**Project Proposal by Yan Zhu**

As a scientist, I understand that machine learning can have far reaching impacts on society. Machine learning is a highly adaptable tool that can use big data to improve efficiency and accuracy in a wide range of disciplines, as well as in multiple facets of life. Seeking to maximize impact, I chose a topic that has high relevancy for every person in the world: eating. While depending on your personal and cultural habits you may eat a different number of meals each day, the bottom line is that everybody eats. When dining out, Yelp is a widely used tool for determining where that next meal will come from. For my project, I am using machine learning to create an automated algorithm for predicting the usefulness of a Yelp review. What makes my algorithm unique is the inclusion of useful words as a factor.

These days, people’s decisions on where to eat breakfast, lunch, or dinner are highly influenced by online reviews in platforms like Yelp. The number of reviews a restaurant has, in combination with a high rating, is often an indication that a restaurant is consistent and good. However, when a popular restaurant has hundreds or even over a thousand reviews, it becomes difficult for consumers to quickly identify the reviews with the most useful information. As a result, finding the perfect restaurant for a special occasion requires considerable research, and restaurants may lose potential customers or miss important feedback because the most useful reviews are not displayed first.

While Yelp does automatically sort reviews, the sort does not appear to take into consideration when useful words are mentioned. Yelpers have the ability to review their peers by marking reviews as Useful, Funny, and Cool. Also, reviews of friends typically appear first. However, if you are seeking a restaurant unknown to your social circle, or if you are somewhere without a strong Yelp community, these factors are not as important. What is most important are the actual words being used to describe the food. Therefore, Yelp would benefit from using an automated algorithm that incorporates useful words to predict the usefulness of a review.

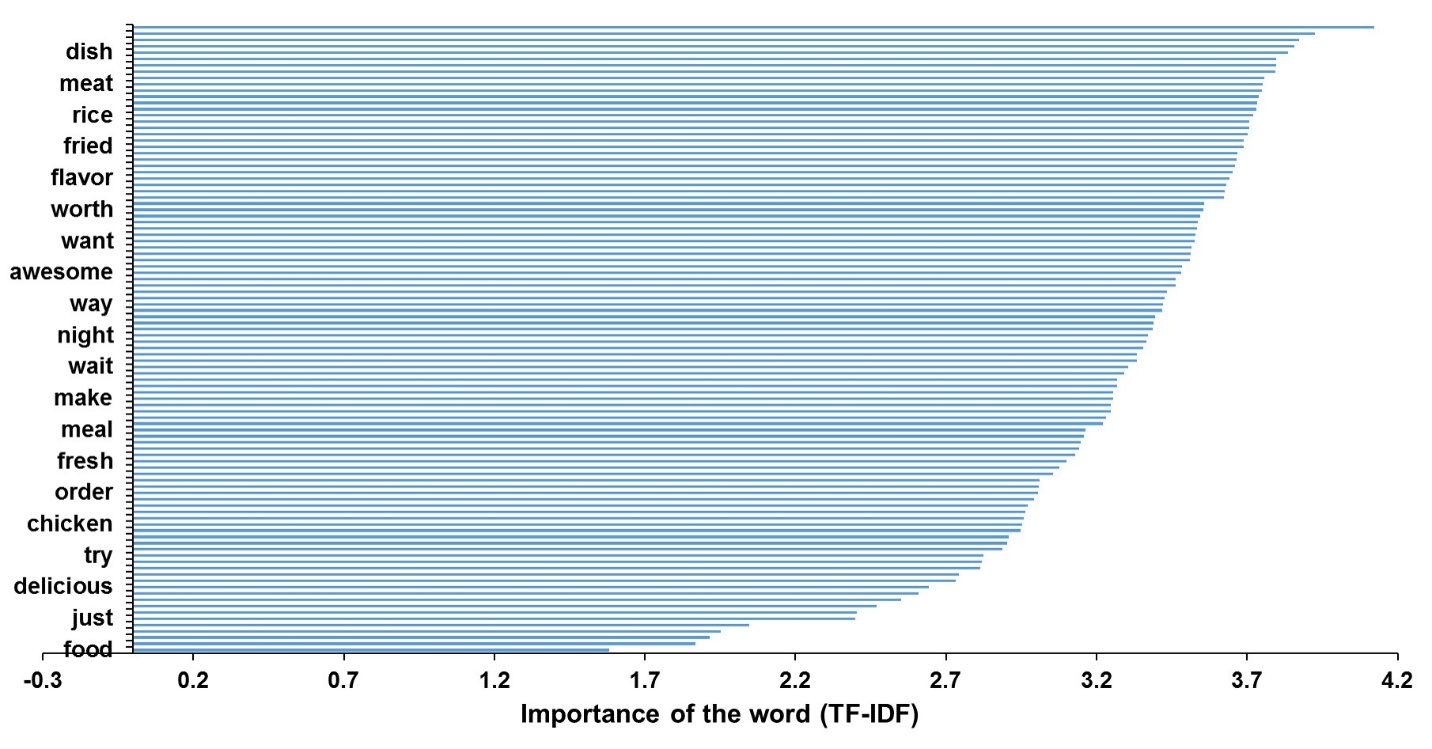
The project utilizes natural language processing (NLP) approach to feature select the importance of words used in reviews by term frequency-inverse document frequency (TF-IDF). The program tests different machine learning models (e.g., Logistic Regression, Random Forest Classifier, Linear Discriminant Analysis, K-Neighbors Classifier, Decision Tree Classifier, Naive Bayes Classifier) to assign an accurate category of usefulness to each review. Usefulness is categorized into Not Useful, Slightly Useful, Moderately Useful, Highly Useful, and Extremely Useful based on the number of times a fellow Yelper marked the review as Useful. The prediction algorithm is based on restaurant rating, quantitative features of the reviews (e.g. importance of words, review length, number of pictures in the review), and elapsed time since the review was posted.

As proof of concept, I scraped ~100 Mb of data on Yelp reviews on Orlando, FL restaurants from 2005 to 2017. The data includes 907 restaurants and 134,491 reviews. I cleaned the data by extracting the features as described above. I then generated a matrix that shows the most frequently used words in each review (**Figure 1**), along with other quantitative features. It is worth noting that the top most frequently used words were the names of common dishes, such as pizza, burger, sushi, fries, steak, followed by adjectives and verbs, such as happy, bad, sweet, awesome, excellent, tasty, visit, recommend, love, wait, want. Some nouns, such as price, atmosphere, experience, and staff, also frequently appeared in the reviews.

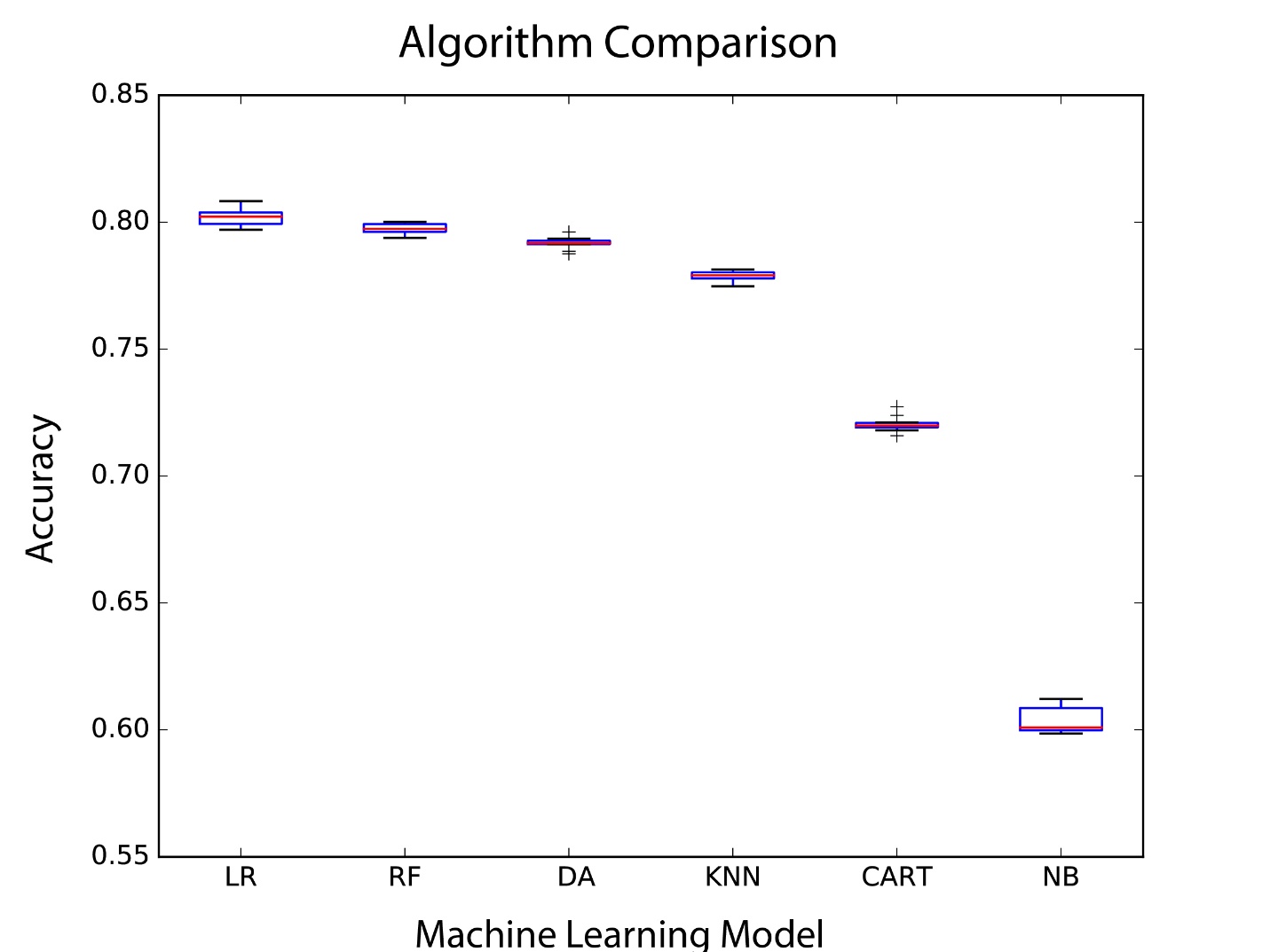
Next, I split the dataset into 80% training set and 20% validation set. To determine which ML algorithms would perform better, I tested the accuracy of each ML prediction on a mixture of simple linear and nonlinear ML algorithms with a 10-fold cross validation on the training set (**Figure 2**). The Logistic Regression model yielded the best accuracy (80.19%).

My next step was to run a Logistic Regression model on the unseen validation set, which resulted in a final accuracy score of 80.55%. Finally, I calculated the logistic regression coefficient for the most frequently used words, restaurant rating, review length, number of pictures in the review, and elapsed time since the review was first posted to determine which variable is most important in the Logistic Regression model prediction (**Figure 3**). Interestingly, the frequency of words like fries, burger, atmosphere, favorite, and flavor played a more important role in predicting usefulness of a review than restaurant rating, review length, number of pictures in the review, and elapsed time since the review was first posted (labeled in red).

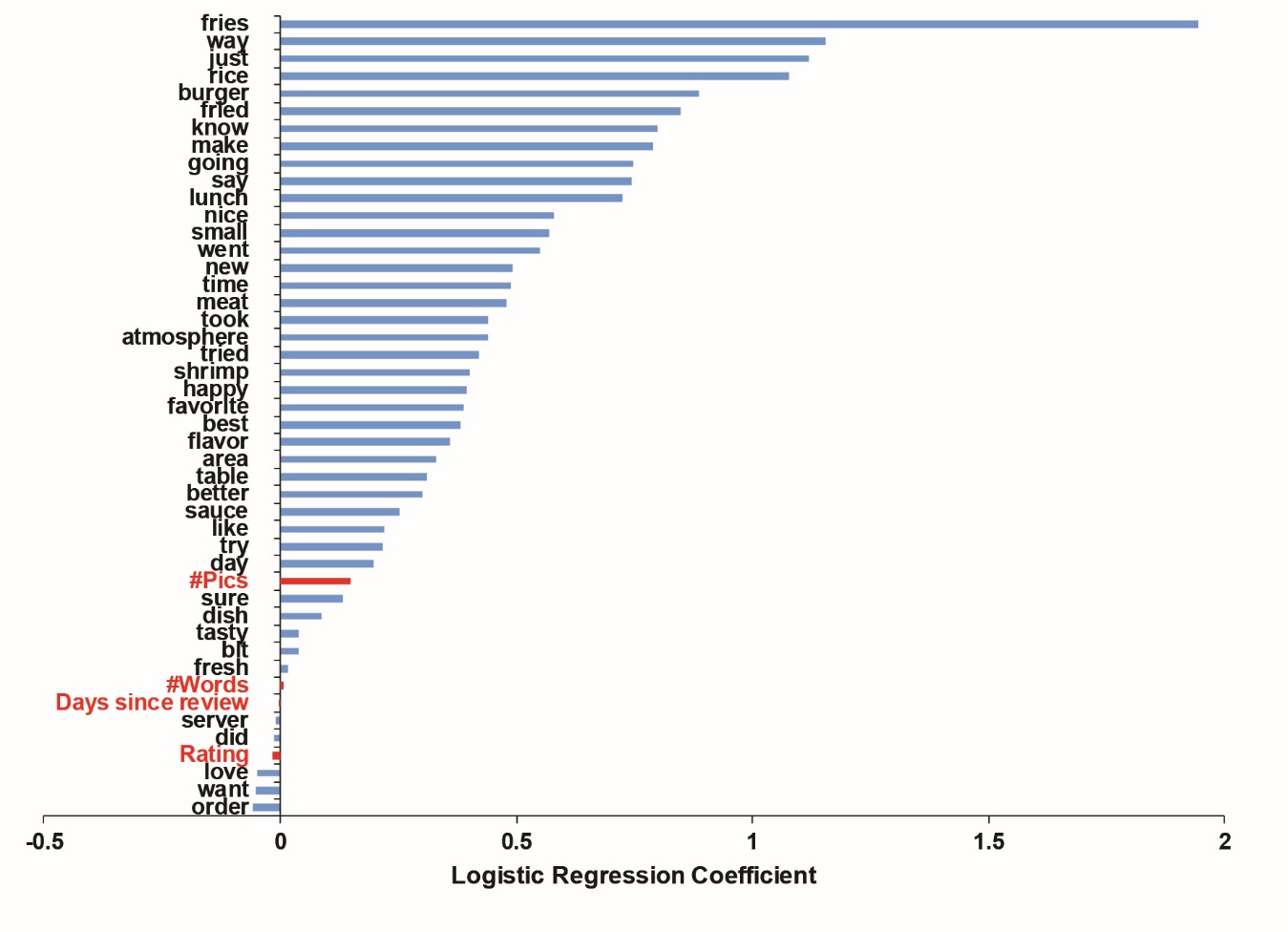
My preliminary results indicate that my proposed project would be successful. As aforementioned, the Logistic Regression model yielded high accuracy. Furthermore, by calculating the logistic regression coefficients, I determined that the frequency of useful words was the most important factor in predicting usefulness of a Yelp review. I hope to be given the opportunity to continue this promising momentum as a Data Science Fellow.



**Figure 1.** Term frequency-inverse document frequency of 100 words extracted from 134,491 reviews representing the importance of each word in context of entire reviews. Only 20 words (out of 100) are labeled on Y-axis for illustration purpose. Top 5 most frequent words are pizza, burger, sushi, fries, dish.



**Figure 2**. Accuracy of different machine learning models (Logistic Regression (LR), Random Forest Classifier (RF), Linear Discriminant Analysis (LDA), K-Neighbors Classifier (KNN), Decision Tree Classifier (CART), Naive Bayes Classifier (NB)) generated from training set.



**Figure 3.** Logistic regression coefficient for most frequent words, restaurant rating, review length, number of pictures in the review, and elapsed time since review first posted.

Figure 1 <https://github.com/vincent625/data_incubator_challenge/raw/master/Figure%201.jpg>

Term frequency-inverse document frequency of 100 words extracted from 134,491 reviews representing the importance of each word in context of entire reviews. Only 20 words (out of 100) are labeled on Y-axis for illustration purpose. Top 5 most frequent words are pizza, burger, sushi, fries, dish.

Figure 2 <https://github.com/vincent625/data_incubator_challenge/raw/master/Figure%202.jpg>

Accuracy of different machine learning models (Logistic Regression (LR), Random Forest Classifier (RF), Linear Discriminant Analysis (LDA), K-Neighbors Classifier (KNN), Decision Tree Classifier (CART), Naive Bayes Classifier (NB)) generated from training set.

Figure 3 <https://github.com/vincent625/data_incubator_challenge/raw/master/Figure%203.jpg>

Logistic regression coefficient for most frequent words, restaurant rating, review length, number of pictures in the review, and elapsed time since review first posted.