COMP 598 Final Project - Data Science Project

Assigned Nov 13, 2020
Due Dec 7, 2020 @ 11:59 PM

This is a GROUP assignment. This document contains a project description and a fine print section.

Overview

Your team has been hired by a not-for-profit that wants to understand candidate discussion in the days following the US election. They have indicated that they are especially concerned with perceptions of election legitimacy. Specifically, they want to know:

1. The salient topics discussed around each candidate and what each topic primarily concerned
2. Relative engagement with those topics among liberals and conservatives

topics for each candidate

tf-idf posts (10 most commonly occuring word, and then engagement with those topics

You will conduct this analysis and submit an 8-10 (double-spaced) page report discussing your findings.

Analysis Details

Your analysis will draw on the Reddit posts (not comments) from two subreddits: /r/politics and /r/conservative. While not perfect, these have been found to roughly line up with liberal and conservative US communities.

Task 1: 333 from each subbreddit Deciding between Hot and new posts

To inform your analysis, you should collect 1,000 posts from each over a 3 day period. Keep only posts that mention either Trump or Biden. Of these, conduct an open coding on 200 posts to develop the topics (approach the exercise requiring each post to belong to exactly one topic). You should aim for between 3-8 topics in total.

Task2: Collect Trump or Biden posts (REGEX) Task 3: Topics Task4: Check all the posts

Once your topics have been designed, manually annotate the rest of the candidate-mentioning posts in your dataset. While double annotation would usually be used, for this project (given time constraints), single annotation will be sufficient.

Characterize your topics by computing the 10 words in each category with the highest tf-idf scores (to compute inverse document frequency, use all 2,000 posts that you originally collected). Haven't done in class yet

Report Details

Your report should be written entirely in paragraph-form, Arial 11pt font, 1-inch margins, double line spacing. It should have the following sections (the lengths are suggestions):

1. Overview (0.5 page) - Key findings Analysis of the two main points that the company expects us to find
2. Data (1 page) – describe your dataset. This should include statistics relevant to the project – the number of posts you originally started with, the number of Trump and Biden posts you had post filtering, and any design decisions you had to make around the filtering of this content.

Topics such as hot/new; Duplicates, Number of posts from a specific user (to balance out liberals and conservative)

3. Methods (1 page) - explanation and justification for what you did. Focus on the design decisions you made NOT listed in this document that impacted your results.
4. Results (2 pages) - share all your findings including the topics selected (and their definitions), topic characterization, and topic engagement. result of our manual data annotations and what did we conclude
5. Discussion (2 pages) - interpret your results in terms of what they reveal about the way each candidate was being discussed and perceived. Make extensive use of your results to justify your interpretations.

Will discuss when we have all the data

6. Group Member contributions (0.5 page) - a description of the contributions each group member made to this project.

7. References (< 1 page) - this is an optional section should you reference other works in your report.

Fine Print

- Each group will submit one report which will receive one grade that all members of the group will share. The one exception to this is in the case of strong evidence of delinquent group members. In this case, each member's grades may be adjusted up or down as appropriate. The blind team survey (described below) will be a key means of judging the contributions of individual group members.
- While there are no rules about how work should be divided up, good team participation and fair sharing the workload are absolutely expected parts of this project.
- The team survey will be completed AFTER the report due date by each member of the team. It involves an assessment of the quality of each other team member's contributions. This survey is intended ONLY as a means of identifying serious equity or team dynamic issues and underscoring to all students that this is being taken seriously from the outset. Our sincere hope and expectation is that nothing concerning will emerge from this survey (or from other communications) and everyone's grades can be based entirely on the report grade itself.

Evaluation Rubric

Criteria	**Points (100 in total)**	**Details**
Style	10	Is the text written in a clear, concise way? Is good grammar and spelling employed throughout?
Data collection correctness	10	Was the dataset prepared correctly? Did it have baseline characteristics that would allow this study to deliver meaningful insights?
Topic design validity	15	Was a process followed that would produce valid topics? Insufficient details should be treated the same as if something was not done.
Topic validity	15	Are the topics appropriate to the task? Are they well-defined? Are they defined to minimize subjectivity?
Annotation quality	10	Does the annotation process give us confidence in the quality of the annotations?
Results	20	Are all results requested present? Do the results make sense? Are outliers or unusual trends appropriately explained?
Findings	20	Are insightful candidate-level interpretations provided? Are

		these grounded in results? Do the findings integrate results and prior knowledge in a sound, well-reasoned way?