PPOL 5203 Data Science I: Project Proposal
Group Members: Josahn Oginga, Marilyn Rutecki, Muhammad Saad

Clicks and Crimes: Forecasting the Virality of Crime News in Washington, DC

Problem Statement

Washington, D.C. has high rates of localized crime and its crime rates are rising precipitously each year. This figures prominently in online content generated and consumed by city residents across different social media, news media, and microblogging platforms. Although it is clear that the consumption and generation of crime-focused content likely remains high for DC residents, less is known about why certain types of content garners more engagement, specifically virality, on content platforms. Moreover, not every crime gets reported by media outlets. Instead, coverage is selective and may depend on factors that affect residents' sensibilities.

This project will focus on two aspects. Firstly, what are some of the crime-related content characteristics that likely affect virality. Secondly, what are the factors that may make it more likely for news outlets to cover or report a particular crime. The project will focus on the socioeconomic and psychological factors that cause DC citizens to share or engage with crime related content, or result in coverage by news outlets. These factors constitute the dependent or predictor variable(s). The predicted variables are virality and likelihood of news coverage, including different metrics available on engagement.

Research Design

Predictor Variables

No.	Primary Predictor	Description	Subcategories	Initial Hypothesis
1.	*Emotionality or valence*	The emotional response generated by the content for the user. Although emotionality can both be positive and negative, we focus on *negative sentiments* as they are more likely to influence virality	Focus on three sentiments: 1. Anger 2. Anxiety 3. Sadness	Anger and anxiety likely to be more substantive predictors for virality compared to sadness
2.	*Nature of crime*	The type of crime that was committed. We will rely on the broad categories of crime used by Was	Categories of crime (not exhaustive): 1. Carjacking 2. Burglary 3. Homicide 4. Domestic violence	Violent crimes more likely to evoke a stronger response and online engagement

3.	*Characteristics of the alleged offender(s)*	Personally identifiable characteristics of the likely or convicted perpetrators.	Demographic/identifiable characteristics: 1. Race 2. Gender 3. Age	Racial factors, especially those focused on non-white populations, may predict virality
4.	*Locality*	The particular ward or neighborhood where the crime was committed.		Crime in more affluent neighborhoods or wards will likely generate more engagement

Predicted Variables
The predicted variables include:

1. The number of comments on a specific social media post
2. The number of engagements, specifically likes or dislikes, or upvotes or downvotes, a piece of content generates
3. The number of shares a piece of content receives
4. Whether a particular crime gets covered and reported by news outlets

Control Variables
The model will also control for factors likely to determine virality and news coverage, but not of interest that may also cause endogeneity. These include: DC's population breakdown by race, gender and age, the time when a specific post was made on a content, amongst others.

Data Sources
The project will have a range of data sources. For data related to virality, the project will focus on content platforms such as Reddit, X, and YouTube comments. For news coverage of crime in DC, locally focused outfits such as Fox 5 DC, and NBC DC. The project will also utilize the Washington, DC Metropolitan Police Department's detailed statistics on crime as a counterfactual.

Plans to Obtain Data

- We understand that Reddit has now restricted its researcher API. If not available, the project will rely on third party scrapers/wrappers such as pushshift.io and PRAW.
- Local news outlets, specifically NBC4 and Fox5 that focus on the Washington DC area. These news outlets will not measure virality but will corroborate crimes that made it to the news. Fox5 has a search button that enables news search by keyword, in our case, types of crimes. We will likely manually scrape data from these platforms using BeautifulSoup as they do not have an API.
- YouTube API can be a resource to gauge news engagement as YouTube has comments, thumbs up, and thumbs down buttons.
- Additional data can be obtained by scraping X (Twitter) using hashtags or profiles of DC crime trackers such as @RealTimeNews10, @crimedatadc, and @dccrimefacts. Same criterion on impression, retweets, and likes can be used to gauge virality.
- Metropolitan Police Department data

Data Analysis Components

- Machine learning or predictive analytics component, specifically to predict the virality or likelihood of coverage by news outlets for a particular type of crime.
- Text analysis component to analyze sentiments on content platforms related to particular crimes. Here, we want to be able to code for sentiments such as anger, anxiety and sadness.
- Descriptive statistical analysis of DC crime data.

What Will Success Look Like?

Success would be comprehensive scraping of web data, statistical analysis of DC crime data, and successfully predicting the virality or news coverage of a particular crime using our model. We understand that we may be overcommitting and casting a wide net with our initial proposal. However, we hope to have more clarity as we progress with our data collection and research. We will consider the project to be a success if we develop a sound predictive model, manage to scrape data as required and also do robust sentiment categorization. Overall, we hope that our study will shed light on the specific attitudes and socioeconomic/psychological factors that determine citizens and news outlets' sensibilities around crime in the city.