**Name: Jonathan Lawrence**

**Date: 5/15/2020**

**Project 3: Disaster Tweets**

Any surprises from your domain from these data?

* There were no surprises as this dataset included all of the necessary variables that we needed. It included a labeled dataset which we used to train our machine learning algorithm, and an unlabeled dataset which we used to test our model. The files were pre-split from a singular dataset containing the same header configuration, meaning we didn’t have to modify either file to work with the other. The dataset was clean and simple so it wasn’t bogged down with useless data. The dataset also included a codebook which was immensely helpful in determining the scale of each datapoint.

The dataset is what you thought it was?

* This data included almost 22,000 tweets which were evenly divided between labeled and unlabeled data. The data contained two variables that we didn’t require (id, keyword, location) so we removed them which helped to speed up our cleaning process. There were a lot of tweets containing potentially biasing text, so we performed many cleaning steps. We encountered 2,594 values represented as “NA” and replaced them with blank values in order to prevent bias in our model and improve the accuracy. This included removing HTML, stop words, brackets, URLs, hashtags, special characters, and more.

Have you had to adjust your approach or research questions?

* Since the data contained lines of text rather than numbers, we had to come up with additional research questions in order to figure out how to learn what we didn’t know. Two of these research questions were “Does the quantity of characters in a text show any patterns? How about the quantity of words?” We played with a few different options and determined that a bar graph could best represent these trends. We also discovered that certain cleaning methods actually lower the model’s accuracy, leading us to question our initial approach to cleaning the data. As of now, we are running numerous A/B tests to determine which cleaning methods alter the context, or ease of understanding, for the tweets, and which preserve the context and understanding.

Is your method working?

* Yes and no. Our method is working to the point that it is able to predict a majority of cases accurately. However, the accuracy is still a bit lower than we would like so we’re researching additional ways to properly clean and format the text so that the model has an easier time trying to understand the data. While our plan is still the same, we had a few things come up that we’re still working through in our analysis. For starters, we had to consider what kind of linear regression model we wanted to build that would be best for this project. We knew that we needed a model which would be able to adapt to new words without becoming overly sensitive to them. We also knew that the model would have to work with a nearly unlimited number of tokens and still be able to make a proper prediction. We ended up choosing a Ridge Regression model because of its ability to regularize, or shrink, the coefficients so that the algorithm will produce low bias and low variance regardless of how many new tokens were seen.

What challenges are you having?

* Our initial challenge was trying to determine the best way to represent the data in a visual form that could be useful in identifying trends in the data which we mentioned earlier. Also, we had to consider what kind of linear regression model we wanted to build that would be best for this project. We knew that we needed a model which would be able to adapt to new words without becoming overly sensitive to them. We also knew that the model would have to work with a nearly unlimited number of tokens and still be able to make a proper prediction. It took some time but we eventually remembered about the Ridge Regression model. We chose it because of its ability to regularize, or shrink, the coefficients so that the algorithm will produce low bias and low variance regardless of how many new tokens were seen.