X Data as a Predictor of Political and Ideological Alignment

Eloragh Espie
eae2273
eloraghespie@utexas.edu

Rylan Vachon
rmv764
rylanvachon@utexas.edu

1 Introduction

In the digital age, social media platforms serve as crucial arenas for the expression and dissemination of political opinions. The ease of use and popularization of web scraping techniques have provided researchers the ability to analyze the immense amount of data generated by online interactions, offering insights into the distribution and dynamics of political ideologies. This paper delves into the utilization of web scraping methods to collect data from the social media platform X, with the aim of plotting individuals onto a political compass.

The political compass, a visual framework separating political ideologies along two axes—authoritarian vs. libertarian social ideals and left-wing vs. right-wing economic ideals—provides a more nuanced understanding of political orientations compared to a simple left vs. right scale. Using Python's BeautifulSoup and Selenium libraries, text data is extracted from the online platform. The extracted text is used to train two separate logistic regression models to classify political sentiments and opinions.

This research seeks to contribute to the understanding of the contemporary political landscape by offering an analysis of individuals' political stances through their use of social media platforms. Through keyword extraction, individuals are assigned an x and y value within the four quadrants of the political compass, allowing a granular examination of their ideological leanings.

The significance of this study lies in its potential to further exemplify and examine the complexities of online political engagement, shedding light on the prevalence and polarization of ideological perspectives in online spaces. By visualizing the distribution of individuals across the political compass, this project aims to show the dynamics of political discourse in the digital sphere, offering valuable insights for policymakers, scholars, and practitioners alike.

2 Data

The data collection process for this study involves the use of web scraping techniques to gather text data from X. The individuals we will scrape data from will come from a pool of politicians that we have political compass coordinates for.

The dataset is curated to include individuals for whom political compass coordinates (x and y values) are available, thereby enabling the mapping of their ideological positions onto the political compass. The Official Political Compass maintains an archive of their official Political Compass values for major politicians or political groups in recent elections across the world.

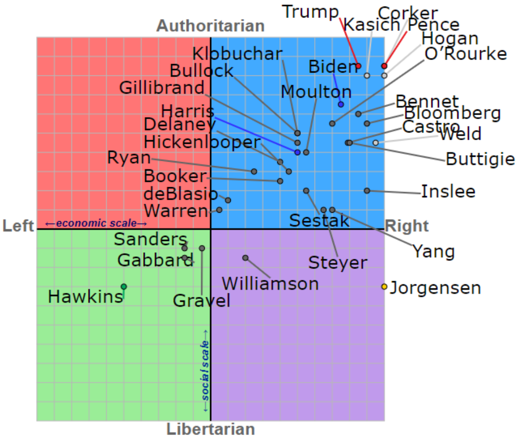

Figure 1: Political Compass's Alignments for early candidates in the 2020 American General Election.

The Political Compass has 36 unique elections with data points from the United States, United Kingdom, Latin America, Italy, France, Ireland, Australia, and New Zealand. We plan to scrape as much data as possible from the individuals that can be attributed to these values since their x and y coordinates are coming from the ”official” source.

There will be some issues that we will run into when attempting to scrape this data. Firstly, many countries in the list above do not align with the same election practices that the United States does. The UK, for example, elects a party into power and then a representative from the party is chosen to be the Prime Minister. Therefore, no one individual is ever really elected, and this is reflected in the political compass' of the UK elections, as shown in Figure 2.

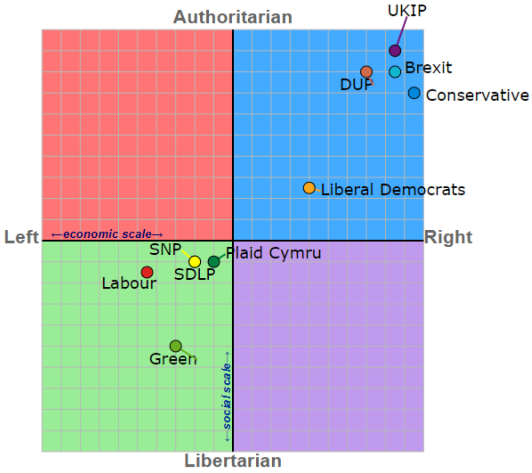

Figure 2: Political Compass's alignments for political parties in the UK's 2019 General Election.

To make use of such data, we would have to be able to accurately identify who would have been chosen as Prime Minister from each party had they won the election. We are unsure if such data can be determined within an acceptable range of confidence.

The second problem arises from the variety of linguistic data we would encounter if we attempted to use data from various international politicians. For several of these countries, such as Italy and France, the national language is not English, and there is a good chance their X posts, if they exist, would not be in English. We endeavor to keep our data as valid and unbiased as possible for this project. To attempt to translate linguistic data from X, a social media company known to inspire new slang and linguistic variations to align with its character limit, would be against our efforts to remain objective. We will not be attempting to use any data not originally in English.

We will use both Selenium and BeautifulSoup to scrape dynamic content, enabling access to X's interactive elements and ensuring comprehensive data retrieval. This combination will allow us to efficiently navigate through the online platform, extract relevant text content, and compile a robust dataset for subsequent analysis with ease.

The dataset assembled through web scraping efforts forms the foundation for our analysis, facilitating the mapping of politicians onto the political compass based on their previously expressed opinions and ideological orientations. This comprehensive dataset enables us to explore the distribution and dynamics of political ideologies across a diverse spectrum of political figures, offering valuable insights into the online political landscape.

2.1 The Value of X's Data

X, formerly known as Twitter, has long been an invaluable resource for a myriad of natural language processing (NLP) tasks. The platform offers real-time insights, enabling immediate access to unfolding political discourse and sentiments. This real-time feature facilitates the analysis of current events, trends, and public reactions as they manifest, providing invaluable insights into the ever-changing political discourse occurring in online spaces.

Additionally, X boasts a broad corpus of text data, reflecting a broad spectrum of political opinions, ideologies, and linguistic nuances. With millions of users globally, the platform generates a new, rich, and varied linguistic dataset every single day. This diversity enhances the robustness of analyses, allowing for comprehensive examinations of political discourse across different regions, languages, and demographic groups from only one source.

Additionally, X's unique character limit incentivizes users to convey their thoughts and opinions concisely within the confines of a single tweet. This succinct communication style, characterized by brevity and informality, lends itself well to NLP analysis such as sentiment analysis and topic classification. The compact nature of tweets facilitates efficient processing and analysis, allowing us to derive meaningful insights from thousands of records of data in a computationally efficient manner.

These characteristics underscore the significance of X data in NLP classification efforts within the realm of political analysis. By leveraging the wealth of text data available on the platform, we can gain deeper insights into public opinion, ideological trends, and the intricate dynamics of political communication in real time.

3 Methodology

Our methodology involves data collection from online platforms, text data preprocessing, feature extraction, and logistic regression model training to predict political compass coordinates for individuals. We evaluate model performance, map predicted coordinates onto the political compass, and conduct validation and sensitivity analyses.

3.1 Data Collection

The data collection process involves scraping text data from X. We utilize Python's BeautifulSoup library in conjunction with Selenium for dynamic content scraping to extract text data. The dataset is curated to include politicians for whom political compass coordinates (x and y values) are available, enabling the mapping of their ideological positions and providing training and testing data for the models.

3.2 Preprocessing

The extracted text data undergoes preprocessing steps to enhance its quality and suitability for analysis. This includes tokenization, lowercasing, punctuation removal, and stop word removal to standardize the text and reduce noise. Additional measures such as stemming or lemmatization may be applied to further normalize the text if found to be necessary or beneficial.

3.3 Feature Extraction

For the text data scraped from each politician's X account, we will extract features from the preprocessed text to represent their political opinions and sentiments. These features may include word frequency counts, n-grams, sentiment scores, and other relevant linguistic features. The aim of the feature extraction process is to capture the linguistic patterns that best indicate political ideologies.

3.4 Model Training

We employ two distinct logistic regression models to predict the political compass coordinates (x and y values) for each politician. The first logistic regression model is trained to predict the x coordinate, representing the authoritarianism vs. libertarianism axis of the political compass. The second logistic regression model is trained to predict the y coordinate, representing the left-wing vs. right-wing axis.

3.5 Evaluating the model

The performance of each logistic regression model is evaluated using appropriate metrics such as accuracy, precision, recall, and F1-score. We employ techniques such as cross-validation to ensure the robustness and generalizability of the models. Additionally, we conduct exploratory data analysis to gain insights into the distribution of predicted coordinates and their alignment with known political ideologies.

3.6 Predicting

Ideally, once trained and evaluated, the logistic regression models will be able to predict the political compass coordinates for any X handle. The predicted coordinates are then plotted on the political compass, providing a visual representation of their forecasted ideological positions. This mapping enables the classification of individuals along the authoritarianism vs. libertarianism and left-wing vs. right-wing axes, facilitating the analysis of ideological trends and distributions.

As a broader goal, we aim to create a simple front-end system that would enable users to enter their X handle and allow the logistic regression models to predict their Political Compass values. Depending on the accuracy of our models' end states, this interface would give users insight into which ideologies their text data seems to align them with.

3.7 Ethical Considerations

Throughout this methodology, ethical considerations are paramount. They pose a relatively difficult challenge for this specific project as we aim to plot unique individuals on a Political Compass. To ensure data privacy and anonymity, we are aggregating and anonymizing the text data. No individuals used to train or test the logistic regression models will be named or identified by online handles, aliases, or any other form of identification. Additionally, we adhere to ethical guidelines in handling sensitive political topics and opinions, maintaining neutrality and objectivity in our analysis and interpretation.

4 Team Structure

4.1 Eloragh Espie: Data Acquisition and Processing

Eloragh will be responsible for acquiring and preprocessing textual data from online platforms where politicians engage in discourse. She will employ web scraping techniques to gather data and perform preprocessing tasks such as cleaning and functionalization. By extracting linguistic features from the data, she will ensure that it is ready for analysis.

4.2 Rylan Vachon: Model Training and Evaluation

Rylan will be focused on training the logistic regression models to predict political compass coordinates based on the featurized data provided by Eloragh. He will evaluate model performance and be responsible for identifying any issues and iteratively improving the models as needed. Additionally, Rylan will test the trained models on separate datasets to assess their effectiveness, ensuring the quality of the final analysis.

4.3 Collaborative Effort

Ultimately, this project will be a collaborative effort and depends on the unique abilities of both parties. Both team members will work closely together throughout the project, sharing insights and findings. Eloragh's primary responsibility will be to provide clean and featurized data to Rylan for model training. Rylan's main objective will be to offer feedback and request input on model performance and potential improvements over iterations. Both partners will be responsible for the end result and accuracy of the models. This collaborative approach ensures the successful completion of the project objectives as well as a valuable opportunity to build off of each other's strengths and prior education.