A circuit board

Description automatically generated

## Scenario

You work for a US software company working with the US government. Your client is concerned with interactions on social media related to civil unrest and political protesting in order to help identify themes that can inform improved policing and policies. Historically the client would directly interview protesters, asking for social media handles directly. After which the organization would research the posts of the individuals looking for qualitative insights related to their research. Since not all user posts are relevant, the research required time intensive manual labeling of social media posts from a particular user. The labels would be related to civil, political discourse or not. Below is a labeled example from a twitter user.

Related to Political Discourse Not Related to Political Discourse

A screenshot of a cell phone

Description automatically generated A close up of a dog

Description automatically generated

You have been asked to explore the existing dataset of manually labeled documents then create a document classification system. This will save money by avoiding human labeling and enable quicker identification of users and their related texts.

## Technical:

* Provide some insights such as frequent words, averages and other EDA about documents etc.
* Create a document classification model to identify political discourse documents
* Calculate the Accuracy of the model for a training and validation set.
* Build the model with 2000 documents, and classify a test set of 1000 documents, and provide their probabilities.

## Project Deliverables include

1. R scripts for exploration “lastName\_TM\_HW2.R”
2. R script for model construction “lastName\_TM\_HW2.R”
3. CSV of the 1000 scored documents “lastName\_TM\_scores.csv”

## Example Data

|  |  |  |
| --- | --- | --- |
| doc\_id | text | Label |
| 1266537590984851456 | "RT @HKrassenstein: @JoeBiden As Trump cowers in the White House, Joe Biden is actually on the ground where the protests are taking place" | 1 |
| 1267649652704243712 | RT @day6siete: Uhm. Hi! I'm new to the fandom. I make fanarts. Please be nice to me~” | 0 |
| 1267310955089797120 | "RT @AhmedBaba\_: Houston Police Chief Art Acevedo gives a speech you'd think was from a protestor. THIS is how you do it. | 1 |