Alex Murphy User Profile

Profile

Age: estimate late 20's to mid 30's (has a child and worked for a few years post undergrad)

Role: Data science master's student doing a thesis

Degree: M.Sc. in Computing (TU059) from Technological University Dublin

Primary Data Tools used: R and excel

Dissertation Title: "The Effects of Disinformation Upon National Attitudes Towards the EU and its Institutions" link (chapters 3 and 4)

Goals and Objectives of Thesis

Main Research Question:

The thesis aims to explore changes in the national attitudes of Germany, the UK, and Poland during a specific study period, given that these nations have been targeted by disinformation. This is inferred from the analysis of two datasets: one from Kaggle and a scraped disinformation dataset.

Objective:

The thesis seems to investigate the effects of disinformation on national attitudes towards the European Union (EU) and its institutions. It appears to focus on understanding the impact of disinformation campaigns and how they might influence public opinion in different countries.

Data used

EUvsDisinformation Dataset:

Kaggle EUvsDisinformation Data:

It contained over 7K observations with 37 columns, with 7.3K rows having 59K missing values

Scraped Disinformation Data (Scrape_Disinfo):

This dataset, referred to as Scrape_Disinfo, was used in conjunction with the Kaggle data.

The scraped data consisted of 66K rows with 4 columns. The date ranges of the scraped data were not initially aligned with the Kaggle data, requiring adjustments to align the datasets.

Eurobarometer Data:

Eurobarometer data was also used, loaded from EB83-2015 to EB97-2022. The data was loaded as a 'Large List' by the openxlsx package in R, and it was accessed through the list index. The data was not initially in dataframe format and was listed as separate years.

Other Data:

Other datasets were considered but eventually eliminated due to not meeting specific criteria. For example, a Twitter-based database in French and English and a text-based fake news dataset were within the scope of the research but did not satisfy the criteria.

Issues regarding Data Prep:

Handling Missing Values:

The Kaggle dataset contained a significant number of blank cells. Within the 7.3K rows, there were 59K missing values and cells marked as "['None']". These were converted to 'NA' during the import process

Data Wrangling:

The Eurobarometer survey data required extensive wrangling and preparation. Each added question created a significant amount of work. Challenges included cleaning data effectively over thousands of pages contained in long lists that were not easily accessible in R. There were also issues with French text in the data that was incorrectly encoded in RStudio.

For question D78 in the Eurobarometer data, there were many columns with different changes that required a careful approach to ensure data fidelity

Data Separation and Extraction:

Data was initially atomic, meaning any row that could have a separated value was separated. However, it was later determined that a lot of the data would not be used, so unnecessary separations were avoided to prevent the creation of datasets with hundreds of thousands of rows.

Data Conversion and Formatting:

There were challenges related to converting and formatting data. For example, dates needed to be correctly stored as date values, and factors were applied where necessary. There were also challenges related to handling large text data and comma-separated lists in certain columns.

A function was written to search for specific string values, remove rows, set new column headings, convert values to numeric, and reset the index for new DataFrames. Challenges were encountered due to the inner workings of R, especially when searching for string values and handling NA values

Data Combination:

Combining column titles using strings was necessary to maintain data across different years. This process required careful attention to detail to ensure the accuracy and integrity of the data