Social Media Data Science Pipelines Project 2: Measurements and Analysis

November 9, 2021

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 answer research questions with your data.

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 to answer *at least three* research questions. Please note that I do not necessarily expect you to be able to answer all the research questions you ask, but I will help ensure you have a shot, and you should make an attempt to answer everything you can.

3 Project Deliverables

There are three deliverables for this project.

1. Project proposal.

2. Project implementation.

3. Project report.

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, I suggest your proposal have several sections:

- An introduction that *motivates* the experiments you will perform.
- A section that describes your research objectives, and includes your (at least) three research questions.
- A proposed methodology section which sketches out how you intend to answer your research questions.
- 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 analysis code you created. While there are essentially no restrictions to what libraries you might use, there are some ground rules:

- No Excel. If I 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. I would heavily discourage the use of these tools over Python or R, but there are situations where it makes more sense to use them.
- I 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.

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 I 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 I would be surprised if there wasn't new things to add/update. **NB:** This section should explicitly note your (at least) three research questions, as well as the specific *contributions* your work makes.
- 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 dataset(s).
- A methodology section, which describes the more technical details of *how* you perform the experiments. E.g., if you build a data processing pipeline, a description of each component, any specific techniques or algorithms they uyse, etc. should appear in this section. Another way of thinking about this section is "Experimental Setup."
- One or more sections on "results." How you organize these (different sections or subsections) is mostly up to you. While the presentation of a written paper is very important, there is no exact formula and I personally prefer if people are able to express themselves in their own "voice" as much as possible.
- A discussion and conclusion section (these could be two sections). This section should contain a rough summary of the rest of the paper, as well as explicitly discuss the *implications* of your findings. What are the more abstract things we can take away from your results? How can your work be built upon by future work? What are the limitations of your work, and how might those limitations be addressed by future researchers building upon your wokr, etc.?
- A references section. This is mandatory. Cite things!!!

There are some additional requirements:

- Your report **must include** *at least one table* and *at least four figures* that describe your dataset/results. These figures must be properly captioned, labeled, and referenced in the text.

- In addition to the above requirement, your report must contain a figure that plots on the x-axis time and on the y-axis the number of tweets that came in over the sample stream starting Nov 11th, 2020 through Nov 21st, 2020. The x-axis should be binned hourly. I.e., the plot should be the number of tweets that came in each hour from Nov 18th, 2021 to Nov 25th, 2021 (inclusive).
- **NO PIECHARTS**. You will receive a **ZERO** if you submit something with a piechart.
- **NO EXCEL PLOTS**. You will receive a **ZERO** 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.