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.

I. **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.

II. **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] It was through the groundbreaking work of Levitsky and Way that our understanding of these regimes underwent a profound transformation.[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

---

[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.

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

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] While NLP has been commonly applied for sentiment analysis of new media artifacts, its potential in political science remains largely untapped. [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.

III. **Data and Methods**

We constructed a corpus consisting of the translated versions of ten national constitutions. These translations were obtained from the Comparative Constitutions Project (CCP), an initiative aimed at collecting national constitutions for the purpose of studying constitution-making processes and their outcomes. The corpus is divided into two groups: ten democracies (D) and ten competitive authoritarian (CA) governments. The selection of nations was based on their widely recognized classification as either a democracy or a competitive authoritarian government. Moreover, we ensured representation from diverse regions around the world to minimize the influence of regional biases. To see the nations included in our study see Table 1.

Each constitution in our corpus comprises tens of thousands of words, thousands of sentences, and hundreds of paragraphs. Our analysis focuses on the document level, which results in a relatively small sample size of 20 documents. Despite the limited number of documents, the richness of the data within each constitution provides ample material for in-depth analysis and the extraction of meaningful insights.

Preprocessing:

To transform the artifact of data into usable text data, it is necessary to translate and clean the content before extracting relevant features for our models. We employed Python's NLTK tools to develop an NLP pipeline for cleansing the corpus of constitutions and extracting features. Our NLP pipeline consisted of several steps, as outlined in Table 2.

The initial step in our NLP pipeline involves converting all characters to lowercase to ensure consistency and prevent the classification of the same word as multiple entities based on capitalization. Subsequently, we eliminate stop words using the NLTK stop words list, which includes additional removal of special characters and punctuation. Lastly, we employ stemming and lemmatization techniques to reduce words to their root forms, resulting in a collection of

[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.

[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).

cleaned documents suitable for feature extraction. This process produces a set of cleaned documents that are now ready for feature extraction.

Feature Extraction:

We extract two sets of features from our documents: simple counts and name entity recognition (NER) counts. The simple counts consist of variables that count the number of characters, words, sentences, and unique words in each document.[5] To standardize these counts, we calculate the relative frequency of each variable by expressing it as a percentage of the total number of characters.[6] The second set of features we extract is the counts of four categories of NER using the spacy package. These categories include NORP (nationalities, religious, or political groups), Org (companies, agencies, institutions, etc.), Person (individuals), and GPE (countries, cities, states). These NER counts provide additional variables to examine within the constitutions. Similar to the simple counts, we calculate the relative frequency of each NER category by dividing the count by the total number of NERs identified in each document. Lastly, we create a binary dependent variable to identify democracies (D = 1) and competitive authoritarian (CA = 1) governments.

IV. Exploratory Analysis

For our exploratory analysis, we assessed the performance of three different models in accurately classifying national regimes based on their constitution's features. The models we tested were logistic regression, artificial neural network (ANN), and random forest. Prior to model input, we divided the data into training and testing datasets. We experimented with various ratios, including 60-40, 70-30, and 80-20, and ultimately settled on a 70-30 split to maximize training data while allowing for meaningful testing. After splitting the data, we trained each of the three models using the training dataset and evaluated their predictive performance on the test dataset. The results are summarized in Tables 3 and 4. To assess model performance, we examined precision, recall, and accuracy metrics, providing multiple perspectives on the model's effectiveness. The logistic regression model achieved an accuracy score of 0.66, but it had a recall and precision score of 0.0. The ANN classifier achieved an accuracy score of 0.66, a recall score of 1.0, and a precision score of 0.5. Finally, the random forest model performed the best, with perfect scores of 1.0 for accuracy, recall, and precision.[7]

V. Discussion

The training of our models was indeed constrained by the small sample size of only 10 documents. This limited dataset necessitated a 7-3 ratio for training and testing, respectively.

---

[5] 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.

[6] For the characters variable, we calculate the number of characters over 100,000. We choose the value 100,000 as a near mean value that would limit the magnitude impact of the characters variable.

[7] 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.

While this ratio is acceptable for an exploratory analysis, it is important to note that a more comprehensive explanatory or predictive analysis would require a substantially larger sample size. The impact of the small sample size is evident in the precision and recall scores, which resulted in values of zero or one for the models. With a larger sample size, we would expect to see more nuanced and varied scores, reflecting a more accurate representation of the population. The round scores obtained in our analysis are indicative of the limitations imposed by the limited number of documents.

The random forest model demonstrated superior performance in classifying the constitutions, achieving perfect scores for all three-evaluation metrics. This indicates its strong ability to predict the regime type of nations based on the natural language (NL) of their constitutions. However, it is important to note that all three models exhibited promising predictive capabilities in this study. Despite the limitations imposed by the small sample size, they were able to successfully classify the regime type based on the NL features of the constitutions. The results of this exploratory analysis highlight the promising nature of NLP techniques for regime classification based on constitutional texts.

VI. Conclusion

In this exploratory analysis, we aimed to assess the effectiveness of utilizing natural language processing (NLP) to predict regime type based on national constitutions. Our findings indicate that even with the utilization of basic NLP features, such as simple counts and named entity recognition (NER), NL elements extracted from national constitutions can successfully predict regime type.

To advance this research, future studies should first focus on expanding the corpus of national constitutions to enhance representativeness. This would enable more robust investigations into the predictive power of NLP models. Second, exploring advanced NLP features, like relationship extraction, can provide deeper insights into the dynamics within constitutional texts, contributing to a nuanced understanding of regime types. Finally, establishing a gold standard for named entity recognition would enhance the accuracy of NER counts, ensuring consistency across studies. Expanding the corpus, exploring advanced features, and establishing a gold standard are crucial steps to advance NLP-based regime type prediction.

Overall, by expanding the corpus size, exploring advanced NLP techniques, and establishing a gold standard for NER, future research can build upon the foundation laid by this exploratory analysis and contribute to a more comprehensive understanding of the potential of NLP in predicting regime type based on national constitutions.

VII. 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.

VIII. Appendix

Table 1:

CA:	D:
Turkey	United States
China	France
Russia	Germany
Egypt	Australia
Venezuela	Japan
Belarus	Canada
Cambodia	Chile
Iran	Portugal
Kazakhstan	South Korea
Nicaragua	Sweden

Table 2:

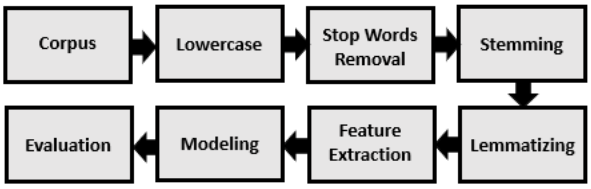
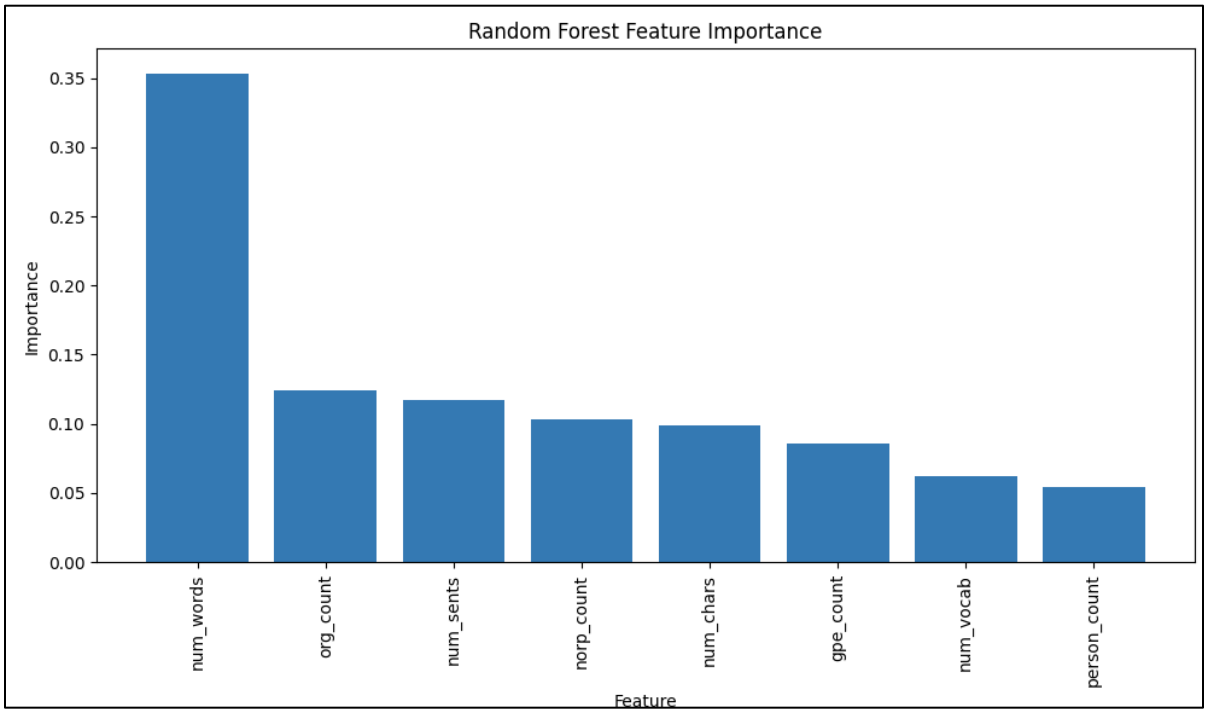


Table 3:

	Logistic Regression	NN Classifier	Random Forest
Accuracy	0.66	0.66	1.0
Recall	0.0	1.0	1.0
Precision	0.0	0.5	1.0

Table 4:


Random Forest Feature Importance
Importance
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
num_words
org_count
num_sents
norp_count
num_chars
gpe_count
num_vocab
person_count
Feature