Cade Wiley and Konghao Zhao

CSC 673: Data Mining

Dr. Khuri

18 November 2023

Assignment 5: Final Project Proposal

Background

We found our data on ICPSR, a public data archive that stores data specific to social and political research. Our dataset comes from a study originally uploaded in 2017 named “The Attack on America and Civil Liberties Trade-Offs: A Three-Wave National Panel Survey, 2001-2004” which was authored by Darren Davis of Notre Dame and Brain Silver of Michigan State University. The dataset is composed of responses from individuals in the United States between 2001 and 2004 to questions concerning terrorism, public policy, and personal identity (Davis and Silver).

We are particularly interested in researching this dataset due to its relevance today with respect to the Patriot Act, islamophobia, and the United States’ involvement in international anti-terrorism policy. Our goal is to investigate how privilege influences someone’s willingness to support pro-security or pro-civil liberty policy.

Research Question

What is the relationship between privilege and one’s preference towards pro-civil liberty or pro-security policy after the 2001 attack on September 11th? In this context, privilege is defined as having a structural advantage due to unearned social identity such as belonging to a particular age group, gender, race/ethnicity, or socio-economic background.

Preparation

For our Project 5 submission, we have done some preliminary data processing and have inspected our data for potential challenges when conducting our research. After inspecting the

features of the dataset and its profile on ICPSR, we downloaded the dataset and used our previously developed code to uncover the qualities of our dataset.

First, we audited the dataset for duplicate columns and rows as well as rows and columns that had sufficient levels of null values. When removing columns with more than fifty one percent null values we retained 1031 samples and 257 rows. After filtering our data, we send our data to both PCA and UMAP separately. The PCA and UMAP output plots can be seen in our Jupyter Notebook file; however, due to attribute selection, they are likely to change.

Since there are lots of attributes in this dataset and we are looking to isolate the relationship between privilege and pro-security or pro-civil liberties policy, we need to be careful to preserve relevant data even if it is sparse, and remove irrelevant data. This requires us to revisit our attributes selection process as we will not be able to simply filter based on null values as in previous assignments. To make these decisions we looked at publications using our dataset to see what has already been done.

To classify different types of attributes, we follow the classifications in “Civil Liberties vs. Security: Public Opinion in the Context of the Terrorist Attacks on America” by Davis and Silver, the original collectors of the dataset. Attributes are broken down into three major classifications with subgroups, where the major classifications are core explanations, social, psychological, and political attributes, and demographic factors. Core explanations are attributes concerning how threatened an individual feels by terrorism, how much they trust their government, or how much terrorism has affected their financial well-being. Some examples of social, psychological, and political attributes are how much the subject trusts others, agrees with dogmatism, or how much they engage with national pride. Demographic attributes are those focused on race and ethnicity, education, age, and financial background (Davis and Silver).

Addressing Foreseen Challenges

Since we are interested in privilege we must focus on demographic features, particularly those attributes that are independent of choice. As stated in the research question, attributes

related to race/ethnicity, age, gender, and socio-economic will be critical to our success. One major challenge of our research will be dealing with missing values in key attributes such as race/ethnicity and how these attributes can be responsibly imputed if at all. We must also revisit our attributes before we filter for row sparsity as some valuable samples may currently be filtered out even though they are missing values for irrelevant attributes. Although we have played around with our data before this submission, we recognize that significant time will need to be spent engineering appropriate features for a successful research project.

Aside from the remaining work that needs to be done in feature selection, we are also researching how to augment tabular data containing human survey data. Currently, data augmentation is very popular in image processing, so has not widely been developed with tabular data in mind. This leaves us with two major challenges. One we must find a data augmentation method that seems reasonable and appropriate for tabular data. Two we must think critically about, and justify how we augment data that represents human survey responses. As a model we are going to research how data scientists have manipulated similarly sensitive data in datasets such as COMPASS.

The final major challenge we anticipate is finding an appropriate clustering algorithm. Relative to other datasets, we are using a small dataset where we are looking to uncover network-like relationships. This is leading us in the direction of graph-based clustering algorithms similar to PhenoGraph, but one that is not specific to single-cell RNA reads. In search of an appropriate clustering algorithm, we will continue researching graph-based algorithms and plan to meet with Dr. Berenhaut.

Work Cited

Davis, Darren W., and Brian D. Silver. “Civil Liberties vs. Security: Public Opinion in the Context of the Terrorist Attacks on America.” *American Journal of Political Science*, vol. 48, no. 1, 2004, pp. 28–46. *JSTOR*, https://doi.org/10.2307/1519895. Accessed 19 Nov. 2023.