Tweet Analysis:
2018 Camp Fire

Justin Fischer & Matt Burke

May 15, 2020

GA-DSI-DEN-Project5

Agenda

- Context
- Problem statement
- Data collection and feature engineering
- Modeling and analysis
- Challenges and constraints
- Recommendations and next steps

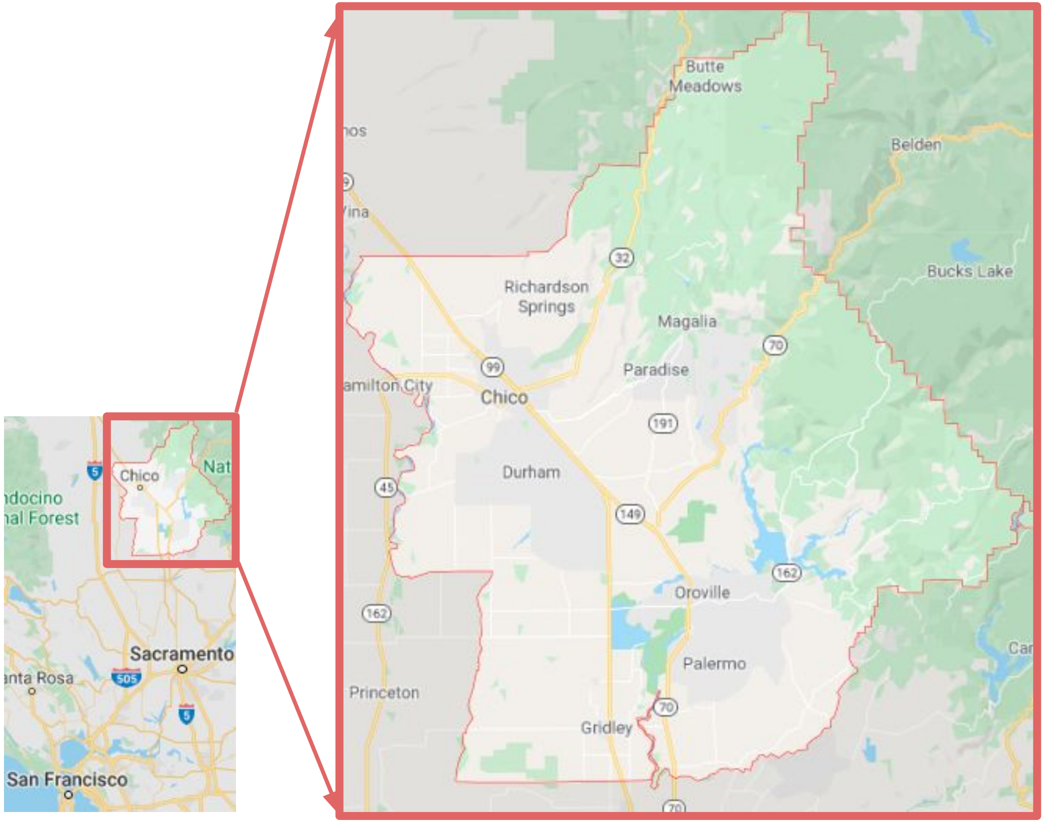
Context

Background on project

- Work should focus on:
  - preparing for emergencies
  - rapidly responding to emergencies, and/or
  - estimating the economic impact of disasters
- Our topic selection criteria:
  - Pulling data from social media
  - Relatively recent disaster
  - End result something that can be generalized to other disasters
  - Unsupervised learning
  - Sentiment analysis

Case Study

- Camp Fire
  - Paradise, CA (Butte County)
  - November 8-25, 2018
  - 85 fatalities
  - 17 non-fatal injuries
  - 52,000 evacuated
  - 153,336 acres
  - $16.5 billion worth of damages
- Reasons for selection
  - Timing
  - In US
  - Media coverage
  - Footprint of impact
  - “Slow” disaster



source

Problem Statement

Refining our problem statement

Problem statement v1

Can we use location-based social media data and a mapping API to map the 2018 California wildfire?

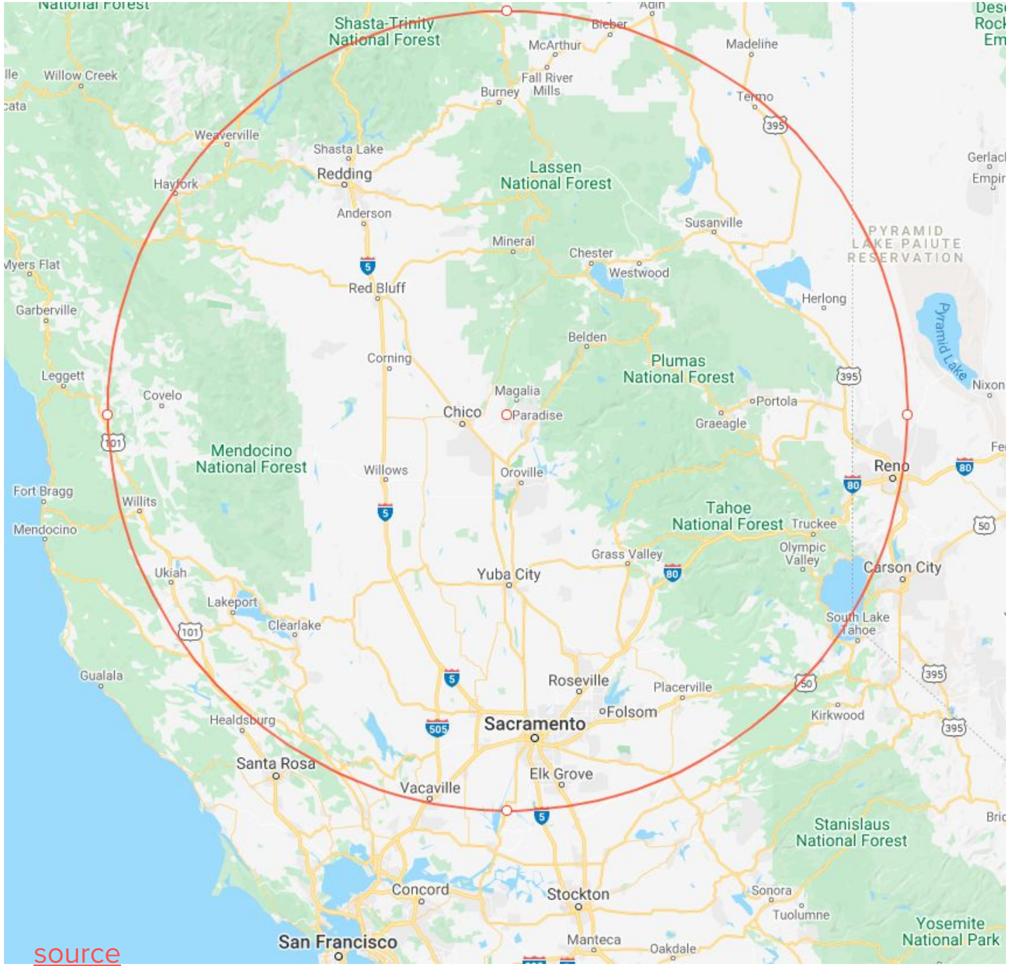
Problem statement v2

Can we build a list of keywords to help detect that an event is happening from social media posts?

Data Collection and Feature Engineering

Data Collection

- GetOldTweets3 (API)
- Cities of interest
  - Butte County
  - Paradise
  - Chico
  - Magalia
  - Oroville
- Date range: 11/1 - 11/26
- Pull all tweets within 100 mile radius
  - Remove duplicates



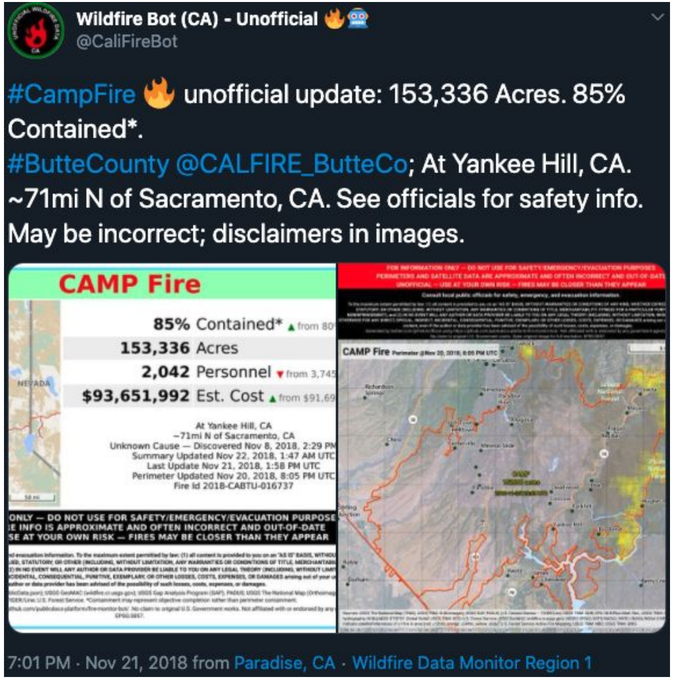
source

Feature Engineering

- key_score
  - **0 -** *656*; **1 -** *353*; **2 -** *84*; **3 -** *19*; **4 -** *11*; **5 -** *7*
- from_locations
  - 32 unique locations
  - 95% from Butte County
  - 80% from Chico and Paradise alone
- is-fire-related
  - 46/54 split between is not/is
- during_fire
  - 20/80 split between is not/is
- sentiment
  - 65/35 split between positive/negative

- keywords
  - fire
  - evac
  - smok
  - burn
  - wild
  - blaz
  - hell
  - department
  - inferno
  - help
  - alone

Sample tweets



key_score: 1
from_location: Paradise, CA
is-fire-related: 1
during_fire: 0
sentiment: 1
source

key_score: 5
from_location: Paradise, CA
is-fire-related: 1
during_fire: 1
sentiment: 1
source

Modeling & Analysis

Overview

- k-Means clustering
    - *What types of things people are tweeting about during the fire?*
    - Fire-related tweets only (612)
    - Tested to find optimal amount of clusters
    - Balance interpretability with granularity of clusters
- Sentiment analysis
    - *How does tweet sentiment change over the course of the fire?*
    - Library: Twitter NLP Toolkit
    - Module: Tweet Sentiment Classifier
- Key score analysis
    - *How is the key score of a tweet affected by factors such as sentiment, location, and time horizon of the fire?*

k-Means Clustering

3 clusters

- 1: Traffic and road related
- 2: Emotional
- 3: California

5 clusters

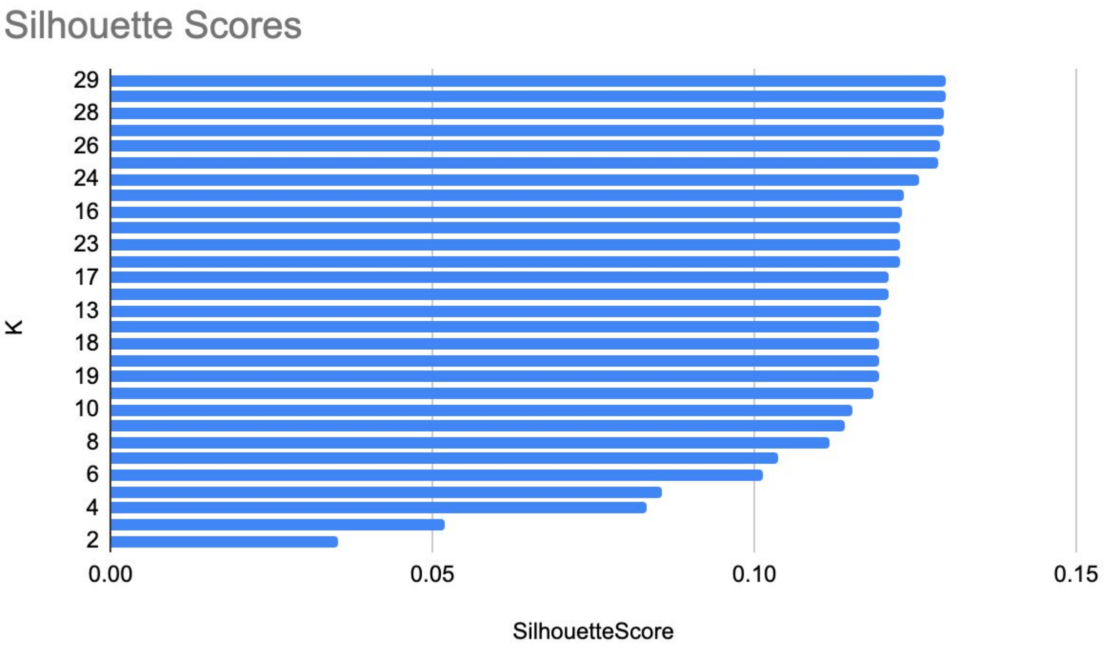
- 1: Traffic
- 2: Emotional
  - negative sentiment
- 3: California
- 4: Traffic
- 5: Informational

7 clusters

- 1: Photo
- 2: Fire
- 3: Emotional
  - positive sentiment
- 4: Informational
- 5: Emotional
  - negative sentiment
- 6: California
- 7: Traffic

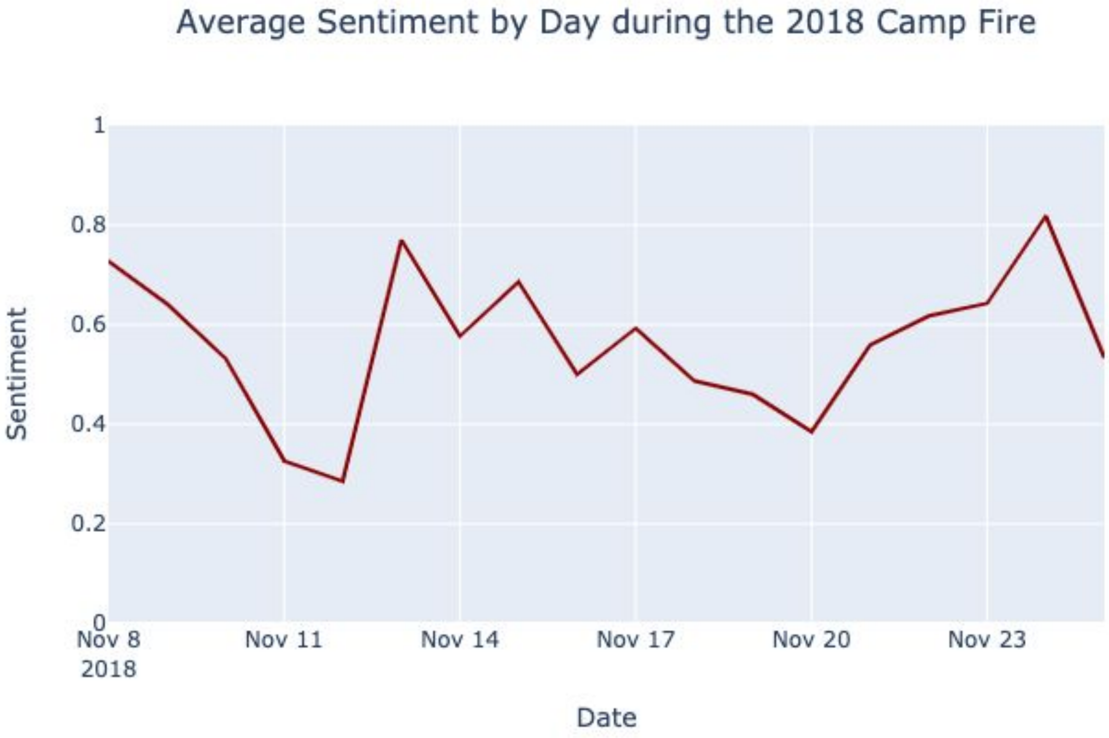
k-Means Clustering

- Cluster evaluation
  - Surprised at the amount of traffic-related tweets
  - Emotional nuance clearer with more tweets
  - 5 clusters as sweet spot between interpretability and granularity
- Silhouette scores
  - Poor silhouette scores
  - Largest increase from 4 to 5 clusters

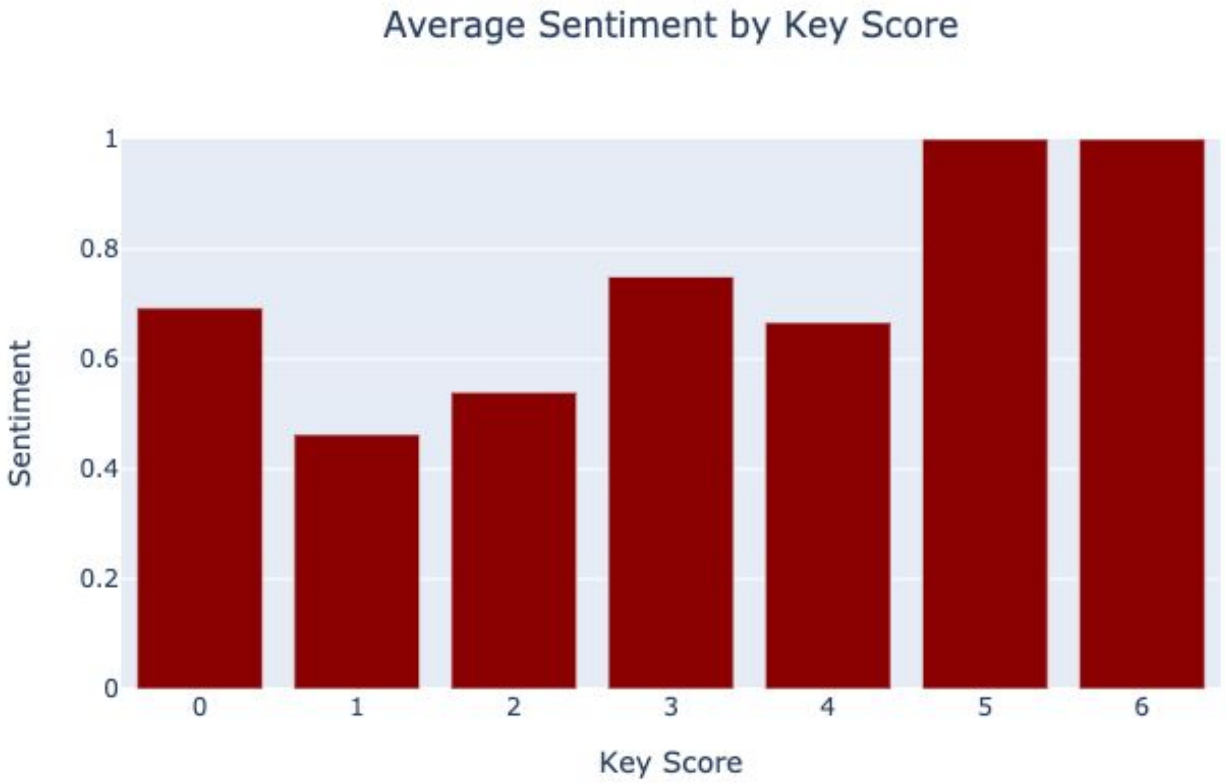


Sentiment Analysis

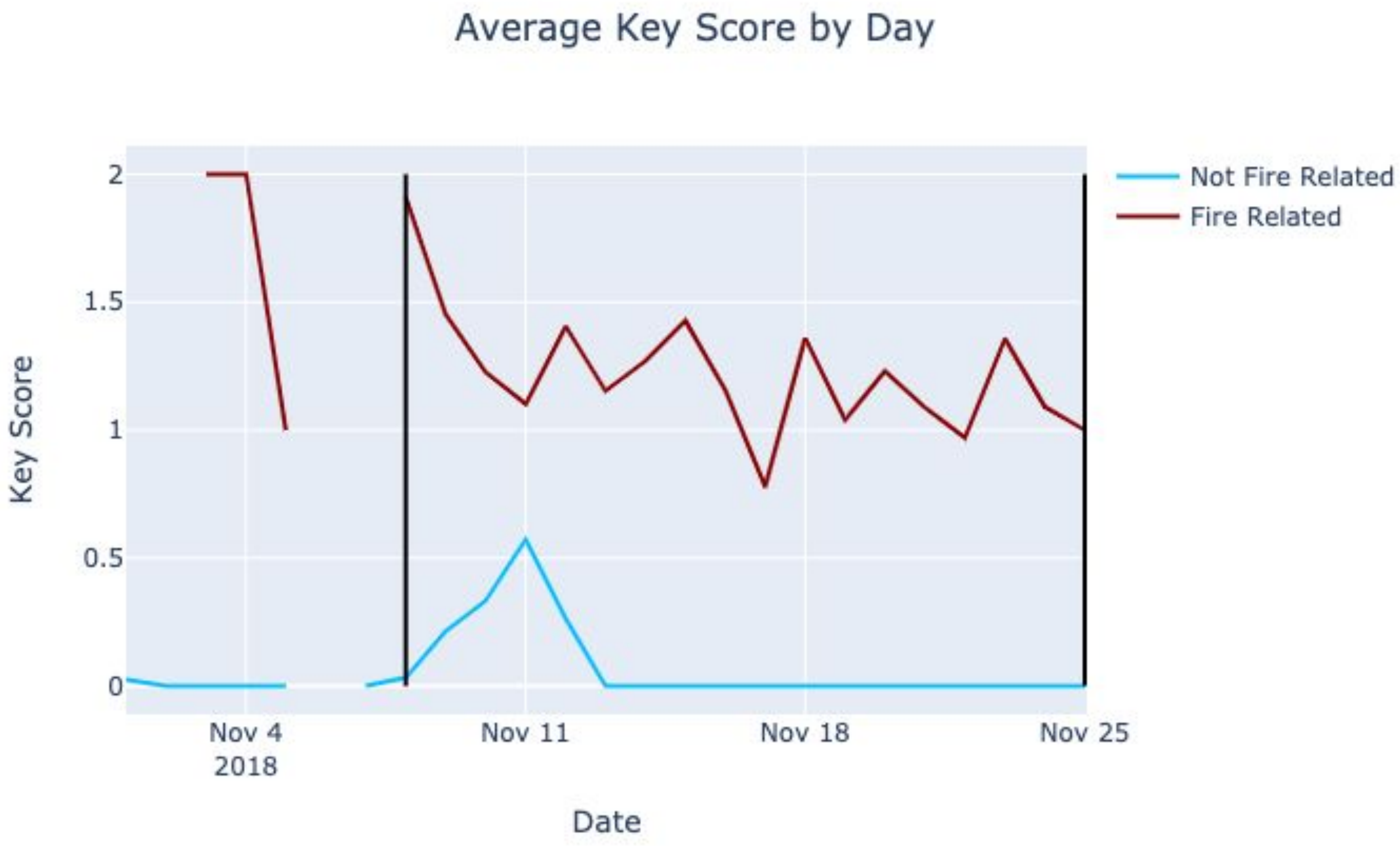
- Positive / negative split
  - Pre-fire: 86/14
  - During fire: 60/40
  - Total: 65/35
- Lowest 4 days after fire started
- Not all tweets negatively classified actually were
  - "Even in the tragic moments of life there are reasons to celebrate...#HappyBirthday #IggyPup. We <3 you!! @Chico, California" (source)



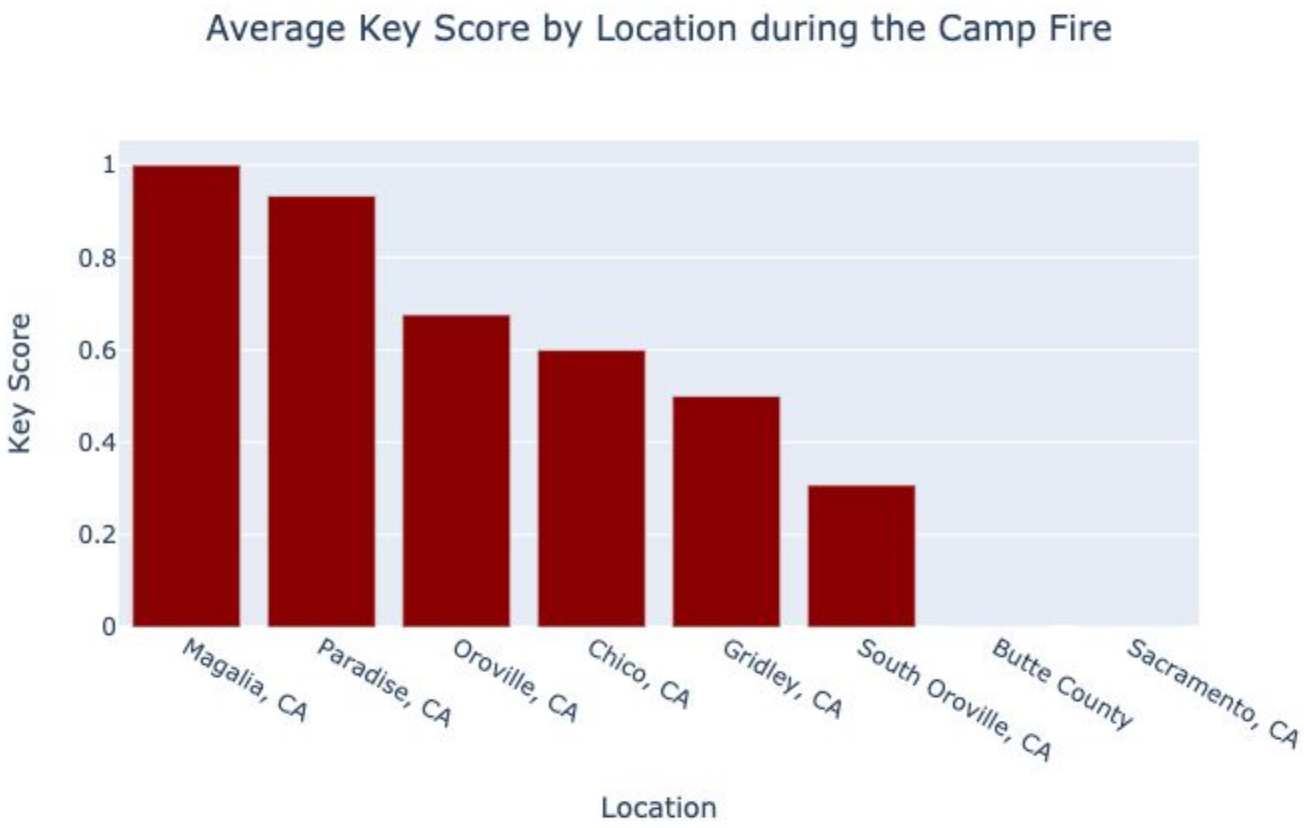
Sentiment by Key Score



Key Score by Date



Key Score by Location



Challenges & Constraints

Challenges & Constraints

- Pulling a sufficient amount of data
- Data from single social media source
- Only text data analyzed
- Not all tweets within 100 mile radius captured
    - Only those with location tag

A question of ethics

- Where is the line between the right to data privacy on using media and using social media data in times of disasters to potentially help save lives?

Recommendations and Next Steps

Whats next?

- Pull in additional Twitter Camp Fire data
- Aggregate data from additional social media sources
  - e.g., 880 of the 1,130 tweets analyzed were originally posted to Instagram
- Improve keyword list
- Try other models
  - e.g., integrate image recognition
- Test and improve on other disasters
  - “Slow” disasters (e.g., fires, floods, hurricanes)
  - “Fast” disasters (e.g., tornados, shootings)
- Develop proof of concept
- Ideal state: deploy product that agencies can implement to monitor social media data

QUESTIONS?