LEVERAGING SOCIAL MEDIA TO MAP DISASTERS

GENERAL ASSEMBLY – DATA SCIENCE IMMERSIVE

TEAM:
Dan Kim
Cameron Bronstein
Patrick O'Connell

1. Problem Statement

Implementing efficient and effective methods for disaster response is an important initiative particularly for geographical locations that may be more vulnerable to natural disasters or any sort of disaster in general. When responding to disasters, it is crucial to know which locations are in the most need of assistance in order to properly allocate resources. Currently, satellite and aerial imagery are a couple of tools used to locate and map areas that are affected by a disaster. There are other methods and tools, but the goal is to enhance these disaster response techniques to provide prompt and strategic assistance while being conscious of those areas that are in more dire need of help than others.

With the increasing popularity and integration of social media platforms into everyday life, in the event of a disaster, people turn to social media as a means to call for aid, stay updated and update others on the state of their current location via photos or text, and share other relevant information regarding the disaster or their situation. All of this data can be leveraged to build upon existing tools to map locations that are affected by disaster in real time. Using historical social media and other disaster related messages, our team set out to build a model that would identify people in need of immediate assistance versus people with less urgent needs and attempt to create a geospatial map to illustrate those areas that required help.

2. Data Collection & Methodology

For the purpose of this project we gathered tweets for Hurricane Michael and Hurricane Harvey to use as test sets for our model and a third-party text based dataset consisting of disaster related messages to train our model on. Twitter’s API was stringent which made data collection cumbersome because we needed to pull historical data, but Twitter’s Standard API limits the kind and quantity of data able to be requested. With the Standard package we were only able to pull data from the last seven days that the request was made and only 100 tweets at a time which made pulling large quantities of data a tedious process. To leverage Twitter’s database as best as we could, we managed to retrieve some data related to Hurricane Michael. One of the limitations that we faced with this data was that these tweets were not tweets that occurred during the hurricane so the tweets did not reflect individuals in need of assistance as much as we would have preferred.

Since requesting from Twitter's API was limited, we sourced third-party data related to disaster messages. Kaggle had a sizeable dataset for Hurricane Harvey readily available and Figure Eight, an online machine learning platform, provided us with an expansive dataset containing disaster response messages from multiple natural disasters such as earthquakes in Haiti and Chile, Hurricane Sandy in the United States and floods in Pakistan. This dataset contained text data that most appropriately reflected individuals in need of assistance and thus we trained our model on this data.

The Figure Eight dataset categorized its text data into 3 distinct categories:

- Direct calls for help
- Personal social media posts
- News related social media posts

While these categories were helpful, the data was not labeled to represent which observations were urgent needs versus not urgent. To go through each one of these messages and classify whether it was an urgent message would be very time consuming so for the purpose of our classification model we labeled all direct calls for help as urgent needs and all other observations as non-urgent.

3. Model Selection & Findings

Ultimately, we created a Logistic Regression model to predict whether a given message reflected a person in urgent need of assistance. The Logistic Regression model allowed for the most interpretability of results which we felt was most important because what we interpret from this model will allow us to better inform modeling decisions or "search words" in the future. Before modeling, we processed our observations by Count Vectorizing our text data using these hyperparameters: (n_gram = (1, 2), max_features = 50,000, stop_words = None). Through NLP we were able to sift out text data that had mentioned the disaster in part, but did not explicitly say that an individual or a group required assistance. We ran a GridSearch to test the hyperparameters of CountVectorizer, TFI-DF Vectorizer and the Logistic Regression model and found that the model performed best under the CountVectorizer with a test accuracy score of 0.9735.

4. Conclusions & Recommended Next Steps

Our model resulted in a strong test accuracy score for our training data (the data we built our model on). For future iterations on the project we want to focus on optimizing for sensitivity, as we want to reduce for false negatives.

It is important to define the limitations that we faced throughout our process in order to tailor our recommendations for future use. First, as mentioned earlier, it was difficult to access Twitter's database for historical tweets given that the scope of our project didn't allow for a paid Enterprise Twitter package. But even with the data that we did manage to retrieve, the tweets did not come with geolocation tags which made it difficult to map the general affected area if a given tweet reflected an individual or group in need of assistance. Also, without geolocation information specific to the origination of the message, our models would be rendered useless because we would not be able to determine where the message came from even though we classified that an individual or group required assistance. We recommend that disaster response or organizations partner with social media platforms like Twitter and others to access their geolocation data if possible. An issue with user privacy may come into question, but this can be addressed by providing the user with an opt-in choice in their social media settings (i.e. "Do you agree to give access to your location in the event of natural disasters with emergency response organizations?").