Coffee Break Experiment #2

Charles Hutchinson, Sam Kalman

Due 2022-11-20

Introduction

One of North America's most noticeable symptoms of climate change has been a dramatic increase in the volume and intensity of wildfires, particularly in its western regions. August and September were previously defined by clear skies and fair weather, but fire and smoke have become a recurrent factor in this area. According to the National Interagency Fire Center (NIFC), more acreage burned nationally in each of the last seven years than any single year but one in the first ten years this data was collected (1983-1993, ex. 1988).[1] We sought to analyze if and how public sentiment has shifted over time with respect to these cataclysmic events, with the hopes of gaining some insight into public perceptions of climate change.

Data

We analyzed sentiment towards wildfires by pulling publicly available tweets through Twitter's API. We chose to pull tweets made around the time of four of the largest, most recent wildfires in California. These events were the 2021 Dixie Fire, the 2020 Bay Area Fires, the 2018 Camp Fire, and the 2017 Tubbs Fire.

It was difficult to find relevant data. Tweets that were well-tagged (e.g., including #DixieFire for the Dixie Fire) tended to be associated with news organizations, which were overwhelmingly neutral and rarely "organic" reactions. However, loosening the search parameters inevitably lowered the data quality (e.g., searching for "smoke" mostly returned tweets referencing marijuana usage). It was especially challenging to develop searches with consistent methodology across separate wildfires, while still getting useful data.

We found our best results using the following approach:

- Use the commonly accepted fire name and "fire" as search terms. For example, the search terms for the 2021 Dixie Fire would simply be "dixie fire".
- Limit the search results to English. Creating sentiment dictionaries in other languages was out of scope for this project.
- Remove retweets.
- Limit the tweets to a month time frame. Some fires lasted longer than this, so we picked the interval in which the smoke was most impactful.

The size of each fire's corpus was unfortunately still far from uniform, but we found this to be a reasonable compromise among a set of unappealing options.

Table 1: Relevant Tweets per Recent Wildfire.

Fire	Tweets
Tubbs Fire	3530
Camp Fire	50469
Bay Area Fires	4680
Dixie Fire	26139

[1] See https://www.nifc.gov/fire-information/statistics/wildfires.

Methods

We used a sentiment analysis algorithm from the syuzhet[2] package to analyze our Twitter data. The syuzhet package returns sentiment scores on a spectrum of positively/negativity using a sentiment dictionary developed in the Nebraska Literacy Labs. We removed all links, web addresses, and twitter handles from the tweets. Sentiment scores above 0, below 0, and equal to 0 were respectively labeled as positive, negative, and neutral. Any tweets from Twitter accounts with the words "news" or "meteorology" in their biographies were filtered out. We found that most tweets from news organization contained neutral, non-opinionated information, which wasn't useful for our research question. In an attempt to filter out bots adding unnecessary noise to the data, we only used tweets from users with at least one follower, and the tweet itself needed at least one external interaction (a like/retweet/comment).

Results

In Figure 1, the sentiment classification breakdown is plotted for each of the four wildfires. To account for the tweet volume difference between the four fires, percentages of sentiment type are calculated based on the total number of tweets pertaining to each fire. The fires are in chronological order, beginning with Tubbs Fire and ending with the Dixie Fire. The error bars reflect a 95% confidence interval for each proportion.

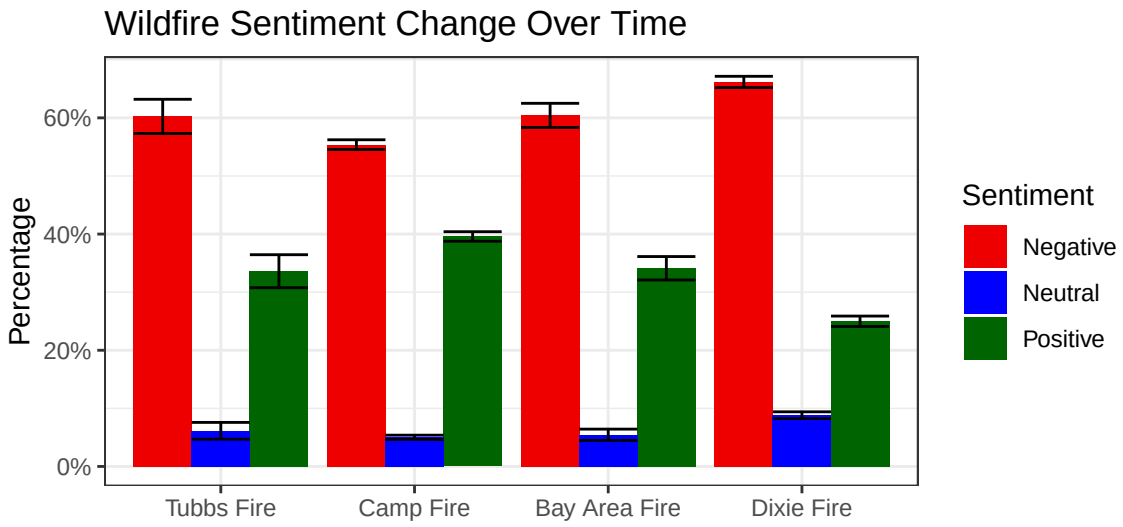

Figure 1: Shift in Wildfire Sentiment Over Time.

Discussion

Even though the tweet volume was not consistent throughout time, Figure 1 gives us preliminary evidence that there may have been a shift in tweet sentiment towards wildfires in the last five years. Beginning with Camp Fire, the amount of negative tweets significantly increased for each wildfire.

As mentioned previously, data quality was the limiting factor for this analysis. It was difficult to find a set of relevant tweets using a consistent methodology for each fire. Finding a balance between organic and relevant reactions ended up requiring more judgment than we would have liked.

It was surprising that we could not find more tweets for the Bay Area and Tubbs fires, given their proximity to major urban centers. We originally suspected that the Camp Fire tweet set contained many false positives, but it appears that fire generated more engagement than expected due to the almost total devastation (and subsequent press coverage) of the towns of Paradise, Concow, Magalia, and Butte Creek Canyon. The Dixie Fire was similarly destructive. As the largest ever single wildfire in the state's history, part of its impact included popular Lassen National Park, the coverage of which likely increased engagement on Twitter.

[2]See https://www.rdocumentation.org/packages/syuzhet/versions/1.0.6

Our next area of research will be studying location-based elements of the wildfire tweets. Data quality issues were compounded in our initial attempts to collect this data, since only a subset of Twitter users enable this functionality. However, knowing the location of each tweet will allow us to better understand to which extent these changes in sentiment were localized.