Proposal for Twitter Data Analysis Project

Linsey Sara Passarella, Emily Joyce Herron, Maria Mahbub and Gerald Leon Jones

***Abstract*— We propose a project in which data from tweets will be used to compare conditions in different areas (e.g. cities, counties, neighborhoods) within an affected region before, during, and after a particular natural disaster. The goal will be to investigate what kinds of information Tweets can reveal about things such as the preparedness of or impact of the disaster on certain areas compared to others. To accomplish this, we will gather Tweets related to the disaster and compare the subjects, topics, attitudes, and sentiments present in these Tweets changes during the disaster or differs between communities in different stages of the disaster.**

I. MOTIVATION

Hurricanes pose both an economic and social threat to communities. By assessing tweets concerning the impact of a hurricane across multiple cities, we can better understand the differences in attitudes towards preparedness, disaster clean-up, and damage costs in relation to the socioeconomic status of an area. An article published by Masozera et al [1] in 2006 showed that low-income areas in New Orleans were more vulnerable during response and recovery efforts after Hurricane Katrina. While flood damages occurred regardless of area or neighborhood, a number of limiting factors including transportation issues, lack of quality insurance (if any) and a higher chance of poorly constructed houses and buildings lead to a greater relative impact for low-income areas [1].

II. PROPOSED APPROACH

A. Data Collection

For this step, an API such as BeautifulSoup or an open-source tool such as DisasterMasters TweetScraper or https://github.com/taspinar/twitterscraper (format later) will be used to directly download disaster-related tweets along with their metadata. The scope of the collection will be limited tweets that both contain keywords or hashtags related to the disaster and were posted within a time frame beginning two weeks before the disaster and ending one month after the disaster. We expect that each data item downloaded will include attributes such as the tweet's text, time stamp, and author. For example, if we look at tweets surrounding hurricane Florence, we will look for the keywords "florence" and "hurricane". The time frame for collecting tweets is going to be dated from August 15th to September 26th. We will gather economic data and information for a given zip code using the United States Census found at their website https://census.gov.

B. Data Processing

To prepare for analyzing the tweets, we will remove hash tags '#', stopwords, punctuation marks and user-mentions which is usually started with a symbol '@' followed by a username. We will also focus removing the embedded URLs which might be in the texts. After that, we will work on replacing contractions with words; such as can't will be replaced by can not, didn't will be replaced by did not and so on. We will use elongation replacer to replace the elongated words with proper English words. For example, sometimes tweets will include words like, 'goooooddd' or 'noooooo' to express strong emotions towards something. We will keep an eye for those words and replace them with appropriate meaningful ones. We will also lemmatize words if necessary.

C. Exploratory Analysis

After collecting and cleaning all relevant tweets, we will evaluate their content for the purpose of intelligently targeting relief effort and identifying infrastructure improvement areas. Using a list of cities, counties, and states in the affected area, tweets will be tagged and grouped by location. We also plan to identify the most frequently tweeted words, particularly words describing topics such as resources and recovery, at different time intervals before, during, and in the aftermath of the disaster and compare how these frequencies differ by location. Similarly, the tweets will be categorized by sentiment using a pre-trained sentiment classifier and compared by location and changes over time.

D. Communication & Visualization of Results

Following the exploratory data analysis process, we plan to summarize our findings by presenting our data in a series of visualizations. These will include graphs

of the most frequently identified terms and sentiments and how they differ over time and by location. In addition, we plan to create maps labeled or color-coded by sentiment or most frequently used words using an API such as geopandas or basemap.

E. Expected Results

By analyzing the tweets we hope to answer the following questions :

- Whether useful information can be gained about the varying effects of natural disaster on different areas
- How much destruction and damages have been caused by this disaster
- In case of getting relief or aide, how does it depend on area or locality
- Is there any certain types of tweets that lead to an increase in the speed or amount of help to an area
- What is the sentiment of people regarding this particular disaster

III. General Proposed Timeline

- **September 28th:** Proposal due
- **October 1st:** Begin data scraping with an API. Cleaning and processing will also take place as we gather twitter data.
- **October 29th:** Finish processing data and begin working on evaluation and analysis.
- **November 12th:** Begin data visualization step and prepare for presentation late November.
- **Late November:** Presentation and research paper is finalized.

IV. Member Responsibilities

Members will plan to meet at least once a week to discuss and work on the project. Our work will be divided evenly between all members at the beginning of the week and will be dependent on what step of the project we are on and the different skill sets needed to complete the step. By the end of the week, we will give an update on what was accomplished and review the work of all team members. Issues can be raised in the project repository if a team member is stuck or needs additional help/advice while working their tasks for the week.

References

[1] C. K. Michel Masozera, Melissa Bailey, "Distribution of impacts of natural disasters across income groups: A case study of new orleans," in *ECOLOGICAL ECONOMICS 63*, pp. 299–306, 2007.