Final Project Proposal

Team: Forhad Akbar, Adam Douglas, and Soumya Ghosh

Introduction

The old saying goes: "*How can you tell a politician is lying? His lips are moving*". Although this humorous statement is a bit of an exaggeration, there is a seed of truth to it when it comes to politics. One needs to look no further than the website www.PolitiFact.com to see how difficult it is to take at face-value what we read and hear. The spread of dis- and mis-information has only become more of an issue in our society over the past several years.

For these reasons, we are exploring a dataset of political statements and the associated fact-checking, as done by Politifact[1].

Data Source

Data was scraped from the PolitiFact website[1] and was made available on the Kaggle website[2]. The dataset contains the source and text of statements made regarding a variety of subjects important to politics and public policy. It also contains the fact-checkers' score rating the statements' veracity and the text the reviewer posted explaining their scoring.

For this project, we will analyze the text of both the original statements and the fact-checkers' research to identify the major theme(s) for each record (e.g. coronavirus, economics, etc.). Then, we will perform network analysis on the sources and subjects to look for patterns of false information (e.g. person A most often posts false information about subjects X and Y) and accurate information (e.g. person B is more trustworthy than all others when posting about subject Z).

We also fully expect that interesting patterns will emerge from both the text analysis and the network analysis, and will explore them as time (and the data) allows (e.g. does person A's false information with respect to subject B change over time?).

Data Concerns

The largest concern is that many records may fit into more than a single topic, and the outcome of the overall fact-check may be partially true, but have different outcomes when considering each subject in isolation. For example, the statement "The 2020 election was stolen, but at least the vaccines are beginning to be distributed" may be labeled as mostly false overall, but when viewed through the lens of two different topic categories (coronavirus and elections) it is a bit of a dichotomy.

1 - https://www.politifact.com
2 - https://www.kaggle.com/shivkumarganesh/politifact-factcheck-data

Work Plan

The work plan, in general terms, looks like this:

Task / Step	Primary Team Member
Data loading and cleaning	Adam
Text Analysis (categorization)	Forhad
Network Analysis (relationships)	Adam
Final Write-Up and Validation	Soumya
Presentation	TBD