A circuit board

Description automatically generated

## Scenario

You work for an US based non-profit thinktank concerned with interactions on social media related to civil unrest and political protesting. Historically the organization would directly interview protesters, asking for social media handles directly. After which the non-profit 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 let the non-profit 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 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.

## Non-Technical:

* Describe the preprocessing steps and why are they are applied to the documents
* Describe the modeling technique in a powerpoint to a non-technical user

## Project Deliverables include

1. R scripts for exploration “lastName\_TM\_case\_exploration.R”
2. R script for model construction “lastName\_TM\_case\_model.R”
3. CSV of the 1000 scored documents “lastName\_TM\_scores.csv”
4. Powerpoint of any visualizations, findings and descriptions of non-technical results as if presented to the thinktank’s leader. “lastName\_TM\_case\_presentation.pptx”

## 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 |

## Criteria for Success

The case material will be evaluated according to the following criteria. Each is worth 5pts for a total of 20pts.

## **Organization of content**– Logical ordering of ideas, modeling artifacts, applicable visualizations in slides

## **Organization of code**- R Code is well organized, concise, and free from error

## **Data mining Process** – Recognize the type of data mining problem, adherence to established main data and text mining steps.

## **Completeness** – Understood impact, and mined the data for relevant insights/recommendations