Analysis of Natural Disaster's Impact on Public Sentiment Towards Climate Change

Akash Ilangovan, Dharmpreet Atwal, Vijay Sambamurthy, Issam Arabi

Monday, Dec 14, 2020

Introduction: Problem and Research Description

Over the last couple of decades, awareness about climate change has increased significantly. In a survey done this year, 60% of Americans believe that global warming is a major threat to their country, whereas only 44% did in 2009[1]. Disasters from this year like the California fires (which consumed over 6,500 square miles and damaged 10,400 structures[2]) and the Australian bush fires (which burned 110,000 sq km of land[3]) have garnered much attention from mainstream media, particularly social media.

With the invention of social media, it has become increasingly easier to connect ourselves to catastrophes occurring around the world. The change in public attitudes and sentiment towards climate change seems to be largely driven by this increase in connectivity. Could exposure to extreme weather events through social media affect public sentiment towards climate change? This idea fueled most of the questions behind our research. The main question we sought to answer was: **How has climate change sentiment expressed on social media changed over time? Are natural disasters the cause of any such changes?**

This question is important for various reasons. An analysis of sentiment expressed in social media allows us to accurately predict sentiment expressed in real life. Such information would allow us to gain insight into how much (if at all) do these disasters affect public opinion. For instance, an increase in the negativity of public opinion concerning climate change during or after a natural disaster indicates that people are not happy with the status quo, and are more likely to push for green policies, vote for politicians that call for such plans, or simply just incorporate eco-friendly habits into their lifestyle. The inter-connectivity of social media makes this analysis all the more important. This is because negative (and positive) emotions spread on social media rapidly, so influencers and accounts with many followers have the ability to shape public opinion like never before, and it would be interesting to see whether the domain of influence of such accounts includes public opinion concerning climate change.

To investigate this topic, we plan on surveying one of the world's most used social media platforms: Twitter. We will select and analyze tweets from key time frames where a natural disaster had just occurred (or was just about to occur) and compare the average sentimental value of these tweets to the average from those from periods of regular climate change discourse. In the days before a disaster, people are usually aware of the impending situation, and in the days after, people are taking to social media to express their opinions. In our analysis, we set that any climate change tweets within 3 days of a natural disaster were disaster tweets, and the rest weren't. We hypothesize that the average sentimental value of climate change tweets in periods of natural disasters would be more negative than that of other climate change tweets.

Dataset Description

Climate Change Tweets Ids: a dataset provided by Harvard Dataverse with tweet ids from 39,622,026 tweets related to climate change between September 21, 2017 and May 17, 2019 from the Twitter API using Social Feed Manager. There's a gap in the collected data from January 7, 2019 - April 17, 2019. The tweets were collected using Twitter's API and searched for using keywords like, "#climatechange, #climatechangeisreal, #actonclimate, #globalwarming, #climatechangehoax, #climatedeniers, #climatechangeisfalse, #globalwarminghoax, #climatechangenotreal, climate change, global warming, climate hoax"[8] The tweet ids used depend on the input of the user of the program. The dataset only contains 1 column, 'tweetids'. We used the 'tweetids' column and hydrated each tweet the id represented in the dataset.

Disaster Declarations Summaries: a dataset that contains all the reported natural disasters in the US since 1953. It contains other information, but we only used the 'incidentBeginDate' column. The full dataset spans many decades but we only used the years for which we had tweet data.[9]

Computational Overview

We first take our dataset of 40 million tweet IDs and shorten it a manageable size. We have two options for this processing[1]:

1. Chose a random number of IDs from the original 40 million
   (a) User chooses size of list
   (b) For example, if the original dataset was 100 elements and the user wanted a shortened dataset of 20 elements, we return 20 random elements.
2. Choose every n^{th} Id from the original 40 million:
   (a) the n value depends on the size of list the user wants
   (b) For example, if the original dataset was 100 elements and the user wanted a shortened dataset of 20 elements, we return every 5^{th} element.

Next, we convert all the tweet Ids in the shortened dataset to actual tweet content (text, date published, etc.). This is done using the hydrate function from the TWARC module[7] using private API keys that let us access the tweets from the ID values[2]. The tweet content is then saved to a CSV to use for later. Hydrating is necessary because Twitter's terms of service does not allow tweet content to be publicly available in collections like these.

We then covert the content of the CSV file into a dictionary for cleaner use later on. The reason for creating a CSV file and then creating a dictionary later anyway is so that the tweet hydration part only needs to be done once in case the program needs to be run multiple times (the tweet hydration part takes quite a bit of time). The hydrated tweets are stored in a csv file that can be accessed at any point without having to do any more processing.

Next, we pass the dictionary into a function that groups the tweets by day and finds the average sentiment (positive or negative on a scale of -1 to 1) for each day. And we finally store these values in another csv file. For later use in plotting.

Then, we take natural disaster datasets and average sentiment value per day dataset to find tweets that are within three days (before or after) a natural disaster. This is done by comparing the date values that are present in both datasets to ensure they are no more than 3 days before or 3 days after the date of the disaster. All of these tweets are stored separately from the non-disaster tweets.

[1]The dataset we created to send was only 20,000 evenly distributed tweets.

[2]We also have a secondary method of hydrating tweets that is a little bit faster but requires the command line to be executed. The command line hydrate function runs faster but requires JSONL files as its input which is why we have various conversion functions in our code to accommodate for that. But they are all unnecessary if you choose to run it purely on PyCharm using only CSV files.

Finally, we plot the sentiments of each tweet type, with the disaster tweet points being a different color from the non-disaster tweets to differentiate. For a numerical output, we calculate the average sentiment of each type of disaster tweets and non-disaster tweets separately to see if there is any meaningful difference in sentiment as a result of natural disasters. The data is visualized as a scatter plot of sentiment over time with the disaster tweets being in red while the non-disaster tweets being in blue.

Instructions for obtaining data sets and running program

Please make sure all imports in the *requirements.txt* file have been installed.

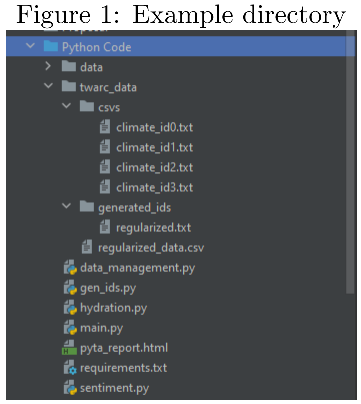
Dataset link : https://doi.org/10.7910/DVN/5QCCUU

Before proceeding, please download the 4 files in the link above named *climate_id.txt.00,climate_id.txt.01*, *climate_id.txt.02,climate_id.txt.03*. This will take a while. (While we could have made a program for this, we find that chrome/any browser downloading these files tends to be faster than the request library)

Download the zip file containing all the other modules and folders. Make sure that *main* is copied into this folder such that *main.py,data,twarc_data* and all the other *.py* files are at the same level.

Move the 4 downloaded files from above into *twarc_data/csvs/* and rename all of them such that they all end with *.txt*. Figure 1 shows what the directory should look like as seen in pycharm.

Figure 1: Example directory



Note: *generated_ids* folder contains the ids of tweets. Data also has the raw csv tweet data, this has been provided in the zip file in case the person running is unable to obtain an API Key for twitter.

Then run the *main.py*. Read the main statement in the program for more information on what to run.

Changes between proposal and final submission

In our proposal, we planned to measure the ideological differences between two of the US's major political parties, the Republicans and the Democrats. By completing a sentiment analysis of the tweets from prominent figures within in each party, we hoped to quantify the divide in how each party talked about climate change, in either a positive or negative light Unfortunately, we could not find enough tweets from either party. Therefore, we decided to change the direction of our project and instead focus on comparing the average sentiment value of climate change tweets posted after a natural disaster to climate change tweets posted at a time when no natural disasters had occurred recently. This shift in focus from a specific group of people, to the general public made it much easier to find tweets to analyse.

Discussion

At the beginning of the project, our group predicted that the average sentiment of tweets pertaining to climate change would be more negative after natural disasters. Intuitively, we thought that natural disasters cause people to feel more pessimistic, leading them to increase their use of words with gloomy connotations in their tweets about climate change. This increase would result in a more negative score being attributed to these tweets by our sentimental analysis library and a corresponding change on our graph of average sentiment value of tweets over time. We were surprised, however, to see that the average sentiment value of tweets related to climate change was relatively constant at around +0.05. We have a few theories for this result.

First, even though our dataset gave us access to around 40 million tweet ids, we only have so much space on the computer to store the tweets - not to mention the time processing these tweets takes. Consequently, we only restricted ourselves to 20,000 tweets (a mere 0.05% of the tweets available). This decreases our chances of finding a significant difference between the tweets tweeted after a natural disaster and other climate change tweets. In addition, the dataset only presented us with tweets that were posted between September 21, 2017 and May 17, 2019. This short time span that is less than two years means that there were not enough major natural disasters for us to perceive a noticeable difference between disaster and non-disaster tweets. This is because most natural disasters don't make it to mainstream media in the first place. This too diminishes the likelihood that we find any trend in the data.

Furthermore, we only know the average sentiment value of tweets of a certain day, and the spread of the values was not examined. Therefore, it is possible that the number of more negative tweets did increase, but a similar number of positive tweets resulted in us getting the same value since we only measured the average sentiment of the tweets of a certain day. These positive tweets could be tweets mentioning news such as an increase in green policies being implemented, people or organisms being saved, or a surprisingly good outcome from the disaster (all of which are likely to be mentioned in optimism).

Moreover, the dataset of natural disaster dates that we used only recorded dates of natural disasters that occurred in the United States, which only represent a fraction of natural disasters that occured around the world, such as the 2019 Amazon rainforest wildfires. This makes it harder to notice a difference between sentiment expressed in the two types of climate change tweets, since what we assumed to be non-disaster tweets include disaster tweets that were not detected by our algorithim since these disasters did not occur in the United States.

Our final - and most probable - theory is that our hypothesis was wrong. So far, the only conclusion we can make is that people don't use more pessimistic words when tweeting about climate change. Although this is not what we expected, we have a few explanations for this too. Perhaps what happens after a natural disaster is not that the number of negative words increases in people's climate change tweets after a natural disaster, but that the number of such tweets posted increases, albeit with the same level of optimism. Also, maybe some of the emotions exuded by climate change tweets did change after a natural disaster, just that positivity was not one of these emotions. It would be interesting to see if a similar study could measure any change in the use of words pertaining to emotions such as fear or surprise, for instance.

While our research did not support our initial hypothesis, there is potential for more meaningful findings in analysis conducted on a larger set of data and in an algorithm that measures the display of another emotion.

References

[1]Funk, C., Kennedy, B. (2020, July 27). How Americans see climate change and the environment in 7 charts. Retrieved December 11, 2020, from https://www.pewresearch.org/fact-tank/2020/04/21/how-americans-see-climate-change-and-the-environment-in-7-charts/

[2]Brackett, R. (2020, December 3). Two Firefighters Injured In California Wildfire Fueled By High Winds: The Weather Channel - Articles from The Weather Channel. Retrieved December 11, 2020, from https://weather.com/news/news/2020-12-03-southern-california-wildfire-orange-county-bond-fire-evacuations

[3]Ghosh, P. (2020, March 04). Climate change boosted Australia bushfire risk by at least 30%. Retrieved December 11, 2020, from https://www.bbc.com/news/science-environment-51742646

[4]Monkey Learn. (2020). Everything There Is to Know about Sentiment Analysis. Retrieved November 04, 2020, from https://monkeylearn.com/sentiment-analysis/

[5]Monkey Learn. (2020). A Comprehensive Guide to Aspect-based Sentiment Analysis. Retrieved November 04, 2020, from https://monkeylearn.com/blog/aspect-based-sentiment-analysis/

[6]Rolczyński, R. (2020, October 28). ScalaConsultants/Aspect-Based-Sentiment-Analysis. Retrieved November 04, 2020, from https://github.com/ScalaConsultants/Aspect-Based-Sentiment-Analysis

[7]Summers, E. (2020, October 22). DocNow/twarc. Retrieved November 04, 2020, from https://github.com/DocNow/twarc

[8]Littman, J., Wrubel, L. (2019, May 20). Climate Change Tweets Ids. Retrieved November 04, 2020, from https://doi.org/10.7910/DVN/5QCCUU

[9]F.E.M.A. (Dec 13, 2020). OpenFEMA Dataset: Disaster Declarations Summaries - v2. Retrieved Dec 13, 2020, from https://www.fema.gov/openfema-data-page/disaster-declarations-summaries-v2