**Constitutional Analysis for Regime Classification: An NLP Approach**

Chasen Jeffries

Abstract:

With the rise of competitive authoritarianism, accurately distinguishing between autocratic and democratic regimes has become increasingly challenging. This exploratory analysis investigates the use of Natural Language Processing (NLP) to predict regime classification based on national constitutions. We examine the potential of NLP in predicting regime types by constructing a corpus of 10 national constitutions and an NLP pipeline, which cleans and prepares the corpus for analysis. After preprocessing and feature extraction, three classification models (logistic regression, neural network classifier, and random forest) are trained and tested on the extracted features. The results demonstrate that even simple NLP techniques applied to constitutional text can accurately predict regime types. The findings highlight the efficacy of NLP in political science research and open avenues for future investigations on the use of NLP tools in the field. Researchers can further enhance our understanding of regime classification and political systems by extending this analysis to larger corpora and exploring more complex NLP features.

1. **Introduction**:

In recent decades, a significant number of nations have witnessed a shift from traditional forms of authoritarianism to competitive authoritarianism. Consequently, there is a growing need to accurately identify national regimes, particularly in distinguishing between democracies and competitive authoritarian governments. This study aims to contribute to the existing literature on regime classification by exploring the potential of natural language processing (NLP) in predicting regime types. Specifically, we seek to determine if NLP can effectively distinguish between democratic and competitive authoritarian regimes by analyzing the language used in their respective national constitutions. By examining the linguistic characteristics of these constitutions, we aim to identify unique elements associated with each regime type and assess the feasibility of predicting regime types based on constitutional texts. The findings from this research could offer valuable insights into predicting future regime types when nations undertake constitutional revisions, thereby enhancing our understanding of political transitions.

National constitutions serve as the foundational charters that guide a nation's governance. Given the continuous emergence and transformation of nations, the ability to predict regime types based on constitutional provisions would be highly valuable for policymakers and researchers. In this study, we aim to develop a comprehensive corpus of national constitutions, initially starting with a small sample for an exploratory analysis. By establishing a successful natural language processing (NLP) pipeline connected to this corpus, we can leverage linguistic elements to predict regime attributes. This research represents an initial exploratory analysis of the capabilities of NLP in predicting these attributes and aims to determine the effectiveness of using NLP for the analysis of national constitutions. By examining lower-level NLP attributes such as word counts, sentence structure, and entity identification, we can identify trends and potential avenues for further research in this domain.

1. **Literature**

In the aftermath of the Cold War, scholars encountered difficulties in accurately classifying a category of nations commonly referred to as "semi-democracies." These nations exhibited characteristics that seemed to indicate a transition towards either democracy or autocracy.[[1]](#footnote-1) It was through the groundbreaking work of Levitsky and Way that our understanding of these regimes underwent a profound transformation.[[2]](#footnote-2) They introduced the concept of "competitive authoritarianism," identifying these nations not as transitional governments, but rather as a distinct form of governance known as competitive authoritarian governments. These authoritarian regimes employ various democratic mechanisms, such as holding elections that result in the election of officials to legislative bodies with the power to enact laws. However, they lack one or more essential democratic components, including genuinely free and fair elections, near-universal suffrage, robust civil liberties, and a government that is accountable and responsive to the will of the people.

In recent years, the availability of advanced software and hardware has facilitated the widespread use of Natural Language Processing (NLP) models in computational analysis. [[3]](#footnote-3) While NLP has been commonly applied for sentiment analysis of new media artifacts, its potential in political science remains largely untapped. [[4]](#footnote-4) This study aims to highlight the value of NLP analysis in political research, particularly when applied to political documents. By leveraging NLP techniques, we can uncover valuable insights and deepen our understanding of political dynamics through systematic examination of textual data. This exploration opens up new avenues for research, offering unique opportunities to gain deeper insights into political discourse and contribute to the advancement of the field.

1. **Data and Methods**

We create a corpus composed of the constitutions of ten nations. Each constitution was translated into english. We get our data from the comparative constitutions project (CCP), a project that collects national constitutions to produce data to better understand constitution making and outcomes.[[5]](#footnote-5) This will be split into two groups of five democracies (D) and five competitive authoritarian (CA) governments. We have choses nations based on their relatively agreed status as either a democracy or competitive authoritarian government. Additionally, we selected a representative sample of nations from many regions of the world, to limit the effect of a regional bias. We chose the following nations:

**CA**: Turkey, China, Russia, Egypt, Venezuela

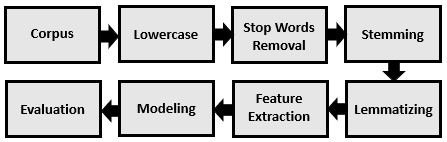
**D**: US, France, Germany, Australia, Japan

Each constitution consists of tens of thousands of words, thousands of sentences, and hundreds of paragraphs. We examine these documents at the document level, which creates a relatively small n research (n=10), but each document has rich amounts of data to extract.

Preprocessing:

This artifact of data must be translated into text data and cleaned before features can be successfully extracted to input into our models. We used Pythons NLTK tools to create an NLP pipeline of cleaned corpus of constitutions for feature extraction. Our NLP pipeline included the following steps as illustrated by Table 1.

**Table 1:**

 The first step in our NLP pipeline is lower case, which takes all characters to lowercase. This prevents the same word from being classified as two different words based on capitalization. The second step removes stop words, we use the NLTK stop words list, we add special characters and punctuation to the list, to remove these elements from our corpus. Finally, we stem and lemmatize the words to take all versions of a word root and stem to the single token. This makes provides us with a set of cleaned documents ready for feature extraction.

Feature Extraction:

We extract the two sets of features from our documents: simple counts and name entity recognition (NER) counts. The simple counts consist of variables which count the number of characters, words, sentences, and unique words that are in each document.[[6]](#footnote-6) We choose to use relative frequency of the count variables by calculating the values as a percentage of the number of characters. For the characters variable, we calculate the number of characters over 100,000.[[7]](#footnote-7) The second set of features we extract are counts of four categories of NER (using the spacy package). The four categories we choose are NORP (nationalities, religious, or political groups), Org (companies, agencies, institutions, etc), Person (people), and GPE (Countries, cities, states). This allows us an additional set of relatively simple variables to examine on constitutions. For each of the NER counts extracted we calculate their relative frequency to total NERs identified in each document. Finally, we create a simple binary dependent variable to identify democracies (D = 1) and competitive authoritarian (CA = 1) governments.

1. **Exploratory Analysis**

For our exploratory analysis, we examine the ability of three different models to correctly classify national regime based on their constitution’s features. We choose to test a logistic regression, artificial neural-network (ANN), and a random forest. Before inputting data into the model, we first split the data into training and testing data. We tested several different ratios including 60-40, 70-30, and 80-20. We settled on 70-30 to maximize the amount of training data, while still allowing for more than 2 tests. After splitting the data, we input the training data into each of the three models before testing the predictive ability on the test data. We obtained the following results, illustrated by Tables 2 and 3. We evaluate the performance of our models by looking into the precision, recall, and accuracy of the model. This allows us multiple angles to investigate the performance of our predictive model. The logistic regression had an accuracy score of 0.66 and a recall and precision score of 0.0. The neural network (NN) classifier had an accuracy score of 0.66, a recall score of 1.0, and a precision score of 0.5. Finally, the random forest performed best with a score of 1.0 for accuracy, recall, and precision.[[8]](#footnote-8)

1. **Discussion**

The training of this model was severely limited by the small sample size (n=10) meaning we had a 7-3 ratio of documents on training and testing respectively. While this is acceptable for an exploratory analysis, a more in-depth explanatory or predictive analysis should have a significantly larger sample size. This limited sample size is best illustrated by the precision and recall scores of zero and one for the models, a larger sample would be unlikely to result in such round scores.

The random forest model performed best at classifying the constitutions with perfect scores for all three-evaluation metrics. However, we highlight that all three models could predict the regime type of nations based on the NL of their constitutions.

1. **Conclusion**

This exploratory analysis investigated the efficacy of using NLP to predict regime type based on national constitutions. Our results demonstrate that, even when using low-level NLP features, NL elements of national constitutions can be used to predict regime type. Future research should build upon this analysis by developing a larger corpus, investigating more complex NLP features (ex. relationship extraction), and the creation of a gold-standard for NER.

1. **References**

Benoit, Kenneth, Kevin Munger, and Arthur Spirling. “Measuring and Explaining Political Sophistication through Textual Complexity.” American Journal of Political Science 63, no. 2 (2019): 491–508.

Constitute. Accessed May 31, 2023. https://constituteproject.org/.

“Informing Constitutional Design.” Comparative Constitutions Project, March 1, 2023. https://comparativeconstitutionsproject.org/.

Fareed Zakaria, “The Rise of Illiberal Democracy,” Foreign Affairs 76 (November– December 1997): 22–41

Gordon P. Means, “Soft Authoritarianism in Malaysia and Singapore,” Journal of Democracy 7 (October 1996): 103–17.

Levitsky, Steven, and Lucan A. Way. *Competitive authoritarianism: Hybrid regimes after the Cold War*. Cambridge: Cambridge University Press, 2013.

Levitsky, Steven, and Lucan Way. “Elections Without Democracy: The Rise of Competitive Authoritarianism”. Journal of Democracy 13, no. 2 (April 2002): 51-65.

Levitsky, Steven, and Lucan Way. “The New Competitive Authoritarianism”. Journal of Democracy 31, no. 1 (January 2020): 51-65.

Németh, R. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *J Comput Soc Sc* **6**, 289–313 (2023).

Terry Lynn Karl, “The Hybrid Regimes of Central America,” Journal of Democracy 6 (July 1995): 72–87

Wiedemann, Gregor. “Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences.” Historical Social Research / Historische Sozialforschung 38, no. 4 (146) (2013): 332–57.

1. **Appendix**

Table 2:

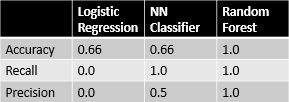
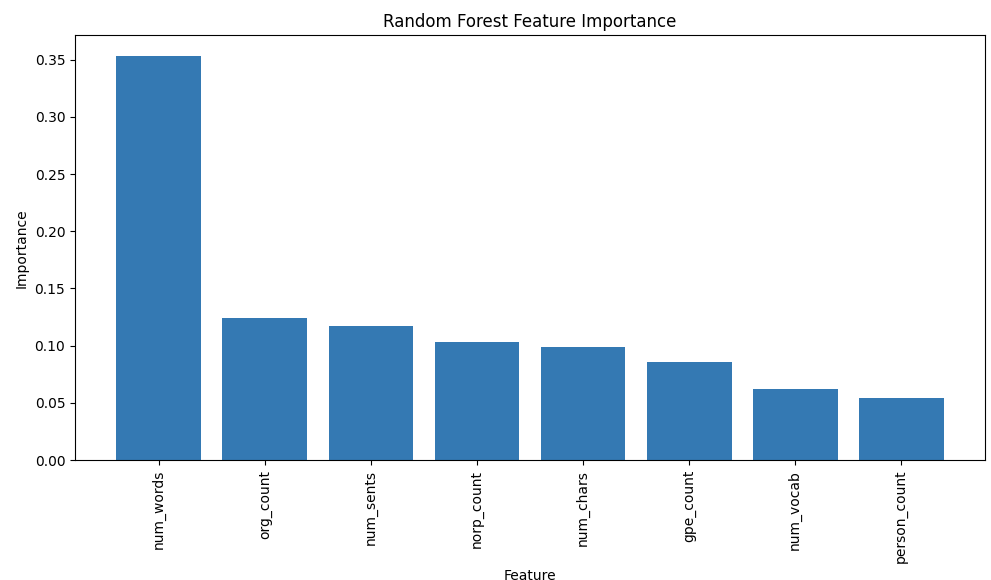


Table 3:



1. Terry Lynn Karl, “The Hybrid Regimes of Central America,” Journal of Democracy 6 (July 1995): 72–87; Fareed Zakaria, “The Rise of Illiberal Democracy,” Foreign Affairs 76 (November– December 1997): 22–41; Gordon P. Means, “Soft Authoritarianism in Malaysia and Singapore,” Journal of Democracy 7 (October 1996): 103–17. [↑](#footnote-ref-1)
2. Levitsky, Steven, and Lucan Way. “Elections Without Democracy: The Rise of Competitive Authoritarianism”. *Journal of Democracy* 13, no. 2 (April 2002): 51-65. [↑](#footnote-ref-2)
3. Wiedemann, Gregor. “Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences.” Historical Social Research / Historische Sozialforschung 38, no. 4 (146) (2013): 332–57. [↑](#footnote-ref-3)
4. Németh, R. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *J Comput Soc Sc* **6**, 289–313 (2023). [↑](#footnote-ref-4)
5. Constitute. Accessed May 31, 2023. https://constituteproject.org/. [↑](#footnote-ref-5)
6. Note that that all constitutions were translated into english which represents a possible source of bias.

   The unique words variable differs from our words variable by counting the number of unique words in our document. A higher score will indicate a higher lexical diversity in our document. [↑](#footnote-ref-6)
7. We choose the value 100,000 as a near mean value that would limit the magnitude impact of the characters variable. [↑](#footnote-ref-7)
8. Reviewing the variable importance for the random forest model highlights that the number of words variable was by far the most important variable in the NLP analysis. The next group of important variables were: Org counts, number of sentences, NORP counts, number of characters, and GPE counts. [↑](#footnote-ref-8)