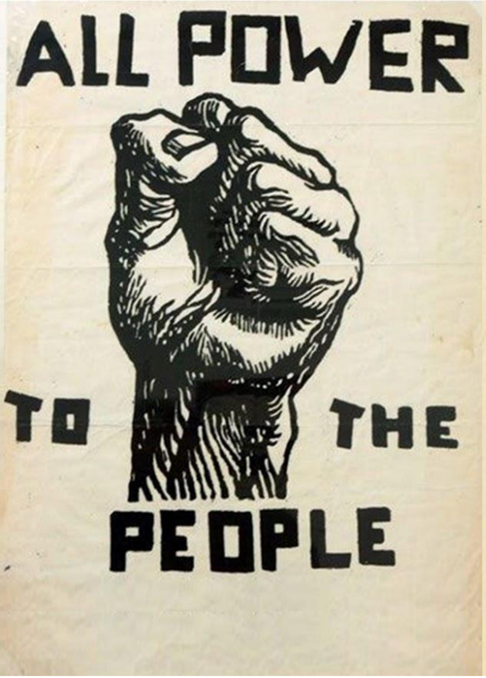
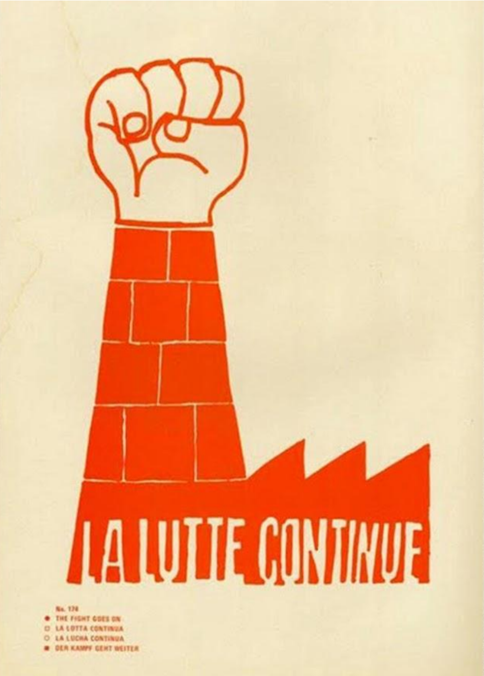
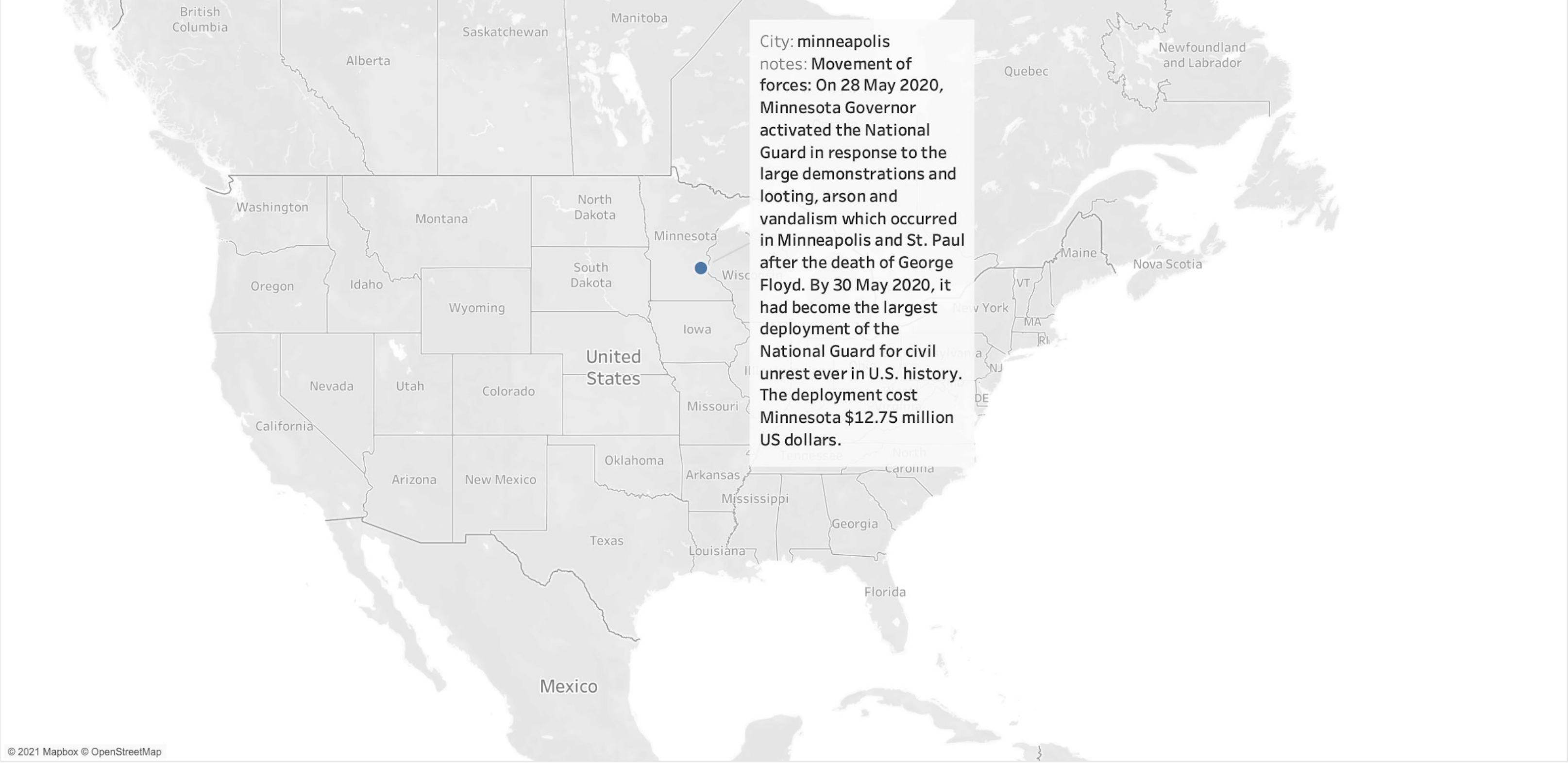




El Popular
DIARIO AL SERVICIO DE LA NACION
MEXICO, D. F., SABADO 1o. DE MAYO DE 1948.

Predicting Violence at Protests

Asia Roy, Gabrielle Burgos, Aida Rahim
National Civic Alliance



City: minneapolis

notes: Movement of forces: On 28 May 2020, Minnesota Governor activated the National Guard in response to the large demonstrations and looting, arson and vandalism which occurred in Minneapolis and St. Paul after the death of George Floyd. By 30 May 2020, it had become the largest deployment of the National Guard for civil unrest ever in U.S. history. The deployment cost Minnesota $12.75 million US dollars.

Is it possible to predict violence erupting at protests within the U.S.?

Background

- May 25, 2020: George Floyd was killed by police in Minneapolis
- National (and international) civil unrest and protest arose in response
- Protests across 2000 cities and 60 countries in support of Black Lives Matter (BLM) movement
- Most protests were peaceful, though some escalated to violence, either from protestors or police

Masked protesters in Philadelphia on June 2
Credit: RGB. - https://www.flickr.com/photos/46437876@N06/49965512681/

Minnesota State Patrol troopers stand in formation.

Credit: Tony Webster - https://www.flickr.com/photos/87296837@N00/49977235136/

Motivation

- Help people decide on protest participation based on predicted occurrence of violence
- Help businesses decide on whether or not to keep stores open during protests, and whether extra protection is necessary for employees
- Help cities decide on implementation of physical barriers and other passive crowd control methods

Problem Statement

Is it possible to predict violence erupting at protests?

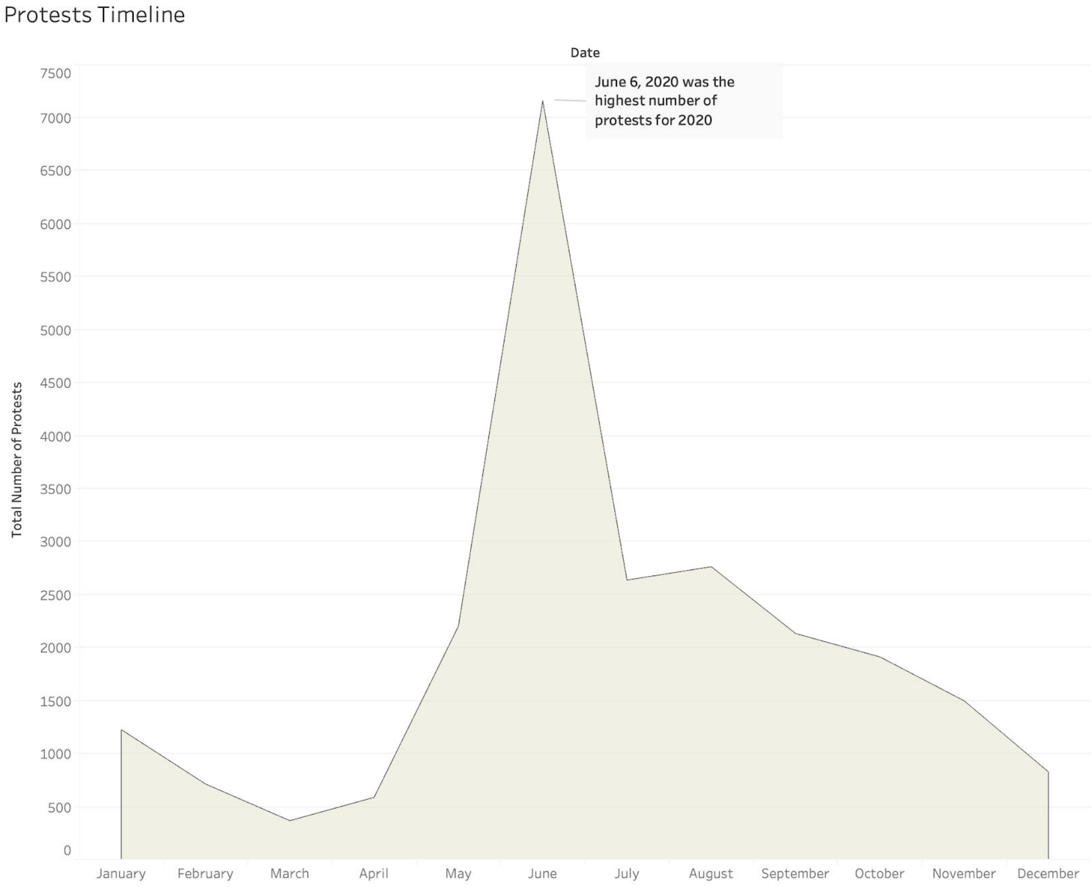
Data Sources

- Protests in the US (2020-2021): The Armed Conflict Location & Event Data Project (ACLED)
    - Protest locations, dates, and classification as well as participant groups
- Numbers of protesters at each protest event in the US (2020-2021): Data.World
    - Protest locations, dates, and protester count
- Population information (from 2014): Dataverse
    - Demographic information for different locations in the US

Data Wrangling

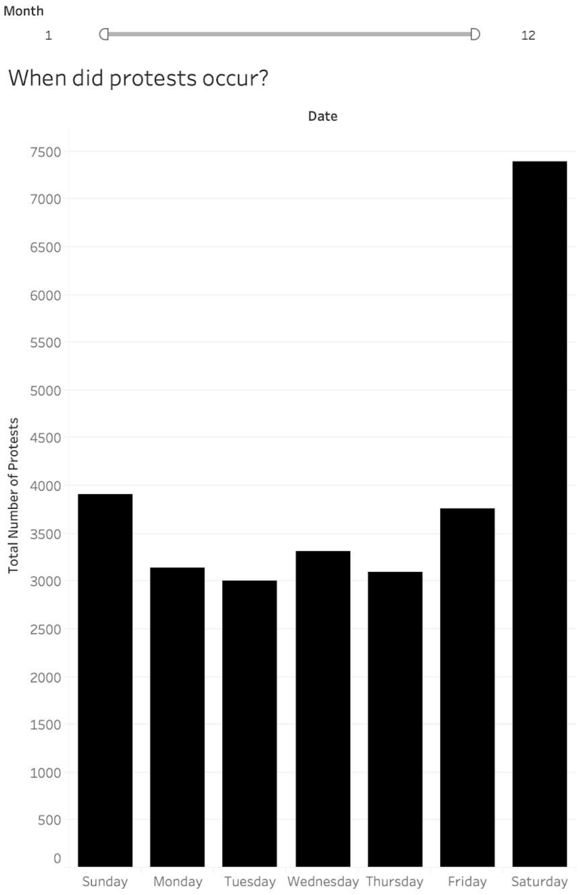
- Create unique IDs to merge dataframe
- Dummify feature of protest participants (organizations)
- Merged dataset had ~27,000 rows (protest events)
- Remaining issues not handled in data preparation:
  - Certain geographical locations have 2 identifiers e.g. New York-Manhattan vs New York-New York; Seattle vs Seattle-CHOP
  - Missing population data
- Drop null values
- Final useable dataset has ~15,000 rows (protest events

Stories The Data Tells Us



Protests Timeline

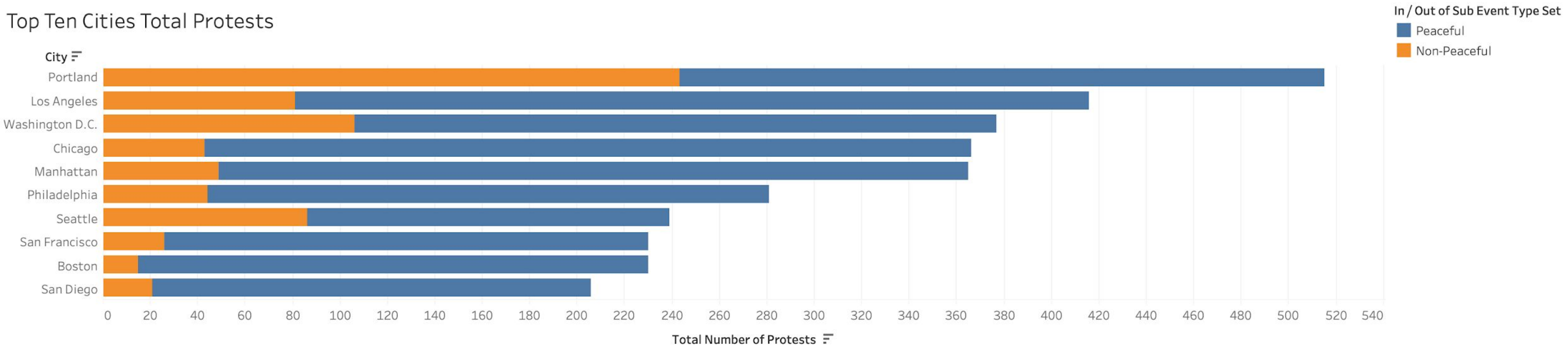
June 6, 2020 was the highest number of protests for 2020



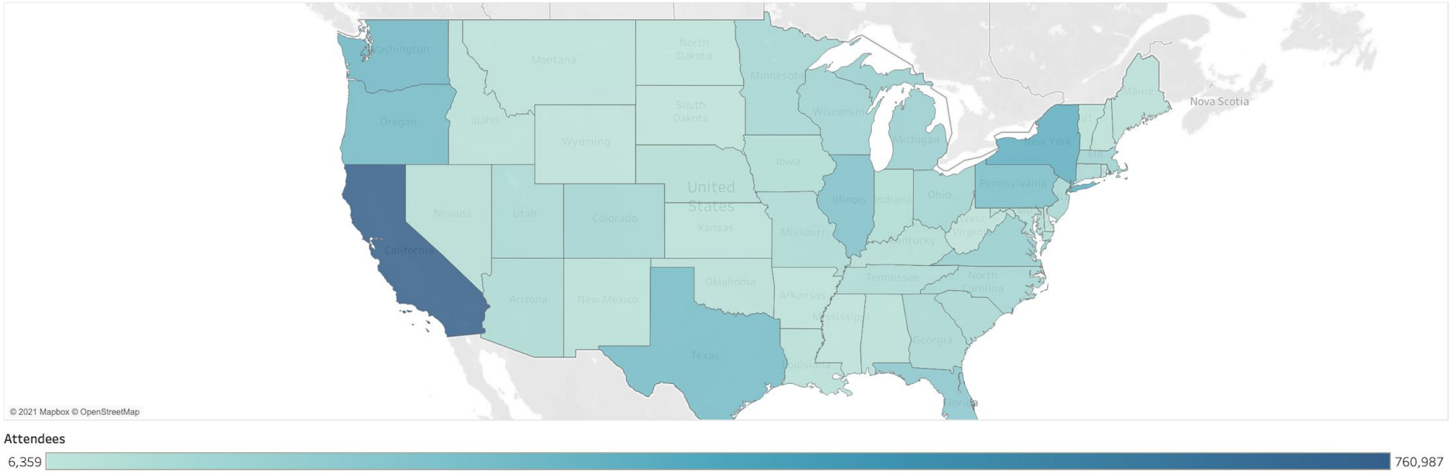
When did protests occur?

Stories The Data Tells Us

Top Ten Cities Total Protests

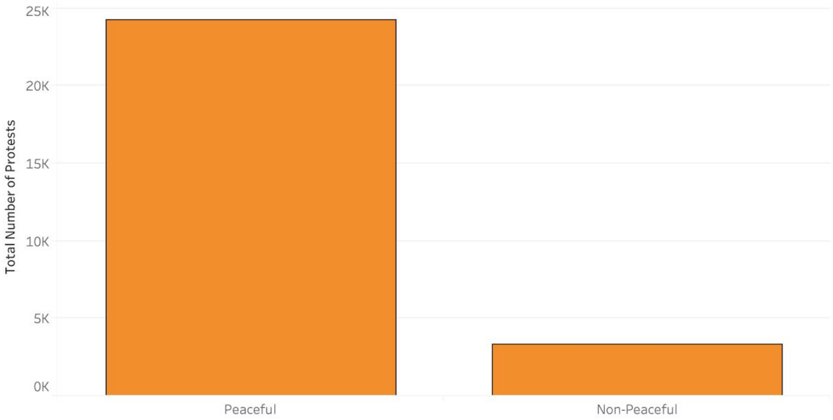


Protests Attendance by State

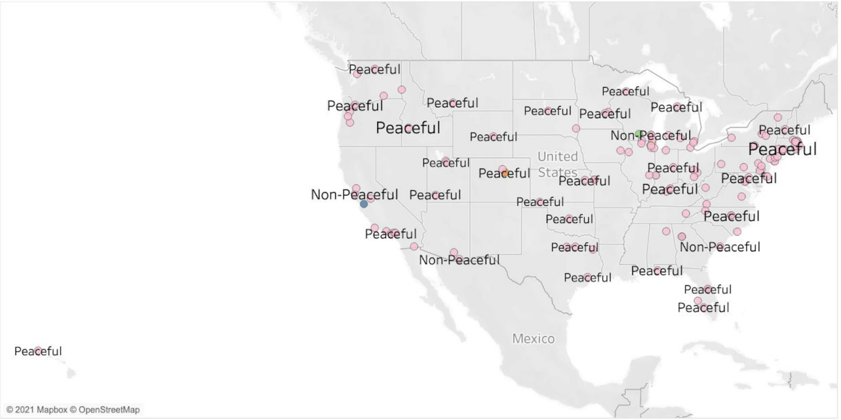


Stories The Data Tells Us

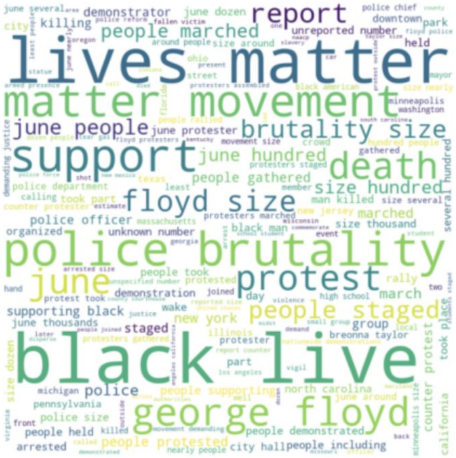
Protests May- Sept 2020



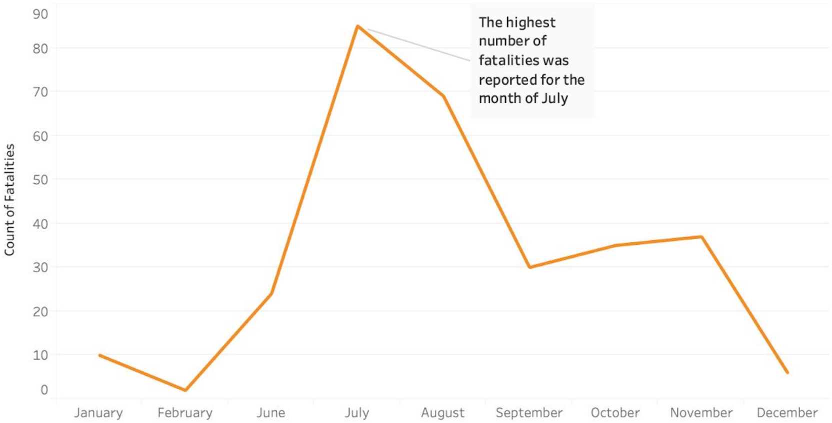
Black Lives Matter Protests



© 2021 Mapbox © OpenStreetMap



Reported Fatalities at Protests



The highest number of fatalities was reported for the month of July

Modeling Methodology - Using Numerical and Categorical Features

1. Features considered: protest host(s), total population, poverty rate, percent of population holding bachelor degrees, mayor status, population political affiliation, historical unarmed death records
    a. Rationale: Perhaps underlying social structure influences presence or absence of protest violence
2. Build models using dataset from June, and apply predictive models on test set as well as protests from other months throughout the year and into 2021
    a. Rationale: Use current data to predict future occurrences
    b. Assumption: Future occurrence is related to current events

Modeling Methodology

	Best Test Score	Model Observation	Best extra-Test Score	Model Observation
Unmodified	0.90 (bsl 0.89)	~ baseline	0.87 (bsl 0.87)	~ baseline
Downsample majority class	0.74 (bsl 0.61)	>> baseline	0.67 (bsl 0.87)	<< baseline
PCA	0.73 (bsl 0.61)	>> baseline	0.63 (bsl 0.87)	<< baseline
Neural network (with PCA)	0.76 (bsl 0.61)	>> baseline	0.66 (bsl 0.87)	<< baseline
Split dataset by city size then build model	0.87 (bsl 0.84)	~> baseline	0.85 (bsl 0.83)	~> baseline

7 classifier models (LogReg, KNC, DecisionTree, Bagging, Random Forest, AdaBoost, SupportVector) + Neural Net

Modeling Methodology - Text Classification Data

Main Feature:

- Protest Data column called “Notes”
- Instantiated both Count Vectorizer and TFID Vectorizer Models
- Applied English language stopwords and additional word frequencies such as: size, 2020, people, june, group

Target Variable:

- Collected from the Protest Data Sub Event column value: (“Peaceful Protest”)
- Sub event column was dummified to create a binary value of 0 and 1 to specify the label.
- Evaluated the baseline accuracy:

Peaceful Protest = 0.878

Non-Peaceful Protest = 0.121

- Addressed the imbalance data by under sampling the dataframe at random.
- Trained and split the undersampled data.
- Pipelined CV and TFID with multiple classification models to predict the outcome of protest during 2020 - 2021.
- Gridsearched each pipeline model to generate best parameters and scores.

Text Classification Pipeline Models

Results	Train	Test	True Positive	True Negative	False Positive	False Negative
CV & Naive Bayes	0.95	0.937	762	817	52	53
CV & Log Regression	0.999	0.967	786	843	28	27
TFID & Naive Bayes	0.943	0.940	767	816	47	54
TFID & Log Regression	0.973	0.965	786	840	28	30
Random F. & CV	0.928	0.923	759	849	55	21
Random F. & TFID	0.951	0.928	746	817	68	53
CV & Support Vector M.	0.995	0.967	788	841	26	29
TFID & Support Vector M.	0.975	0.967	786	844	28	26

Sentimental Analysis

Most Neutral:

“On 18 August 2020, a rally was held in Richmond (Virginia) to call for police reform.”

Negative	Neutral	Positive	Compound
0	1	0	0

Most Negative:

“On 2 July 2020, an unknown number of relatives and friends of a teenager died from gun violence rallied in Atlanta (Fulton, Georgia) to denounce gun violence.”

Negative	Neutral	Positive	Compound
0.448	0.448	0.104	-0.9393

Most Positive:

“On 26 September 2020, about 80 people marched in a rally in Buffalo Grove (Illinois) for promoting peace and love.”

Negative	Neutral	Positive	Compound
0	0.638	0.362	0.8807

Summer 2020 Protest Outcomes

- Over 20,000 protests with very low levels of violence and most violence that did take place was directed against the protesters.
- **93%** of events involved no property damage or injuries.
- Police made arrests in 8% of the protest events.

Recommendations

- Protest participation is an important civic duty and the majority of events are peaceful
- Strong indication that this target (predicting violence at protests ahead of time) can be achieved, but requires additional information:
    - Review data, applying a more extensive data cleanup protocol
    - Apply more intelligent clustering of dataset prior to modeling
    - Reduce imbalance in dataset
    - Apply PCA before modeling
    - Combining text data with more opinionated datasets either from law enforcement or protester's perspective. For example: twitter, subreddit channels, blogs
    - Possibly relate text classification to police brutality datasets provided by protesters and activist