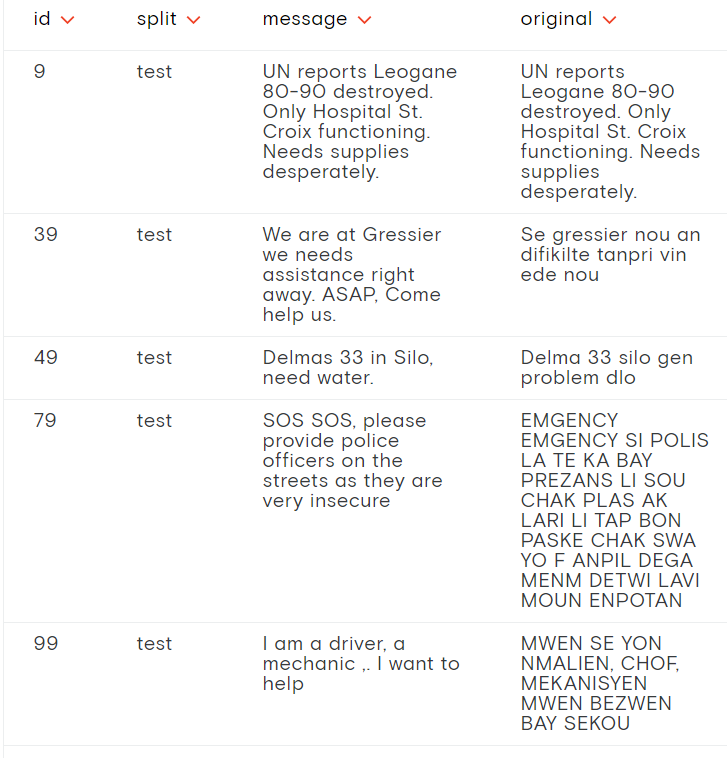
Overview:

Using Natural Language Processing with Naïve Bayes through Google Colab, we sorted messages from direct messages, social postings, and news stories into disaster categories. Trends and prediction results are visualized in tables using Tableau.

Data:

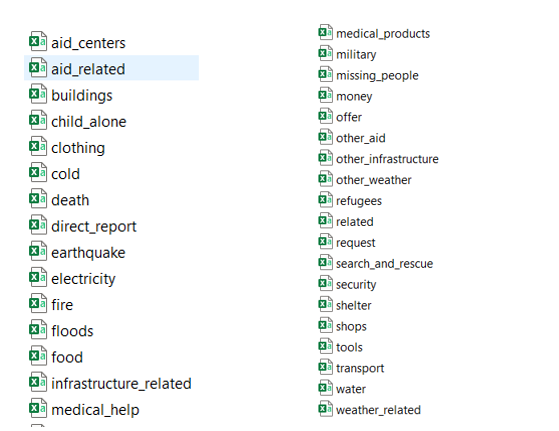
https://appen.com/datasets/combined-disaster-response-data/

This dataset contains 30,000 messages drawn from disaster events including earthquakes, floods, storms, from news articles, social media and direct messages. The messages span over many years and hundreds of different disasters. The data has been encoded with 36 different categories related to disaster response. Training data had already been identified and was somewhat cleaned.

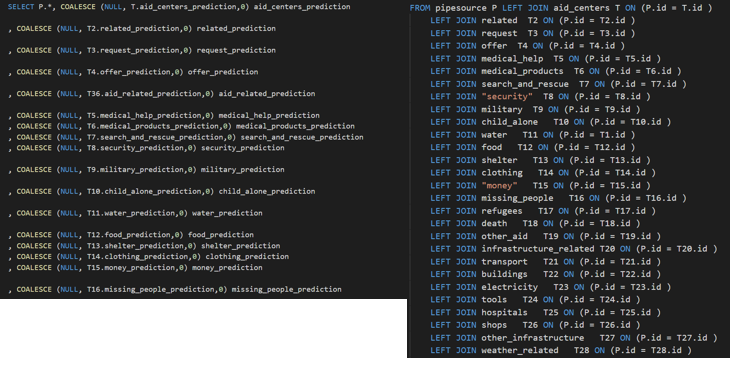


Methodology:

The machine learning algorithm used was Naive Bayes rather than using random forest because our NLP needed to be classified to multiple categories rather than just one. Google Colab was used to run notebooks with PySpark. One the machine was trained on one category it was replicated to be able to run 4 different training/prediction models at once. Each category was trained and tested for accuracy and output 36 CSV files that included the ID of the message, how it was categorized and what our machine predicted.



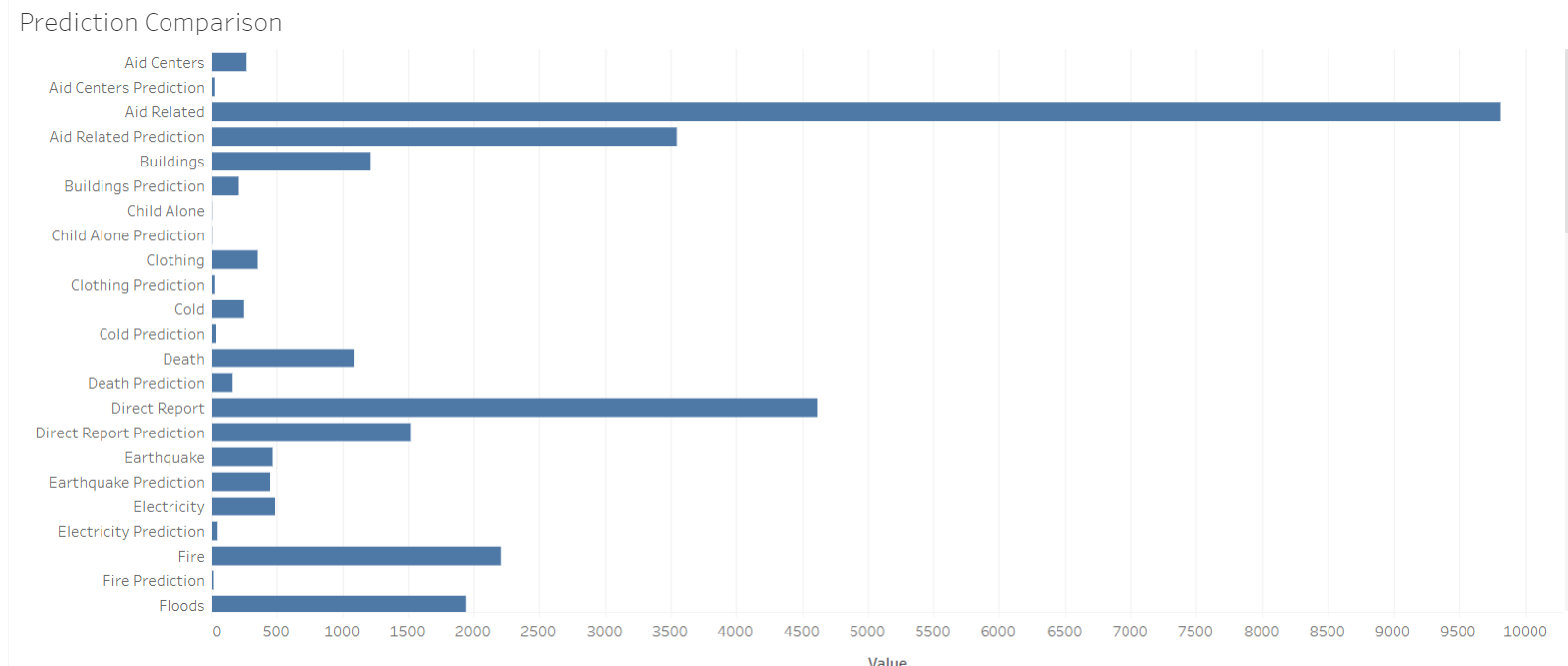
Database:

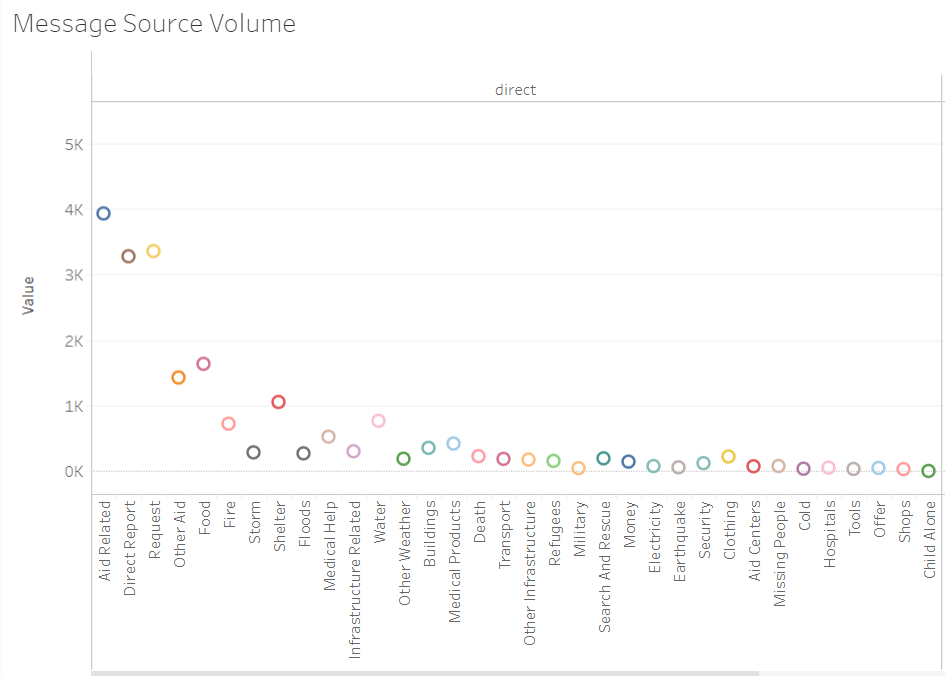
Once the CSV files were produced, they were loaded into Postgres along with a CSV file that contained the ID, message, and source of the message. A table was created for each CSV and then joined on ID. A final table was created and exported to a CSV file to be later used in Tableau to create visualizations. 

Graphs:

Graphs can be viewed:

<https://public.tableau.com/views/NLP_Final_Project/MessageSourceVolume?:embed=y&:embed_code_version=3&:loadOrderID=0&:display_count=y&:origin=viz_share_link>





Results:

When calculating the accuracy in Google Colab, the results we suspiciously high, between 90-99% accurate predictions from the machine on most categories. This was due to the 0 values being so numerous and calculated correctly.

Looking at the data through a table visualization, we discovered that the ratio between how many messages our machine predicted to be in the category versus how many messages were in the category varied wildly. Several variables impacted the accuracy of the machine, mainly how many messages were in each category for the machine to train on initially. With less categories to sort into or with more messages to sort, the accuracy of the machine can be improved.