****

**PROBLEM STATEMENT AND RATIONALE STATEMENT**

**AIDI 1003 – CAPSTONE TERM 1**

**AMIT MARAJ**

AHNCH BALA 100424062

SONAKSHI KARKERA 100720763

SURBHI THAKUR 100732335

ARUN KALAESWARAN 100771700

**September 13, 2019**

**Problem Summary:**

In the year 2018, nearly 147 restaurants in the Toronto region were shut down with no reason given (Daily Hive Staff, 2018). There could be many reasons due to their closure, whether it be poor customer service, inadequate location choice or ultimately the lack of marketing. The reasons and assumptions can be infinite, but with so many resources available why was it so hard for many restaurant owners to figure out where they went wrong? Could they have known beforehand that they were going to close?

As a customer, it is easy for us to go online and search for reviews of specific restaurants to ensure that our experience would meet our expectations. We can even go further and look for specific niches like ambience, music or even if its pet friendly. We can find all this information with a simple search on sites like Open Table or Google Reviews. But perhaps the most informative site of them all would be Yelp.   
  
The satisfaction of customers is tied to the success of restaurants. According to Yelp, there has been an average of nearly 140 million users per month in Q2 of 2019 (Yelp, 2019). Furthermore, as of June 30th, 2019, there have been 192 million reviews contributed to the website (Yelp, 2019). This means that there is an enormous amount of data available for Yelp to use to determine if a restaurant will close before it happens. Yelp can use this information to open up new channels of revenue, specifically in the form of helping failing restaurants come back from the brinks of failure.

**Approach:**

Our goal in this project is to apply machine learning to predict a restaurant’s success based on the customer’s star rating and finding which restaurant features have the most impact on its average rating by performing sentiment analysis on the text of user reviews.

We’ll be using the dataset provided by Yelp as part of their Dataset Challenge 2019 to train and test our classification model. Dataset includes information about local businesses in 10 metropolitan areas across 2 countries.

Since this is a binary classification problem, we’ll be using methods such as SVM, Naive Bayes, decision trees and natural language processing to identify the most informative features from restaurant reviews. For our scope, we’ll just be using business.json and reviews.json files for our analysis.

Business.json contains businesses from selected cities, each having 15 attributes. In our case let’s say we are looking at restaurants in the Toronto region. We’ll look out for restaurants that were open for a particular year and check how many of them closed down after one-year period.Post this, we’ll merge it with reviews.json file based on their business ID. We will then look for words in each review and separate them by removing punctuation so that we have a “bag of words” for each review. Finally, we’ll stem, lemmatize and filter out the stop words in each bag of words using NLP package. Our strategy here will be to:

1. Have some domain specific, positive and negative words.

2. Count their occurrence

3. Co-occurrence in the review

We’ll train our model based on features selected and evaluate the model using 80/20 train and test dataset.

With a balanced dataset we are aiming for at least 65% accuracy. Also, since our aim is to classify whether a restaurant will close or not, precision for “closed” class is more important. Which means, of all the restaurants that our model predicted as closed turned out to be closed after a year.

**Why is this problem important to the organization?**

This problem is important to Yelp, as they want to retain as many businesses on their platform as possible. More businesses on the platform would generate more advertising revenues since it would create more existing web pages to advertise on. Having more businesses on Yelp would make them the industry leader in food recommendations, which would create more trust in their platform and make restaurants want to be on Yelp or have consumers want to search on Yelp.

**Problem Statements:**

**1.** When restaurants are already active on Yelp based on their Yelp profile we will be able to determine if they will close or not.

**2.** Using this model, Yelp can differentiate themselves from their competitors like Google reviews, OpenTable, etc. Yelp would be providing a service that can help their clients make business decisions instead of just being a platform that consumers would use for making food decisions.

**3.** Existing restaurants will be able to increase their customer experience using our recommendation. By determining whether they will close or not would give clients an opportunity to improve their user experience and try and prevent their business from closing.

**Constraints**

There will be some limitations we will face that could prevent us from being as accurate as possible. Below are the constraints that we face:

* **Validity of Reviews**: Although Yelp filters fake or bogus reviews, we are still unsure if 100% of the reviews are genuine.
* **Live Dataset:** The dataset is not live, which would mean that the current situation might be different for restaurant under review.
* **Missing Financial Information:** We don’t know the financial side of the restaurants under review which can help us be more accurate in our predictions.
* **Limited Access to Data:** We only have access to Yelp’s dataset and don’t have any other dataset we can use to strengthen our prediction.

Although we have some constraints, it does not make a big impact towards our predictions. We believe our predictions will still maintain a high percentage of accuracy. The constraints only prevent it from being better.

**Rationale**

Restaurant owners do not know about the closure of their restaurant early enough to prevent it from happening. We have proposed a solution that can warn business owners of closure early enough so that they can take preventative measures. As stated before, our goal out of this project is to look out for features that better predict the future of restaurant.

Our prototype will be able to predict this based on the customer’s star rating and finding which restaurant features have the most impact on restaurant’s average rating by performing sentiment analysis on user reviews. Our key metric for model evaluation will be precision and accuracy. Though accuracy will indicate how well our model is performing, we’ll be targeting on getting high precision value for “closed” class.

There will be 3 beneficiaries of our prototype. Yelp, our main beneficiary, will get to open a new revenue stream through consulting. Secondly, clients/restaurants get to receive expert advice from Yelp in order to prevent closure and financial loss. Finally, consumers will receive a more fine-tuned and catered food experience.