**Project Proposal**

**Yelp Analysis Group**

Team: Yelp Analysis

Team Member Names: Krupa Nagori, Anthony Chestnut, Qais Elian, Edgar Lorenzo

**Motivations:**

This project proposal is motivated in two parts for our project team. First, as students first of data mining and now applied data science we want to further explore the concept and implementation of machine learning and natural language processing. With this concept in mind the team focused on looking through datasets that would provide the base for the project. The Yelp Dataset provided by Yelp contains business metadata as well as reviews for the businesses. The motivation in choosing this dataset was due to the types of text we would be able to analyze and the variety of questions we would be able to address.

The motivation to implement this project from a business perspective comes from the interest of the organization to increase the quality of the rating system utilized by users to filter through a variety of restaurants and shops to determine which best fits their needs. The overall goal of this project would be to provide greater insight into the data the organization has collected and potentially provide methods to implement this data in new or interesting ways. The text of the reviews provided will be analyzed using natural language processing to identify information about the review from the type of the rating should be given to whether the rating should be positive or negative.

**Hypotheses:**

The first hypothesis would be that there will be a strong correlation between the text found in positively worded reviews and positive ratings and similarly for negatively worded reviews and negative ratings.

The second hypothesis is that the dataset will contain reviews and ratings that do not correlate with each other. This could potentially identify incorrectly completed reviews.

**Technical Approaches:**

We will be creating a classification algorithm in Python to extract a sentiment from text reviews on Yelp and attach a prediction to that sentiment. We will compare that prediction to the actual yelp review rating given with the analyzed text. We will be building various text mining models to analyze the sentiment in the text portion of the Yelp ratings and categorize that sentiment by giving it a 1-5 rating. Through cross-validation and various text mining and analysis techniques (that are yet to be determined) we will be selecting the classifier that has the closest predictions to the actual Yelp ratings. We will be splitting the data given to us into training and test portions to reduce the risk of overfitting.

**Data:**

For this project we will be using the Yelp Dataset provided by Yelp.  The Dataset consists of the following:

* Yelp\_Dataset\_Business - JSOn file consisting of Businesses and attributes related to the business such as times open, reservation, kid friendly.
* Yelp\_dataset\_Checkin - Check Ins on a business.
* Yelp\_dataset\_Review - Contains full review text data including the user\_id that wrote the review and the business\_id the review is written
* Yelp\_dataset\_User - User data including the user's friend mapping and all the metadata associated with the user.
* Yelp\_dataset\_Tip - Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

**Risks and Risk Mitigation Strategies:**

Availability, sufficiency and quality of data are risks yet to be assessed. A proper size and quality data set seems to be available as part of the challenge announcement. Further extraction, cleansing or transformation techniques could be require to prepare the data to the modeling process. Data segmentation could be required as part of the analysis considering that text sentiment/meaning probably depends on the context (external dimensions to the service being rated as category of service, geographic, language or cultural variables). Early data descriptive analysis should be made in order to prevent data inadequacy.

**Goals and Deliverables:**

Our goal is to build a model with text sentiment analysis and classification component that can predict Yelp ratings based on the sentiment in text portion of the rating. We will be selecting and submitting the model that has the closest predictions to the actual Yelp ratings given by real users. The classifier can be used to flag possible incorrect ratings based on the sentiment outlined in the text and could be expanded to detect “fake” ratings based on the mismatch between the textual and numerical portion of the rating (but that’s outside the scope of this project). We will be presenting our final model and testing results to the class and we will possibly submit the model as an entry in the Yelp Dataset Challenge.