Robert Tobias

Olivia Viparina

Final Project Proposal

* 1. **What is your research problem?**

Our research problem is to determine how residents in New York City and Arlington, Virginia feel about Amazon building a facility in their cities. Many cities advocated to win the bid for Amazon to build their facility in their city, citing economic advantages and increased employment opportunities. However, there has been some speculation about whether the residents of these areas are actually supportive of their local government’s decision to allow Amazon to build in their cities. We aim to determine the sentiment of the residents in these cities in regard to the arrival of Amazon.

* 1. **A brief discussion of your data, e.g., data type, number of records, data quality, labels, etc., including necessary visualizations.**

We will collect tweets using RapidMiner. Our search query will be “Amazon, Amazon HQ, Amazon Headquarters” and we will search this query once in New York City using the coordinates (40.7447° N, 73.9485° W) with a radius of 50 miles and once in Arlington, Virginia using the coordinates (38.8554° N, 77.0521° W) with a radius of 50 miles. We will collect 1,000 tweets from each location. We can control the quality of our data by defining our search query, coordinates, and number of tweets. Beyond that, the quality of our data depends on what people are tweeting. We will visualize our collected data through word clouds to show an overview of the collected tweets and a density map to show where our tweets are coming from.

**1.3. What model(s) are you planning to use? Why?**

First we will use Naïve Bayes to classify the tweets based on the language they use. This will allow us to predict if people approve or disapprove of Amazon building a facility in their city. Naïve Bayes can be used for text classification and will allow us to classify our tweets into “approve”, “disapprove”, and “neutral” categories based on the words they use.

We will also use K-Means Clustering to explore trends in our collected tweets beyond the categories of “approve”, “disapprove”, and “neutral”. We will be able to create clusters within the “approve” and “disapprove” categories and create word clouds of these clusters that will show the most common words in each. This will give us insight into the reasons *why* people feel the way they do. This will be more of a descriptive machine learning technique compared to Naïve Bayes.

**1.4. Schedule of your project.**

By November 30th we will research the background of our question, identify our problem, and collect and clean our data. By December 2nd we will explain and visualize our cleaned data using tableau and RapidMiner. By December 4th we will have implemented both of our selected machine learning techniques and create visualizations for both. Throughout the course of research, data collection, and implementation of our techniques, we will update our data and codes to our GitHub page and this will be completed by December 4th. By December 6th we will explore limitations of our research and look for ways we could improve our model. We will write up, format, and submit our Final Report by December 7th. By December 10th we will create, practice, and submit a presentation of our findings.