Cataline Perez,
Isaiah Herard,
Nash Morrison

Project 1: Initial Evaluations

Google Colab Notebook

Coronavirus:

For these set of datasets, the corpus involves COVID-19 related comments during the Pandemic in 2020'. They range from extremely radical negative comments to positive comments. The initial data file looks like a gigantic, jumbled mess with a separation of an "|\*d?*". A comment with a "|0" designates the comment is negative while a "|1" flags the comment as positive. These datasets could distinguish between content related to the actual pandemic and unrelated content which hold value for tasks such as information filtering or trend analysis to see what the consensus feel as the pandemic raged onward. When creating our model, index splicing is used on our training set to have 30,000 records against the 10,000 test records to satisfy our 2/3 and 1/3 ruling for predicting.

Coronavirus	***Train***	***Dev/Test***
Predicting?	**0: Represents Negative Comments 1: Represents Positive Comments**	
How many records are in each file?	**80k records total**	**10k records total**
Each class label has how many records?	**0 (Neg) --> 38,837 1 (Pos) --> 41,163**	**0 (Neg) --> 4,963 1 (Pos) --> 5,037**
Unique Vocab/Tokens in each file?	**Vocabulary 81,647 Tokens**	**Vocabulary 23,035 Tokens**

A large amount of preprocessing had to be done so that the models would be as accurate as possible. Positive and negative remarks are split into their own columns, then split by their respective unique tokens within each dataset.

Cataline Perez,
Isaiah Herard,
Nash Morrison

Triage Dataset:

The data presented in this table represents a triage process, which prioritizes tasks, patients, and the urgency of each matter. In our training and development/test datasets, the texts can be categorized into two groups: aid-related (class aid) and not aid-related (class not). Texts followed by '|0' are classified as not aid-related, while texts followed by '|1' are classified as aid-related. Afterward, we determined the number of records used for training and development/testing. The training set consists of a total of 21,046 records, with 8,685 classified as "aid" and 12,361 classified as "no aid." The development/testing set contains a total of 2,573 records, including 1,048 disasters classified as "aid" and 1,525 disasters classified as "no aid." To determine the unique terms, we applied the ⅔ and ⅓ rule, resulting in a range of 0–5000 for the training set and 0–1500 for the development/testing set. We manage to figure out the unique vocabulary/Tokens for the Train and development/test data sets in each file by using CountVectorizer code. This eventually gave us the result of 31,211 Vocab/Tokens for the Train set and gave us 10,145 Tokens for the Development/Test set. To make this data chart like the one in the code, we just separated the train and dev into their own rows and columns. Then adding their information at the bottom of each one.

Triage	*Train*	*Dev/Test*
Predicting?	**0: Disaster Without aid 1: Disaster with aid**	
How many records are in each file?	**21046 Records**	**2573 Records**
Each class label has how many records?	**Class-Aid: 8,685 Class-Not: 12,361**	**Class-Aid: 1,048 Class-Not: 1,525**
Unique Vocab/Tokens in each file?	**31,211 Tokens**	**10,145 Tokens**