend will get positive rating, but not recommend together will give negative rating. This will further increase the overall accuracy. I will also include stemming so that words with same meaning but different tenses like seeing and see can be minimized.

**Supervised learning:** There will be total 5 models which I will be using for this dataset, they are as:

- Naïve Bayes
- Logistic regression
- Support Vector Machines
- Decision tree classification
- Random Forest classifier

All these models will be used with grid search for optimizing the parameters for the models.

All the above models works excellently for multiclass classification, and hence will help to resolve the given dataset too.

**Performance metrics and evaluation:** The evaluation of the models will be done with the help of F1 score.