**ML Project Proposal**

**Predicting Customer Sentiment from Yelp Reviews**

1 **Introduction**

Yelp is a popular online platform for users to search for local businesses, read and write reviews about them, and find deals. It provides information on various types of businesses including restaurants, shops, beauty salons, home services, and more. Yelp has become a trusted source for local recommendations, with millions of reviews contributed by its active community of users. Therefore, it has become significantly important for the Yelp to predict the sentiment behind the text reviews. We aim to build a classifier to perform positive/negative classification from the text reviews. We have divided 80% of our data as training data set and the rest 20% as testing data set.

2 **Problem**

Given a review of any business, our model should classify this as positive or negative sentiment. The dataset contains reviews as a text string and target contains 1 or 2 that represents negative and positive sentiment respectively.

**3 Input**

The input is a csv file containing reviews from the customers as and the sentiment behind reviews. We will map the target input reflecting 0 as positive and 1 as negative.

**4 Output**

The output will be 0 or 1 reflecting 0 as positive and 1 as negative.

**5 ML Technique**

We are planning to use various ML model such as Multinomial Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Support Vector Machine and ensemble methods such as Random Forest and XGBoost. After loading the dataset in pandas Data frame, we plan to use multiple encoding method such as Tf-IDF and skip-gram encoding. After training, we will evaluate models using cross-validation and metrics such as Precision, Recall and Accuracy among others.

**6 Dataset**

The detailed description and the dataset itself can be found under the following URL:

Dataset source: <https://www.kaggle.com/datasets/irustandi/yelp-review-polarity>

Data points: 560000 (Training Data + Test Data)