Democracy Classification: Analysis and Prediction from the United Nations General Debate Corpus 1946 - 2024

Joao Barbara[1], Ron Lakeman[1], Marco Maier[1], and Chang Zhou[1]

[1] Universiteit van Amsterdam

Abstract. This study examines UN General Debate speeches from 1946 to 2024 using Natural Language Processing (NLP) methods to investigate their connection to democratic values, including government accountability as measured by the Worldwide Governance Indicators. It employs NLP to identify patterns that classify countries as democratic, autocratic, or transitional, in line with which argues that states rarely fall into purely democratic or autocratic categories. The methodology involves text preprocessing, exploratory analysis to find mentions related to democracy, and SVM, SGD, Logistic Regression, and KNN machine learning models to assess predictability. The questions of how different types of government are mentioned in UN speeches over this period and, if and how it might be possible to predict a country's democracy class from the speeches held at the yearly UN General Debate. The exploratory analysis uncovers significant patterns. The classification using the aforementioned machine learning models demonstrate prediction accuracy of a maximum of 88%.

Keywords: Machine Learning · Natural Language Processing · Classification · Democracy · Rule of Law · Sustainable Development Goals · UN General Debate Corpus

1 Introduction

1.1 Research context

Democracy is about much more than fair elections or how a country is governed [3]. Findings from an extensive survey state that its true importance lies in empowering people by giving them a voice and ensuring accountability which assures a sense of public safety and security. The UN's 2030 agenda for Sustainable Development recognizes this connection. Although SDG 16 [14] ("Peace, justice and strong institutions") does not explicitly mention democracy, it emphasizes inclusive decision-making and having accountable institutions (SDG 16.6 and 16.7) which closely aligns with the democratic principles from [3].

Now that democracy currently is under growing global threat [12], understanding the functioning of different government systems is vital to advancing SDG 16. Over the past three decades, several major datasets were created such as Polity IV, Freedom House, and Varieties of Democracy with the aim of measuring and scaling democracy [4].

This research focuses on the Polity IV dataset (see Section 2) but takes a different approach by analyzing UN General Debate speeches (1946–2024) to explore their relation to democratic values, including government accountability for the rule of law from the Worldwide Governance Indicators. Following [15], the study applies natural language processing (NLP) to detect patterns that may classify countries as democratic, autocratic, or transitional, in line with [4], who argue that states are rarely purely democratic or autocratic.

The experimental setup includes text preprocessing, exploratory analysis to identify democracy-related mentions and patterns, and the application of machine learning models to test whether democracy classes can be reliably predicted. This leads to the following research questions:

1. How are different regimes and governance reflected in the UN speeches for countries from 1946 to 2024?
2. Can we predict democracy classes based on UN speeches for countries from 1946 to 2024?

To answer these questions, an exploratory analysis is carried out to find meaningful patterns and identify democracy classes. Building on these insights, several machine learning models were applied, achieving strong predictive performance with accuracies approaching 90%.

2 Data

2.1 Dataset Description

United Nations General Debate Corpus
The United Nations General Debate Corpus (1946–2024), available on Harvard Dataverse [10], contains transcripts of annual speeches delivered at the UN General Assembly. It provides a valuable resource for examining international relations, diplomacy, and global policy trends across nearly eight decades.

Democracy Index by Polity5
The Polity5 dataset, published by the Center for Systemic Peace [11] within the INSCR project, covers political regime characteristics and transitions from 1800 to 2018 across 167 countries with populations above 500,000. It is widely used in political science and international relations for analyzing regime stability and change. The index assigns scores from –10 (autocracy) to +10 (democracy), with intermediate values representing anocracies.

Rule of Law Index by World Bank
The Worldwide Governance Indicators (WGI) dataset by the World Bank [2] covers over 200 economies from 1996 onwards. Among its dimensions, the Rule of Law index which is based on surveys and expert assessments measures confidence in adherence to legal and social rules. It ranges from –2.5 (lowest) to +2.5 (highest) and is widely used in research on governance, institutional quality, and development.

2.2 Data Cleaning and Preparation

To ease the analysis between the Democracy Index dataset and the UN speeches, the numeric index (–10 to 10) was converted into discrete classifications. According to the authors [13], this consists of three regime categories: **Autocracies** (–10 to –6), **Anocracies** or hybrid regimes (–6 to +6), and **Democracies** (6 to 10).

After discretizing the Democracy Index into Democracy, Autocracy, and Transition, we re-examined the dataset and found the following observations (Figure 1). Of the 74 countries classified in 2018 (the latest publication), the majority were democratic, while 16% were autocracies and 27% anocracies. The number of democracies increased steadily and surpassed autocracies after the fall of communism in the 1990s. However, since 2010, the share of democracies has declined, accompanied by a rise in hybrid and autocratic regimes.

The second dataset, the *Worldwide Governance Indicators: Rule of Law Estimate*, was used to establish a standardized relation between regime types and one of the focal points of SDG-16, the promotion of Rule of Law (16.3) [14].

Figure 2 shows the evolution of the index quartiles across all countries. This dataset covers only the last 20 years of the Democracy Index, and the small variation in quartile values may reflect slower regime shifts in recent decades compared to earlier periods. This relation will be further explored in the EDA.

2.3 Data Pre-processing

Regarding the UN General Assembly Corpus, several pre-processing steps were applied. Following standard natural language processing procedures, punctuation was removed, all text was converted to lowercase, and stop words were suppressed. Stemming was then performed to reduce words to their root forms (e.g., *election* and *elected* $\rightarrow$ *elect*), thereby reducing vocabulary complexity. These transformations were implemented with the Python library NLTK [1]. In addition, tokenization was applied to split the text into smaller units (words,

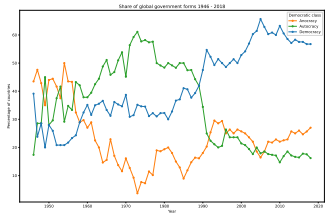


Fig. 1. Share of global government forms 1946 - 2018.

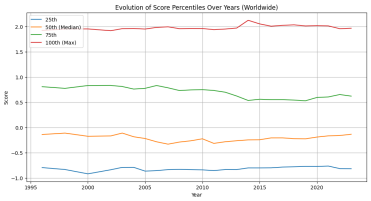


Fig. 2. Time series of quartiles of the Rule of Law index.

expressions, or subwords), providing different levels of granularity for feature engineering and exploratory analysis.

For the *Polity5* and *WGI* datasets, only the relevant columns (indexes, country tags, and years) were extracted. Rows not common across datasets were dropped to ensure consistency. Cases of missing data, often due to political turmoil, were excluded rather than imputed, as they represent a small share of the data and are particularly unreliable in transinitional periods.

3 Methodology

3.1 Feature Engineering

In supervised machine learning, a powerful technique for handling discrete or unstructured data is feature engineering, which transforms raw data into a more effective set of inputs. In our case, the UNGD Corpus consists of unstructured text, which can be encoded into numerical representations to facilitate modeling.

The most intuitive way to do so is by counting the occurrence of each term in every speech or document, essentially *vectorizing* the document and representing it by a vector of a size equal to the number of unique terms. To account for the length of the speeches, we can use the *Term Frequency* by dividing the term counts in a document by the total number of terms in the same document.

However, these approaches disregard the relative relevance of a term across the entire Corpus and might fail to distinguish signals from noises. For instance, in the context of UNGD speeches, terms such as 'nation', 'peace', 'right' are likely to be used frequently by all countries regardless of the government form. Therefore, we introduce the *TF-IDF* (Term Frequency - Inverse Document Frequency) method to rebalance the weights of term frequency:

$$TF\text{-}IDF(t,d) = \frac{\#\text{ term } t \text{ in doc } d}{\#\text{ all terms in doc } d} \cdot \log\left(\frac{N}{n_t}\right),$$

where N is the total number of documents, and n_t is the number of documents containing term t, and the latter term is called *inverse document frequency*. As one can see, the formula penalizes words with high occurrences across all speeches via $1/n_t$. As for the interpretation, a high TF-IDF score may indicate

that the word is particularly important for a specific speech, while being rarely mentioned in other speeches. Words with low TF-IDF scores could be either very frequent everywhere (noises) or absent.

3.2 Exploratory Data Analysis

Having vectorized the UNGD Corpus with TF-IDF, we now merge it with the *Democracy Index* and *Rule of Law Index* to generate integrated insights. To start, a list of democracy-related keywords was created using GloVe. Recognizing that less democratic regimes rarely describe themselves as "autocracies," the vocabulary was refined to capture more context-sensitive expressions.

Using this list, we are able to examine the trends of political mentions across UN General Debates. To identify the most important terms, TF-IDF scores were aggregated across speakers and years, and normalized by the number of speeches per year. Figure 3 presents the top 15 terms. Words such as "human" and "right" gained prominence since the 1990s, coinciding with democracy overtaking autocracy globally (Figure 1). Conversely, "democrat" and "democraci" peaked in the same period but declined after 2000.

For additional granularity, we examined salient terms by regime type (Figure 5). Less democratic countries emphasized "govern" and "independ," while "elect" was absent from their top terms. Autocracies also employed distinctive words like "imperialism." Meanwhile, terms such as "right," "commun," "social," and "freedom" were common across all regime types.

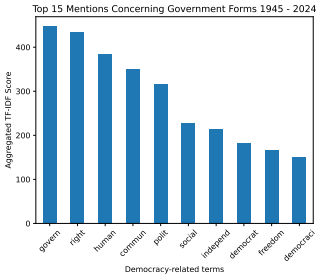


Fig. 3. Top 15 Mentions Concerning Government Forms 1945 - 2024.

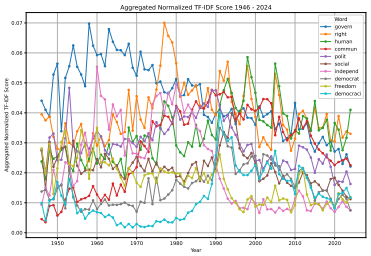


Fig. 4. Aggregated TF-IDF scores normalized by number of speeches per year 1946-2024.

A deeper analysis of the documents from a machine's perspective consists of the search for similarity across different speeches. For this process, different tokenization and vectorization approaches were followed. Namely, *BERTopic*, which organizes the speeches into different topics and determines their topic-based component in the form of a probability. The clusters of vectors that originated from this method did not reflect the expected similarities and proved insufficient for the classification effort that was the goal of the second research question. This

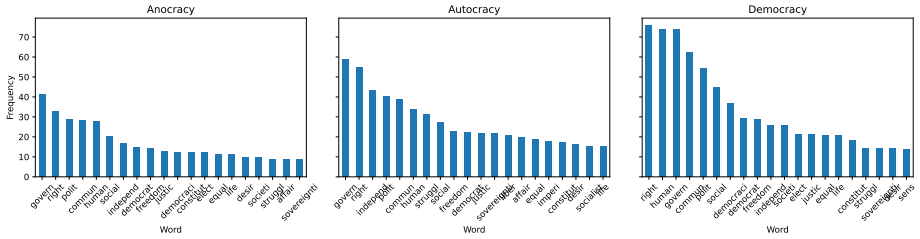


Fig. 5. Most important terms by government form.

is possibly due to the limit of 384 dimensions in the vectorization, which are not representative of the complexity of a speech even when the tokenization was done in split paragraphs.

Having this in consideration, the same approach of looking for proximity in the encoded vectors was applied to a much higher dimension, resorting to the results of the TF-IDF. This vectorization operates with a dimension of the number of stemmed words in the speeches' vocabulary, proving very successful in preserving the sentiment and content of a speech. The vectors are then projected in 2D for visualization, resorting to the PCA technique.

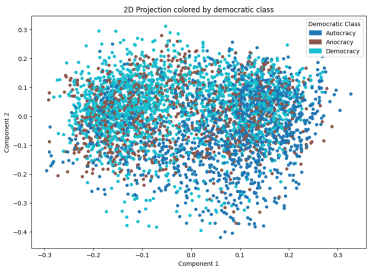


Fig. 6. Scatter plot of 2d projection of TF-IDF vectorization colored by regime.

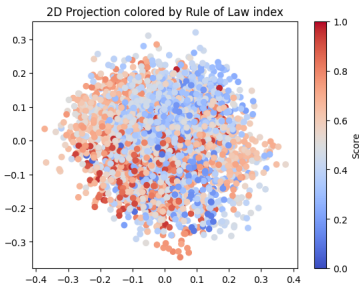


Fig. 7. Scatter plot of 2d projection of TF-IDF vectorization colored by Rule of Law index.

In Figure 6, it is clear why TF-IDF was the chosen method for this analysis. In the projection, vertical and horizontal orientation do not bring any particular meaning, since this is a reduction of a much higher-dimensional space, but the proximity of points does translate to similarity of the encoder vectors, in this case, the rows of the TF-IDF. The coloring according to regime class shows 2 clusters, one to the left of the plot, dominated by democratic speeches, and one to the right, dominated by autocratic speeches. The transition class, *Anocracy* is dispersed throughout the space, not showing clear clusters or similarity between their samples. It is also worth noting the presence of some displaced points

along the democratic and autocratic clusters, indicating that not all documents fit the general trend. This plot gives some confidence to the pursuit of the second research question, about the possibility of classifying a country's regime based on its speech.

The second scatter plot 7 seems coherent with the established idea. It is also possible to distinguish different areas with accumulation of speeches with close index values, although not as clearly, probably due to a smaller dataset. A deeper look into the relation between these 2 indexes will follow to close the EDA.

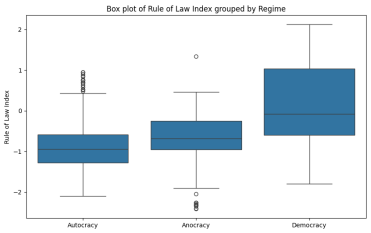


Fig. 8. Box plots of Rule of Law indexes divided by Regime classes.

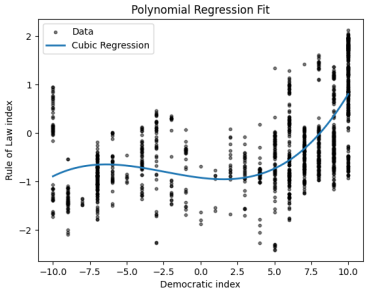


Fig. 9. Scatter plot of Democracy Index vs. Rule of Law index and respective regression.

Figure 8 shows the distribution of the Rule of Law indexes by government form. This follows the predicted behavior: the more democratic a country, the more likely it will comply with the rule of law. The mean increases with the approximation of a democratic regime, as well as the maximal values of the index. It is also notable that the minimum values for all categories stay around the same index of -2. On the right side, Figure 9 shows the scatter plot between the Rule of Law index and the Democratic Index (retaking back its original value between -10 and 10). A cubic regression was fit in the scatter but it is clear that the fit is highly compromised due to the fact that although numeric, the Democratic index is still a discrete value. The regression, as well as the box plots, still shows the purpose of showing that the tendency of the 2 indexes is to move together, even if not as linearly as initially expected.

3.3 Predicting Democracy Class

Model Selection

In order to classify the text data into democracy classes, the literature was reviewed to determine suitable ML models for text classification. The selection of models chosen for implementation was supported by [15], who has a very similar research setup. Each of these models has its own strengths and limitations; the aim of this paragraph is to list each model considered and offer a brief overview.

- The k-nearest neighbor method (KNN) adapts to any feature space and handles multi-class cases. However, its success rate is highly dependent on the choice of the distance function [11].
- Logistic regression models the probability of a categorical outcome by fitting a linear combination of features through a sigmoid function. However, its accuracy depends on the assumption of a linear relationship between the features and the log-odds.[11].
- Support Vector Machines (SVMs) are robust prediction models. They were used for binary classification but can also be applied to multi-class tasks. The choice of kernel and its parameters is key to performance ([7] & [8]).
- Stochastic Gradient Descent (SGD) is an optimization algorithm that efficiently updates model parameters in batches to minimize loss, making it effective for large-scale text classification [5]).

Model Implementation and Hyperparameter Tuning

After preprocessing and merging the speech data with their respective democracy class (Democracy, Autocracy, Anocracy), a supervised learning pipeline was built. The dataset was split into training (80%) and test (20%) sets using stratified sampling to preserve class balance. Hyperparameters were optimized on the training set through stratified 5-fold cross-validation with TF-IDF vectorization, and the best configuration was then evaluated on the unseen test set.

Several models were selected based on insights from the previous section. Hyperparameters were tuned using a combination of manual adjustments informed by the literature and systematic Grid Searches tailored to each model. Table 1 summarizes the parameter grids, final configurations, and best-performing runs.

Table 1. Hyperparameters per model category

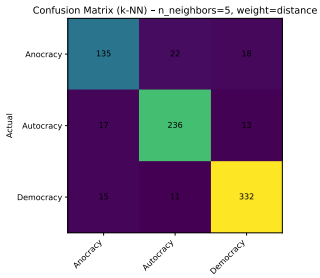
Models	Parameter Grid	Final Parameters	Score (avg)
k-Nearest Neighbors	n_neighbors, weights	n_neighbors = 5, weights = distance	Prec. 0.8798 F1 0.8633
SVM (rbf)	C, γ, class_weight	$C = 10$, $\gamma = 1$, class_weight = balanced	Acc.: 0.8786 F1: 0.8567
SGD Classifier	loss, α, penalty, l1_ratio, learning rate, η_0, power_t, average, class_weight	α: 1e-05, average: True, class_weight: balanced, η_0: 0.1, $l1$ ratio: 0.15, learning rate: constant, loss: hinge, penalty: $l2$, power t: 0.5	Acc.: 0.8586 F1: 0.8367
Logistic Regression	C, penalty, class_weight	$C = 100$, penalty = $l2$, class_weight = balanced	Acc.: 0.8536 F1: 0.8300
SVM (linear)	C, class_weight	$C = 1$, class_weight = balanced	Acc.: 0.8461 F1: 0.8300

Evaluation Metrics

To evaluate the performance of our model, two evaluations metrics were utilized, namely the **Classification report** and **Confusion matrix**.

Accuracy and F1 score are used to assess the models' performance [6]. Accuracy is an intuitive metric that is applicable for multi-class and multi-label problems. However, one must bear in mind that it has its limitations because of its low distinctiveness and discriminability, making it weak at identifying the optimal classifier. It also lacks informativeness and tends to be biased toward majority classes, failing to properly represent minority class performance [9]. The F1-score is beneficial when training on datasets, like ours, that have an unbalanced class distribution. The F1 score represents a harmonic measure of recall and precision. The ratio of true positives to positively classified data is known as precision. The percentage of true positives compared to all positives in the data is known as recall. Overall both metrics complement each other to give a more comprehensive evaluation of the models' performance.

Finally, a confusion matrix was generated to visualize the model's predictions on the test dataset in comparison with the true labels.



The k-NN confusion matrix shows high accuracy for all three classes, with most predictions landing on the diagonal. Anocracy is misclassified as Autocracy (22) and Democracy (18) in a notable number of cases, while Autocracy and Democracy have relatively few misclassifications. Overall, the k-NN model with $n\_neighbors = 5$ and distance weighting performs strongly, with some confusion between Anocracy and the other classes.

Fig. 10. Confusion Matrix for k-NN classifier ($n\_neighbors = 5$, distance weighting) and a brief analysis.

4 Results

Regarding the EDA, most of the results have been discussed directly in 3.2. The clear relation found between the 2 datasets and the speeches motivate the pursuit of the predictive question.

Using the metrics described in the previous section, each model was evaluated individually. The accuracy of all models is summarized in Table 1.

The results show that the KNN model achieved the highest accuracy at 88%, followed closely by the SVM with an RBF kernel (87%), SGD Classifier and Logistic Regression (both 86%). Overall, all tested models showed strong results, with accuracies ranging between 80% and 90%.

Although the accuracy scores of the SVM and SGD Classifier are comparable to those reported in [15], our findings are not entirely consistent with the literature, as the relative ranking of model performance differs across the selection.

In particular, k-NN achieved unexpectedly high accuracy despite its relatively poor performance with extensive datasets. Taking a deeper look at our best performing model, we can look at its classification report.

	precision	recall	f1-score	support
Anocracy	0.81	0.77	0.79	175
Autocracy	0.88	0.89	0.88	266
Democracy	0.91	0.93	0.92	358
accuracy			0.88	799
macro avg	0.87	0.86	0.86	799
weighted avg	0.88	0.88	0.88	799

The k-NN model shows strong overall performance, achieving an accuracy of 0.88 across the three classes. Precision, recall, and f1-score are highest for Democracy (0.91, 0.93, 0.92), moderate for Autocracy (0.88, 0.89, 0.88), and slightly lower for Anocracy (0.81, 0.77, 0.79). The macro and weighted averages indicate balanced performance, although Anocracy is more challenging to classify compared to the other categories.

Fig. 11. k-NN Classification report.

5 Discussion

Although the results are promising with respect to the research question, several considerations must be taken into account when interpreting them. First, the speeches were translated into English, which may have introduced bias. Second, linguistic patterns may vary over time, yet the present study generalizes across the entire period, which could limit the validity. Finally, although it yielded the best resutls in this research, relying solely on TF-IDF as a feature engineering method has its limitations, since it cannot capture sentiment or deeper semantic relationships between words. Word embeddings techniques were utilized, but the model performed poorly with this approach, as explained in previous sections.

Overall, we can conclude this research in a successful note. The discoveries in the EDA were well aligned with the expectations, confirming our hopes related to the SDG-16, and we could establish the relation between regime characteristics and rhetoric in the UN speeches. Additionally, it motivated the predictive question, where we verified that the found relation does hold in a classification effort for the regime.

6 Conclusion

This study examines whether regime characteristics shape speech in the UNGD Corpus, focusing on topics from United Nations SDG-16. Exploratory analysis reveals differences in stances across regime types and clustering of speeches after vectorization. We then train a k-NN classifier on stemmed speeches to predict a speaker's regime class. The model's high accuracy suggests that rhetoric in UN General Debates can reflect a country's level of democracy.

References

1. Nltk: Natural language toolkit. https://www.nltk.org/, accessed: 2025-09-29
2. Bank, W.: World governance indicators (2023), https://databank.worldbank.org/source/worldwide-governance-indicators, accessed: 2025-09-29
3. Baviskar, S., Malone, M.: What Democracy Means to Citizens – and Why It Matters. Revista europea de estudios latinoamericanos y del Caribe = European Review of Latin American and Caribbean Studies p. 3 (04 2004). https://doi.org/10.18352/erlacs.9682
4. Boese, V.A.: How (not) to Measure Democracy. International Area Studies Review **22**(2), 95–127 (2019), https://EconPapers.repec.org/RePEc:sae:intare:v:22:y:2019:i:2:p:95-127
5. Bottou, L.: Stochastic Gradient Descent Tricks, pp. 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg (2012), https://doi.org/10.1007/978-3-642-35289-8_25
6. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics **21** (01 2020). https://doi.org/10.1186/s12864-019-6413-7
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995). https://doi.org/10.1007/BF00994018
8. Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A.: A survey on text classification algorithms: From text to predictions. Information **13**, 83 (02 2022). https://doi.org/10.3390/info13020083
9. Hossin, M.B., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process **5**(2), 1–11 (03 2015). https://doi.org/10.5121/ijdkp.2015.5201
10. Jankin, S., Baturo, A., Dasandi, N.: United Nations General Debate Corpus 1946-2024 (2017). https://doi.org/10.7910/DVN/0TJX8Y, https://doi.org/10.7910/DVN/0TJX8Y
11. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text Classification Algorithms: A Survey. Information **10** (04 2019). https://doi.org/10.3390/info10040150
12. Leininger J., Luehrmann A., S.R.: The relevance of social policies for democracy: preventing autocratisation through synergies between sdg 10 and sdg 16 (2019). https://doi.org/10.23661/dp7.2019
13. for Systemic Peace, C.: Polity project (2018), https://www.systemicpeace.org/polityproject.html, accessed: 2025-09-29
14. United Nations: Sustainable Development Goals 16, https://sdgs.un.org/goals/goal16, accessed: 2025-09-29
15. Yanmaz, A.: Predicting Freedom: An Analysis of the United Nations General Debate Corpus (2023), https://arno.uvt.nl/show.cgi?fid=170614