DSCI521 Project Proposal

Team: Yushan Cai yc844@drexel.edu , Yunxuen Hu yh575@drexel.du , Max Bezahler meb25@drexel.edu

Background Introduction:

Yushan Cai: MS in Business Analytics, specialize in programming and visualization, will take charge of coding and visualization.

Yunxuan Hu: MS in Business Analytics, specialize in storytelling and analysis, will take charge of analysis and drafting.

Max Bezahler: MS in Data science, specialize in modeling and management, will take charge of coding and text wrapping.

From the Project folder in Drexelone

Project: One open-ended group assignment will have two phases:

1. Data Set Identification, Motivational Exploration, and Proposed Methods Implementation

2. Methods Implementation, Evaluation, and Interpretation, with Documentation and Dissemination

Phase 1: Data Set Identification, Motivational Exploration, and Proposed Methods Implementation

1. Data Set Identification

We propose to use the data that was used in a Kaggle competition called disaster tweets. The Kaggle disaster tweets description is available here. The data for the competition is available here under the title "Disasters on social media". Train, test and sample submission are provided on Kaggle here.

2. Motivational Exploration and Application

The data consists of social media tweets that are about a disaster and tweets that are not. For example "The sky was on fire" could be a poetic tweet about a sunset or a description of a conflagration. The dataset provides an indicator of whether the tweet referenced a real disaster or not. We are interested

in exploring this dataset because of the possibilities of using natural language processing (NLP) to determine and develop additional features which could be used as predictive characteristics. Derived features that we would explore would be sentiment analysis, keyword derivation, lemmanization, parts-of-speech, readability and reading level analysis, n-gram analysis and data visualization.

The application of this project is so extensive that can help people drug the information more precisely and correctly. For example, The Centers for Disease Control and Prevention (CDC) is closely monitoring an outbreak of respiratory illness caused by a novel (new) coronavirus first identified in Wuhan, Hubei Province, China. People will type on Sina and Twitters to express the ideas and issues. Using NLP and finding the relevant text, people can grasp the truth at first rather than wait for the news. On the other hand, due to the emergency of the coronavirus and the number of infections exploding , many hospitals lack the medical supplies. People can organize more efficiently and give the help in time.

3. Proposed Methods Implementation

As the data consists of tweets a fair amount of effort will be needed to clean and normalize the data. The computing environment will be Jupyter notebooks using the relevant python modules to create and test data. Currently we are proposing using github as the repository and then mybinder.org to create a virtual machine so that the jupyter notebook can be easily shared amongst the team.

Our method will be to analyze the tweet and determine whether the derived characteristic such as sentiment, part of speech, named entity recognition etc provides a predictive correlation as to whether the tweet was describing a real disaster or descriptive of something else. This will be validated against the provided test set. Currently we expect to use:

- Pandas for data cleaning
- Numpy for mathematical analysis
- Visualization will use seaborn, matplotlib and plot.ly
- Spacy and TextBlob for sentiment analysis, Part of Speech, and Named Entity Recognition, ngrams
- NLTK for stemming, lemmatization
- Textastic for reading level and comprehension analysis
- Wordcloud for wordcloud creation
- Scikit-learn for developing predictive models that best fits the model such as K-nearest neighbors and Support Vector Machine

Conclusion: We hope by detailed NLP analysis on the diaster tweets corpus to find a linguistic feature that provides a best predictor of whether a tweet is disaster relevant or not. We will use python and Jupyter Notebooks and available python modules to do this analysis.

4.Exploratory data analysis:

4.1 Dataset description

The original 'train' dataset contains 7613 observations and 5 columns while the 'test' dataset has 3263 observations and 4 columns. The 5 columns in the train dataset are "id", "keyword", "location", "text" and "target" separately. Below is the explanation of each column.

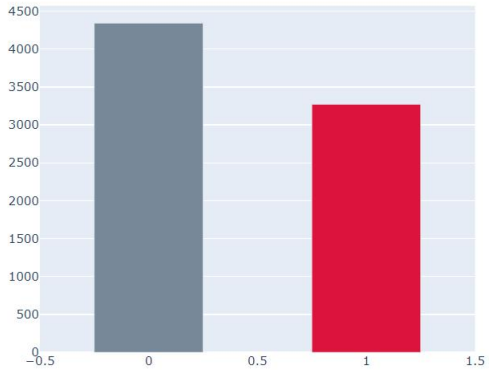
- id : a unique identifier for each tweet
- text : the content of the tweet
- location : the location the tweet was sent from (maybe not the real locations)
- keyword : a particular keyword from the tweet (the disaster tag)
- target : 1 means the tweets align with the disaster,otherwise, it is 0.

Besides, there are the missing value information showing below which we can see the "location" is taking the greatest portion of the missing value both in train(33.27%) and test(33.86%) dataset. Besides, the "keyword" only miss a little(around 0.8%) in the dataset. According to the missing value percentage and distribution, we found the train set and test set have the similar data distribution.

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        7613 non-null   int64
 1   keyword   7552 non-null   object
 2   location  5080 non-null   object
 3   text      7613 non-null   object
 4   target    7613 non-null   int64
```

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        3263 non-null   int64
 1   keyword   3237 non-null   object
 2   location  2158 non-null   object
 3   text      3263 non-null   object
```

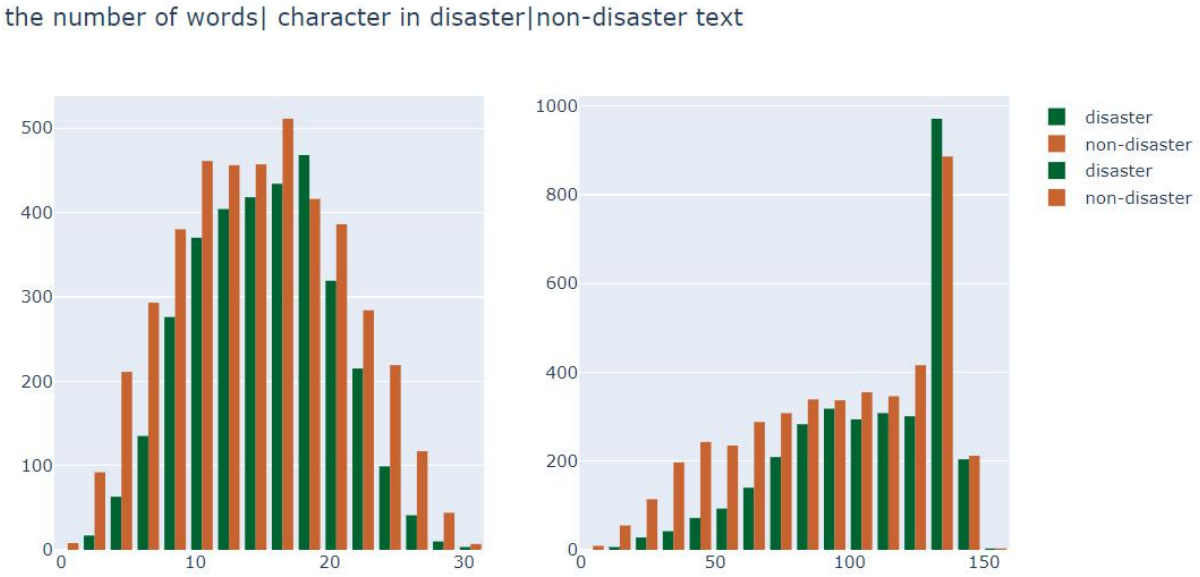
Then we want to know what the target distribution in the train set. According to the graphs, the number of observations = 4342 when target = 0 and the number of observations = 3271 when target =1. It implies the 'disaster' tweets are nearly equal to 'non-disaster' tweets which means we can skip the sampling method like oversampling to avoid a useless dataset.



4.2 Text initial exploration

We take a quick look at the "text" columns and find most of the either disaster tweets or non-disaster text having the character around 130-140. But there is slightly different distribution in other range

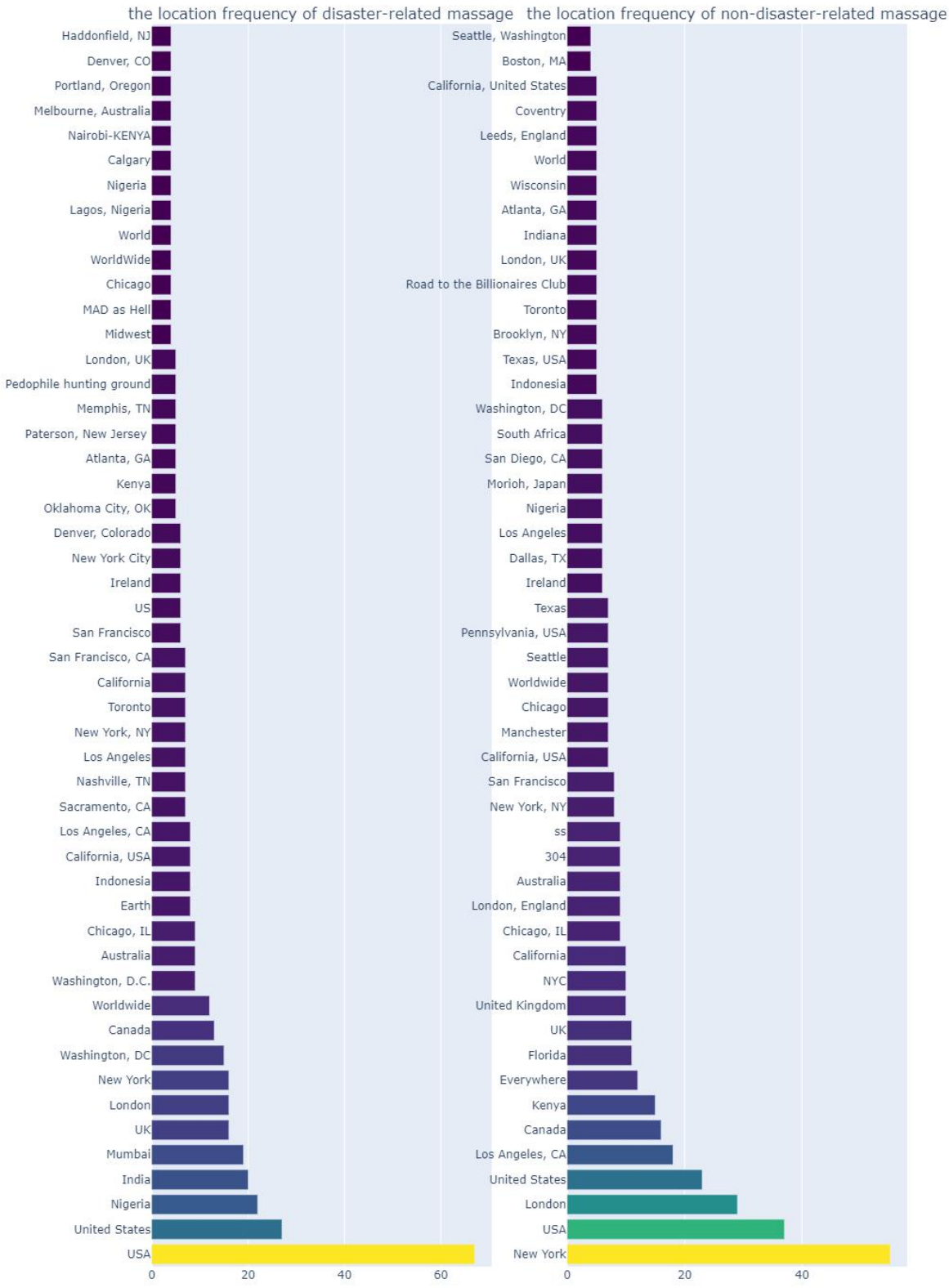
between those two types of text. In addition, when we look at the text in “word” level, we found the number of word in either disaster tweets or non-disaster tweets are between 10 to 20, in other words, there are not significant difference in word
level.



4.3 Location exploration

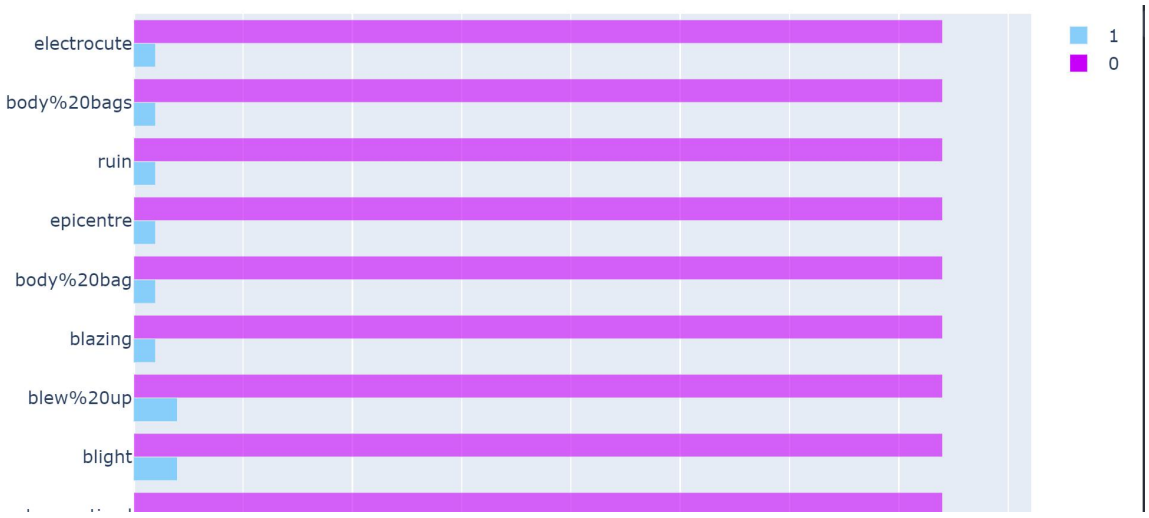
Here is the location appearance in the disaster and non-disaster text. From the 2 pictures showing below we could tell that tweets containing location which represents a real location instead of somewhere uncleared are more likely to be a disaster text. Since in the second graph, there are some uncleared location like “304”, “Road to the Billionaires Club”. Secondly, the disaster tweets are much more happens in the worldwide, because we could find the foreign cities like “Mumbai”, “Toronto” compared to non-disaster tweets often post on United-States and somewhere uncleared. In addition, there are 3341 unique values in training set and 1602 in testing set, which implies “real location” might be a useful feature to determine the target.

The Location of Disaster|Non-disaster Text



4.4 Keyword comparison

The chart below showing the information that which keyword can possibly imply the disaster information. We found that some of the keyword which we are commonly used in the spoken expression situation like “ruin”, “blazing”, “blight” “body bag” and “blow up” are showing less information to indicate the disaster.



However, some other keywords showing high information value to predict if it is related to disaster like “typhoon”, “derailment”, “outbreak” This might because in natural language, these words are high correlated to the disaster and they are not commonly used in the oral expression.