Violent Sentiment and Crime in American Cities, 2000-2021

- Name of the group members and CNetIDs
    - Capptivators:
        - slogan
        - gnogueda
        - ajsaenz
        - rodrigosalazar

- A brief overview of the final project (200 words maximum)

    This project explores whether violent sentiment in a city, as represented by tweets, has any correlation with actual crimes in that city and whether that relationship changes over time. We combined tweets containing violent language with FBI data on crimes in the largest US cities from 2005-2021. The dashboard aims to visually display (time and cross-sectional) trends of violent sentiment on twitter and actual crimes to help to discover patterns that relate both sources of data. Data is disaggregated by type of crime, using FBI categories, and twit category, which was constructed using clustering analysis by topic in Twitter. We used both sentiment analysis (VADER) and clustering techniques (jaccard similarity) to categorize tweets by attitude and category under the umbrella of violent language. This analysis could be a helpful indicator in understanding whether violent sentiment is a lagging or leading indicator of actual crimes and whether a city can change challenging problems both culturally in the way people speak to one another and in terms of actual crime rates.

- The overall structure of the software (1-page maximum). It would be nice to include a helpful diagram of how the modules are connected with each other but this is not required.

    See proj_structure.pdf for a diagram of the project

    Data & Analysis - Twitter

    The twitter analysis was done in three parts. Firstly, we scraped the data using Twint, then we processed and cleaned the data and finally we ran several analyses on the data and aggregated those analyses to the city level.

    To scrape the data we created a corpus with vocabulary related to violence. We wanted to capture data and language related to a broad range of violent topics, from crime to xenophobia (which we will use to cluster into specific groups later). In Twint, we specified a list of cities we were interested in and specific dates (2015-2021). We cleaned the tweets so that they only included text (scraped punctuation and emojis) and excluded ‘stop words’, i.e. commonly used English words that would not give us additional information about the topic or sentiment of the tweet.

To analyze the tweets, we used two Natural Language Processing algorithms, specifically jaccard similarity and sentiment analysis using VADER. We evaluated using advanced clustering techniques such as LDA, but because different categories of violent language may have a lot of overlap, we decided to construct manual corpus' of four categories. They include hate, crime, guns and addictions. We used jaccard similarity to categorize a tweet with a cluster category. We then used VADER to categorize the tweet as positive, negative and neutral. We decided to go with VADER because of how much data the model has been pre-trained on, and how widely it is used.

Lastly, we aggregated the scores and sentiments by city and year. We decided to take the counts of each type of score and the sentiment for each score (positie, negative and neutral).

FBI data

For every year from 2005 to 2019, the FBI provides data for all Offenses Known to Law Enforcement by State by City. This databases also contain information on city population (for each year). We cleaned and merged this data in order to obtain a centralized city-level time series database.

Additionally, we used data on city latitude and longitude from simplemaps (https://simplemaps.com/data/us-cities). This data enabled us to present all data (FBI and Twitter) in a map.

Visualizations

The visualization displays a map and two graphs. It gives you the option to display either the top ten major cities or all cities in the US, but twitter data is only available for the ten major cities. Additionally, you can hover over each city to understand more about the stats for each one. You can also filter the crime data from 2005-2019 and by crime type (arson, property crime, larceny theft, violent crime, aggravated assault, robbery, burglary, motor vehicle theft, rape, murder). You can display either the crime rate per 100,000 inhabitants, or the absolute number of crimes. We also display two additional graphs: 1) a breakdown of crimes by type and 2) time trends of (selected) sentiment and number of (selected) crime. To display these graphs, you can hover over each city and the data will update to display for that city.

- A description on the code responsibilities for each group member (i.e., who was responsible for what module, files, tasks, etc.).
    - Sophie & Rodrigo - twitter scraping, cleaning twitter data, sentiment analysis, clustering
    - Alex - merging and cleaning fbi data
    - Kenia & Alex - visualizations

- Short description on how to interact with the application and what it produces.

    You can run the app by running **python3 -m crime_sentiment** in your command line from the *crime_sentiment* directory. It will open a dash visualization with a map and several charts, which you can toggle by crime type, year and other variables. The visualization runs on pre-scraped, pre-cleaned, and pre-aggregated data, but the code to do all of that data work is within the

sentiment folder (twitter_data.py uses twint to pull the data, tweets_df_pro.py runs the sentiment/clustering analysis, and aggregation.py is the aggregation function). Additionally, the pre-pulled data exists within the data folder.

- What the project tried to accomplish and what it actually accomplished (200 words)

  At first, we wanted to join city budget data to tweets to analyze if people had feelings about how good the public goods in their city were and whether that had a relationship with the budget for those public goods. What we found was that not many people were talking on Twitter about their garbage collectors or how good the tree pruning on their street was. We shifted gears to examining the relationship between how much violent expression there was in a city and the crime level, with a hypothesis that high levels of violent expression might be correlated with higher levels of crime. Visually, the time trends suggests a positive correlation between crime and sentiment data for most cities and types of crime. One of the main limitations of the project is that we were able to obtain sentiment data only for 10 cities and 5 years. Thus, given the preliminary suggestive correlations, we see a potential to conduct more sophisticated time-series analysis, provided more data is obtained. We see many ways to further explore the relationship between public sentiment and objective measeres of crime. Therefore, we see our project as an instrument to sheds light on possible avenues of future research.