Problem Set 4: Predicting Tweets
"A rose by any other name would smell as sweet"
Juliet Capulet

There is an adage that says: *"choose your words carefully."* Words themselves may reveal far more than what we're trying to say. There's mounting evidence that our written words show who we are.

The objective is to predict to whom each tweet belongs. The training dataset contains tweets from three prominent Colombian politicians' accounts: Claudia Lopez, Gustavo Petro, and Alvaro Uribe. The test dataset contains 500 unlabeled tweets. We want to predict which account posted the tweets in the test set.

There are two expected outputs:

1. A `.pdf` document.
2. Submissions with your team's predictions in Kaggle at the following link.

The document must contain the following sections:

- Introduction. The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.
- Data[1]. When writing this section up, you must:
    1. Describe the data, and the sample construction process, including how the data was cleaned, combined, and if new variables were created, how they were created.
    2. Present a descriptive/explanatory analysis of the data using text visualization techniques.
- Model and Results. This section presents the model(s) submitted for evaluation. When writing this section up, include:

[1]This section is located here so the reader can understand your work, but it should probably be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here.

  - A detailed explanation on how it was trained, the selection of hyper-parameters, and any other relevant information.
  - A comparison to at least 4 other specifications submitted to Kaggle.

- Conclusions and recommendations. In this section, you briefly state the main take-aways of your work.

1 Additional Guidelines

- Predictions have to be submitted on Kaggle. Check the competition website for more information.

- Turn a `.pdf` document in Bloque Neón. The document should not be longer than 8 (eight) pages and include, at most, 8 (eight) exhibits (tables and/or figures). Bibliography and exhibits don't count towards the page limit. You are welcome to add an appendix, but the main document must be self-contained. Specifically, a reader should be able to follow the analysis in the paper and be convinced it is correct and coherent from the main text alone, without consulting the appendix.

- The document must include a link to your GitHub Repository.
  - The repository must follow the template.
  - The README should help the reader navigate your repository. A good README helps your project stand out from other projects and is the first file a person sees when they come across your repository. Therefore, this file should be detailed enough to focus on your project and how it does it, but not so long that it loses the reader's attention. For example, Project Awesome has a curated list of interesting READMEs.
  - Include brief instructions to fully replicate the work.
  - The main repository branch should show at least five (5) substantial contributions from each team member.
  - The code has to be:
    * Fully reproducible.
    * Readable and include comments. In coding, like in writing, a good coding style is critical. I encourage you to follow the tidyverse style guide.

- Tables, figures, and writing must be as neat as possible. Label all the variables included. If you have something in your figures or tables, I expect they are addressed in the text. Tables must follow the AER format.