# Social Media Data Science Pipelines Project 2: Dataset Measurements and Analysis

October 24, 2023

## 1 Introduction

Once you have a solid data collection system in place, the next step is to *do* something with it. In this project, you will design and execute measurement and analysis experiments to get a better understanding of your data. Additionally, you will add a new component to your data colleciton system that measures toxicity of content. Finally, you will begin work on answering questions more specific to *your* dataset.

## 2 Project Description

While data collection is arguably the most important part of the data science process, just hoarding data, while fun, has minimal impact. Thus, in addition to collecting data, we must make use of it somehow.

Designing measurement and analysis experiments is *not* a particularly easy task. It involves a lot of thinking (sometimes wishful), and often a lot of *learning* before writing a single line of code. This project is to help you get some experience in doing this.

You will thus design a set of experiments aimed at *describing* your data. You will also come up with a set of *at least three* research questions. While you might attempt at answering these research questions for this project, it is not necessary. Instead, you should focus on how you might answer them for Project 3.

### 2.1 Measuring Toxicity

For this project, you will make use of a 3rd party API to add real time measurements on toxicity to your data collection system: ModerateHatespeech (`https://moderatehatespeech.com`). You must create a key to access ModerateHatespeech, but they are free, and you have unlimited requests.

*You* are fully responsible for learning about ModerateHatespeech and implementing a client that makes use of it. You are not allowed to use any library besides a regular HTTP client (i.e., what you were restricted to in project 1).

Hints:

- There is documentation on the site, however, that documentation is not (in my experience) 100% complete. I would be shocked if you did not experience a variety of undocumented edge cases and errors as you build your system.
- You should not consider ModerateHatespeech to be 100% available. It has been stable for the past couple of weeks at the time of this writing, but I have experienced numerous outages. Your system needs to be robust against this.
- It is a *much* better idea to be accessing the API in approximately real-time with respect to when you are ingesting data. Requests do not instantly complete, and if you don't amortize things, you might not be able to get scores for all your data by the time deadlines hit.
- Project 3 will expect ModerateHatespeech scores to be coming in approximately real-time with data, so it's probably better to do this from the get go.

# 3 Project Deliverables

There are three deliverables for this project.

1. Project proposal
   - GitHub Classroom Section 1: `https://classroom.github.com/a/ZF-ngKtP`
   - GitHub Classroom Section 2: `https://classroom.github.com/a/L8jHqc8y`
   - Due 11:59PM Friday, November 3rd, 2023.
2. Project implementation.
   - GitHub Classroom Section 1: `https://classroom.github.com/a/T3nUFdmQ`
   - GitHub Classroom Section 2: `https://classroom.github.com/a/XRb6DLgP`
   - Due 11:59PM Thursday, November 30th, 2023.
3. Project report.
   - GitHub Classroom Section 1: `https://classroom.github.com/a/Eb0xwJTh`
   - GitHub Classroom Section 2: `https://classroom.github.com/a/Wh97FqLM`
   - Due 11:59PM Thursday, November 30th, 2023.

## 3.1 Project Proposal

The purpose of your proposal is to ensure that: 1) you are not attempting to do something impossible, 2) you are not attempting to do something illegal, 3) you are not attempting to do something too easy. Your proposal should provide enough information that Jeremy can read it and have a rough idea of what it is you plan to do, and with enough detail that Jeremy can help you avoid pitfalls that he has experienced in the past.

To this end, we suggest your proposal have several sections:

- An introduction that *motivates* any experiments you will perform.
- A proposed methodology section which sketches out how you intend to describe your data.
- A section noting what, if any, additional data you need to collect, and also validates that you think you have enough data to perform the experiments.

Your proposal should be one to two pages. **Your report must conform to the two column ACM 'sigconf' format** available here: `https://www.acm.org/publications/proceedings-template` and *must be submitted as PDF*. If your proposal does not conform to this format, or you submit something besides a PDF then you will receive a zero.

## 3.2 Project Implementation

You will be required to submit all the measurement and analysis code you created. While there are essentially no restrictions to what libraries you might use, there are some ground rules:

- No Excel. If we see a plot that was generated using Excel you will receive a **zero (0)** on your report.
- If you need to use tools like SPSS, SAS, Matlab, etc., please talk to me. We would heavily discourage the use of these tools over Python or R, but there are situations where it makes more sense to use them.
- We want to minimize the amount of button pushing, thus if you think there is some off the shelf program you want to use, please speak to me first.
- No pie charts. If we see a pie chart, it is a **zero (0)** on your report.

## 3.3 Project Report

The major deliverable for this project is the report. While there is certainly lots of code to write, the bigger picture is to communicate your results.

This report will be much closer to a "real" research paper, and so we suggest considering structuring your paper as follows:

- An abstract that provides a very high level overview of your report (about 250-500 words probably). Writing a good abstract is something that usually takes some time to figure out, but it's worth giving it a shot.
- An introduction section that *motivates* your work. You can probably re-use a lot of the text from your proposal, but we would be surprised if there wasn't new things to add/update.
- A background and related work section that educates the reader on the problem domain you are working in and illustrates some of existing scientific literature that informs the present work.
- A section describing your datasets.
- A discussion and conclusion section (these could be two sections). This section should contain a rough summary of your report. as well as explicitly discuss the *implications* of your findings. What are the limitations of your work, and how might those limitations be addressed by future work in this area, etc.?
  - **NB:** This section *must* explicitly note your (at least) three research questions that you intend to answer by the end of the course. Ideally, after understanding your data in this project, you will have a better grasp on how to move forward with future work.
- A references section. This is mandatory. Cite things!!!

There are some additional requirements:

- Your report **must include** *at least one table* and *at least six figures* that describe your dataset. These figures must be properly captioned, labeled, and referenced in the text. I.e., you need to tell us something about the figures in your text.
- *At least one* of your figures must have all your datasets on it. I.e., they should allow the reader to directly compare the datasets on the same x- and y-axis.
- *At least one* of your figures must have at least *two* of your datasets on it. This is a separate figure from the above, but it could be another figure that has all three datasets on it.
- *At least one* figure must use data from ModerateHatespeech.
- In general, you are heavily encouraged to *directly* compare your datasets in as many of your plots as possible.
- In addition to the above requirements, your report must contain one additional figure that plots on the x-axis time and on the y-axis the number of *submissions* that were made in the r/politics subreddit from November 1st, 2023 until November 14th, 2023. The x-axis should be binned *daily*. I.e., the plot should be the number of submissions that came in each each day from Nov 1st, 2023 to Nov 14th, 2023 (inclusive).
- In addition to all above requirements, grad student groups must also plot the number of *comments* per day on the r/politics subreddit from November 1st, 2023 to November 14th, 2023 (inclusive) with the date on the x-axis and the count on the y-axis. For this plot, the x-axis should be binned *hourly*.

- **NO PIE CHARTS**. You will receive a **ZERO (0)** if you submit something with a pie chart.
- **NO EXCEL PLOTS**. You will receive a **ZERO (0)** if you submit something that has plots made in Excel.

# 4 Grading

- Proposal is worth 25 points.
- Implementation is worth 50 points.
- Final report is worth 25 points.