Final Project Memo

Ella Bartt, Poli 176, Dec 2025

I. Introduction and Background

The UN is, overall, a forum for international collaboration and problem solving. The UN General Debate corpus reflects that in numerous ways; many countries begin and end their speeches with gratitude for the platform, while also voicing their opinions on the past year's events. Since the UNGD is not necessarily linked to a vote and therefore a more open platform, Baturo, Dasandi, and Makhaylov note: "member states can more freely express their government's perspectives on issues deemed important – including more contentious issues" (Baturo et al. 2017, 2). These "contentious issues" sometimes include the structure of the UN itself– many smaller states use the general debate to voice their concern over the growing power of the world's superpowers, as they "can use the GD to raise more disagreeable political issues" (2). Using the UN General Debate Corpus, I plan to answer the question: Which countries' speeches directly criticize major global powers and the structure of the UN itself with respect to power imbalance? While some countries may criticize a certain policy or country's actions in passing, I hope to look at which countries repeatedly mention UN shortcomings, and which words constitute said criticism. I predict I will find that smaller developing nations either criticize the UN less when asking for aid, or criticize the UN more due to ongoing colonial structures and conflicts in their countries.

In answering the research question, I hope to uncover how greater dynamics between developing and developed countries, regions, and democracies impact the general push for UN reform and negative sentiment towards its management. While the UN includes many smaller states and developing nations, it is not all-encompassing. For example, Palestine is a "Permanent Observer State," but not a full member, despite the state's repeated push for membership ("Palestine's Status at the UN Explained | UN News" 2024). The Palestinian question is a topic of debate throughout the entire corpus– many countries use the General Debate platform to demand peace and show support for Palestinian statehood, going back as far as the corpus reaches. The question of membership and statehood of non-UN nations is just one example of a point of contention amongst member states that may arise in the UNGD corpus, though, and I will be focusing on more general criticism.

While there is little readily available data on whether this has been done before with the UNGD corpus, there are countless scholars that have analyzed pushes for UN reform and conflict within the debate. In "Clashes in the Assembly", Voeten analyzes UN Assembly votes in the 20th century with a focus on the Cold War period and the collapse of the Soviet bloc. Although the data and timeframe differ from those of the UNGD corpus, Voeten's summary of UN post-Soviet hypotheses amongst scholars provide helpful context for this project. For example, he summarizes the "Structuralist view" as that which "argues that weak states share a common interest in blocking adoption of policy positions of powerful states… voting behavior of developing countries is determined by their perception of powerlessness resulting from their position of relative (economic) power in the international system" (Voeten 2000, 189). The "counterhegemonic bloc hypothesis," on the other hand, "states that a counterhegemonic voting bloc has formed consisting of countries that challenge the United States and the (liberal)

principles that rule the international system" (190). The post-Soviet framework is similarly applied to a subset of the UNGD corpus in Zhou and Khurusu's work analyzing post-Cold War speeches from major world powers. The authors applied the Lexicoder Sentiment Dictionary and semi-supervised topic modeling to the Group of Seven's speeches and found that "major powers have divergent value preferences… China and Japan emphasize developmental values, and the US prefers democratic values more than other major countries" (Zhou and Kurusu 2021, 79). Both of the aforementioned papers build a helpful framework on the UN's history of conflict in policy voting and the General Debate.

II. Measurement

Step 1: Hand Coding

In order to analyze criticism of the UN, I will use external data in addition to the corpus. The measure I will be quantifying is criticism of the UN structure generally, classified as texts that directly criticize major global powers and the structure of the UN itself with respect to internal power imbalance. Many speeches classified as critical overall repeatedly mention the need for UN reform and power imbalances between the major world powers, such as the Group of Seven, and developing countries. To accurately categorize developing countries, I use the "Standard country or area codes for statistical use (M49)" dataset from the UN's website and the Varieties of Democracy (V-Dem) dataset ("UNSD — Methodology," n.d.). V-Dem's measures of democracy by country and year are widely cited across political science papers, hence my decision to use their index (Herre et al. 2013). These additional measures provide more detailed regional data, including the UN's Least Developed Countries (LDC), Landlocked Developing Countries (LLDC), and Small Island Developing States (SIDS). See the appendix for the codebook with more details. Once I run the model according to my measure of critical speech, I will determine how the model groups each of these characteristics.

For handcoding, I defined the variable "critical" to indicate whether a specific text was critical of the UN, written as follows:

- 1 = Texts that directly criticize the major global powers and the structure of the UN itself with respect to said power imbalance. Implicit references alone do not count, but repeated criticism (esp. pushing for UN reform) does. References to the "North" vs. "South" may count as long as it is still critical of the UN's imbalance between the two.
- 0 = otherwise.

Once selected, I traded 100 random documents with a classmate, and the results were as follows:

	Handcoder 1	
Handcoder 2	0	1
0	71	0
1	14	15

Krippendorff's alpha:
Subjects = 8093
Raters = 2
alpha = 0.594

I found that intercoder reliability was not the best, but could definitely be worse. There were no documents that handcoder 1 coded as 1 that handcoder 2 did not. However, handcoder 2 was

much more liberal with assigning the "1" label to the selected documents. This was most likely a result of too general a measure, where handcoder 1 coded "0" where unsure and handcoder 2 coded "1" in many of those cases. Upon investigation, I found many of the ambiguous documents do, in fact, call for UN reform and criticize certain international political decisions, but may also express hope that the UN can address the issues accordingly. Other documents may criticize the UN a couple times, in which case one coder deemed it a "1" but it may not have been repeated enough for the other coder.

To boost intercoder reliability, I would definitely go back and split the documents into smaller documents in order to make this much easier in this portion, as the documents were quite long. The length may have contributed to some information being lost when either handcoder was skimming the document. To refine this, I would have liked to make it more concrete– all texts that mention and support UN reform code as 1 and all others code as 0. Given the time constraints and length of the documents, I did not feel comfortable changing the measure at this point and asking the handcoder to go through this again. To hopefully make up for some of these shortcomings, I decided to set the "critical" column in my dataframe as "1" if *either* handcoder coded it as a "1", not both.

Step 2: Measurement

I used the Lasso model to train the classifier in my model. I decided on supervised machine learning because I wanted to see which words the model weighted as being predictive of "critical," and I prefer the supervised method to any LLM (personal bias). In order to do this, I coded the "critical" column as the dependent variable, with the document feature matrix as my independent variable. When testing different models, I found that the Lasso model performed the best by a wide margin when looking at precision, recall, and accuracy, so I chose this model overall. Additionally, I felt that the Lasso model was best equipped for such a large document feature matrix.

```
> print(conf_matrix)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 17  4
         1  2  2

               Accuracy : 0.76
                 95% CI : (0.5487, 0.9064)
    No Information Rate : 0.76
    P-Value [Acc > NIR] : 0.6073

                  Kappa : 0.2574

 Mcnemar's Test P-Value : 0.6831

              Precision : 0.5000
                 Recall : 0.3333
                     F1 : 0.4000
             Prevalence : 0.2400
         Detection Rate : 0.0800
   Detection Prevalence : 0.1600
      Balanced Accuracy : 0.6140

       'Positive' Class : 1
```

Precision

0	1
0.500	1.00

Recall

0	1
1.00	0.3333

When setting up the model, I defined a 75% train-test split for cross-validation, and got this as the out-of-sample performance. The accuracy is by far the highest out of all the possible models, but it is not difficult to see that the model could be much better. I initially ran the model without upweighting true positives and I found that while accuracy increased to .80, the model predicted very few actual positives, which led to a much lower recall of about .167. Since the goal is predicting criticism and not just predicting every document as a 0, I upweighted it. The p-value is relatively high, and while precision is very high (no false positives), recall is still quite low (many false negatives). Once again, I think that this is a result of an unclear dependent variable which would have been solved with a much more concrete measure, such as any support of UN reform. Before moving on to applying this to the entire corpus, I checked the top features in the model to see which words were most "critical":

```
> sort(lasso.1$beta[,40], decreasing=T)[1:30]
  reconsid    virtual      cuban     inclin    prodigi      claim     equiti      magic     servic
2.86646381 1.71479259 1.07267430 0.64965160 0.64603885 0.63579101 0.60799420 0.53211412 0.45173769
    depriv   deliveri        ten      daili     unfair    condemn    heinous       slow      skill
0.36415476 0.32181705 0.30640845 0.27751987 0.26847065 0.22227462 0.21999212 0.21604632 0.20955379
  otherwis  intervent       pain     denial     regret      storm     unjust    obsolet       call
0.20816807 0.20145947 0.19422417 0.12729860 0.12616179 0.11607644 0.11215819 0.10002234 0.09477190
    justif      fruit      level
0.07560439 0.07138721 0.01640108
```

As I will discuss in the next section, this aligns with what I initially thought of as salient themes that may appear in criticisms of the UN, especially with respect to government interventions ("cuban," "intervent") and power imbalances ("equiti," "unfair," "unjust"). I ultimately decided to apply the lasso model to the whole dataset.

Step 3: Application

After evaluating the effectiveness of the Lasso model, I applied it to the remaining texts in the UNGD corpus. Even with upweighting the predicted "critical" texts, the model still underpredicted "critical," as the proportions were slightly different for the whole corpus as they were for the hand coded data. As shown below, the model predicted 19.3% "critical" texts, while the hand coded data contained 29% "critical" texts. However, knowing that the precision was quite high for the training data, one can infer that the positives are very accurate, with a negligible amount of false positives (the same cannot be said for any false negatives).

```
> #Prediction of critical for the whole corpus
> prop.table(table(undf$predict.critical))

        0         1
0.8069937 0.1930063
> #handcoded proportion:
> prop.table(table(undf$critical))

   0    1
0.71 0.29
```

In order to answer my research question, I then looked at the spread of predicted critical values by region, subregion, and country as follows:

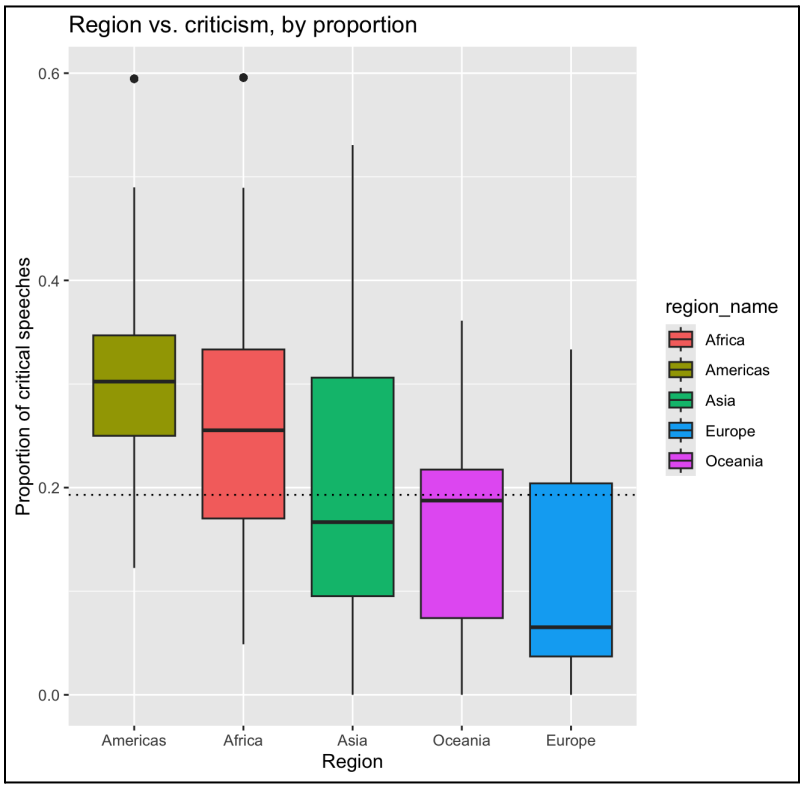


Fig 1. Boxplot of predicted criticism by region, measured as proportion of each country's total number of speeches. The dotted line represents the overall average of predicted critical speeches.

```
# Groups:   region_name [5]
   region_name subregion                       prop_critical
   <chr>       <chr>                                   <dbl>
 1 Africa      Northern Africa                        0.410
 2 Americas    Latin America and the Caribbean        0.312
 3 Asia        Western Asia                           0.275
 4 Asia        Southern Asia                          0.265
 5 Oceania     Melanesia                              0.259
 6 Africa      Sub-Saharan Africa                     0.233
 7 Americas    Northern America                       0.173
 8 Oceania     Australia and New Zealand              0.163
 9 Europe      Western Europe                         0.156
10 Asia        South-eastern Asia                     0.148
11 Europe      Southern Europe                        0.133
12 Oceania     Micronesia                             0.1
13 Europe      Eastern Europe                         0.0863
14 Europe      Northern Europe                        0.0742
15 Asia        Eastern Asia                           0.0547
16 Oceania     Polynesia                              0.0286
17 Asia        Central Asia                           0.00787
```

Fig 2. Table of the subregions with the highest percentages of predicted.critical speeches.

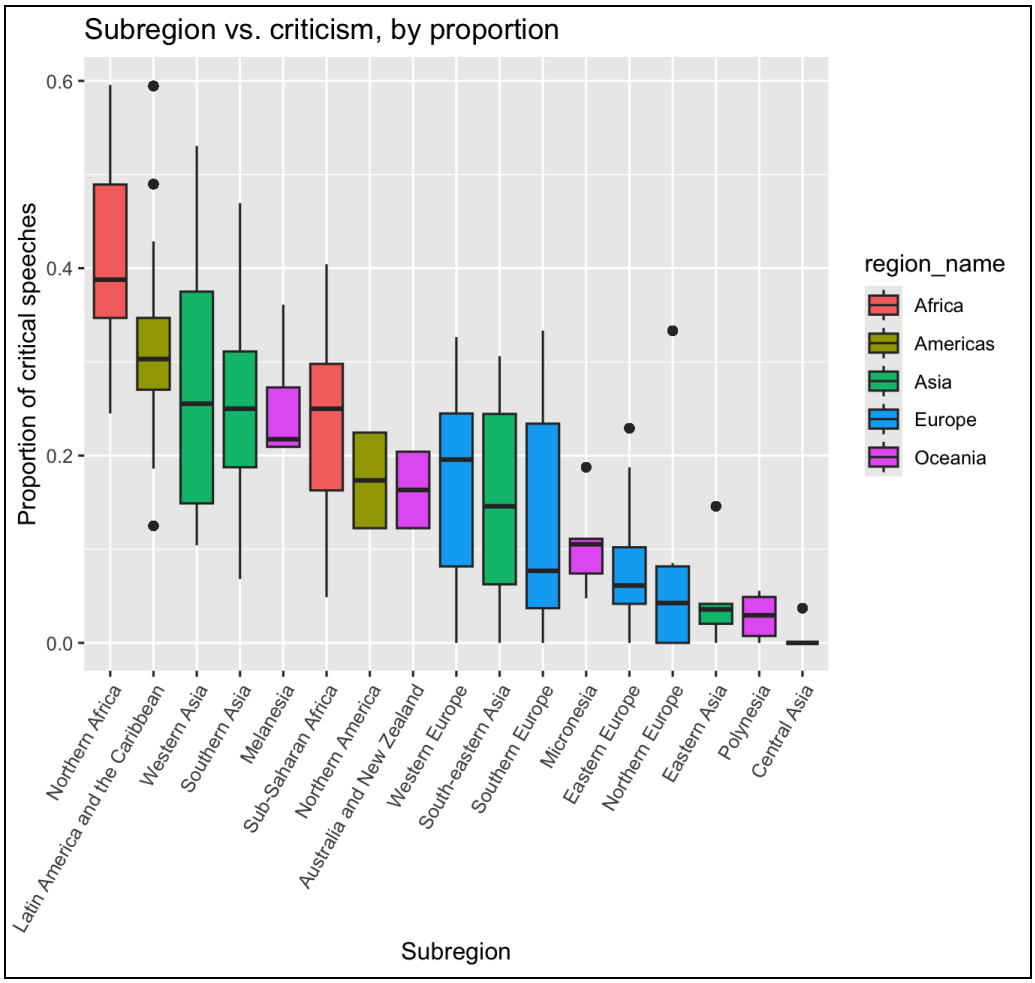


Fig 3. Proportion of predicted critical speeches, by subregion. Grouped first by country then aggregated into the boxplot. Note: The upper outliers for Latin America are Belize (.595) and Cuba (.490). The upper outlier for Northern Europe is Ireland (0.33).

	country_name	country	prop_critical
	<chr>	<chr>	<dbl>
1	Libya	LBY	0.596
2	Belize	BLZ	0.595
3	Syrian Arab Republic	SYR	0.531
4	NA	YDYE	0.526
5	Cuba	CUB	0.490
6	Egypt	EGY	0.489
7	Iraq	IRQ	0.469
8	Pakistan	PAK	0.469
9	Bolivia (Plurinational State of)	BOL	0.429
10	Jamaica	JAM	0.417

Fig 4. Ten most predicted critical countries. Note: YDYE is the former state of South Yemen before merging with North Yemen, now Yemen.

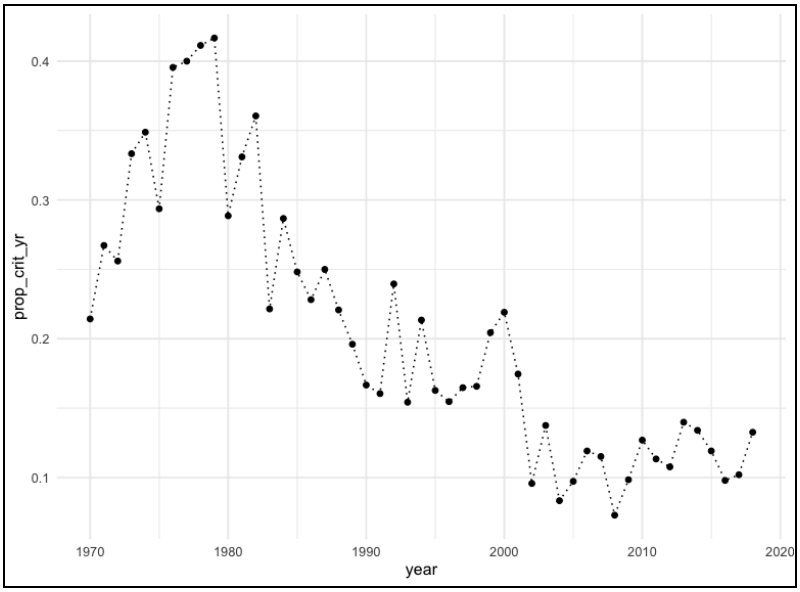
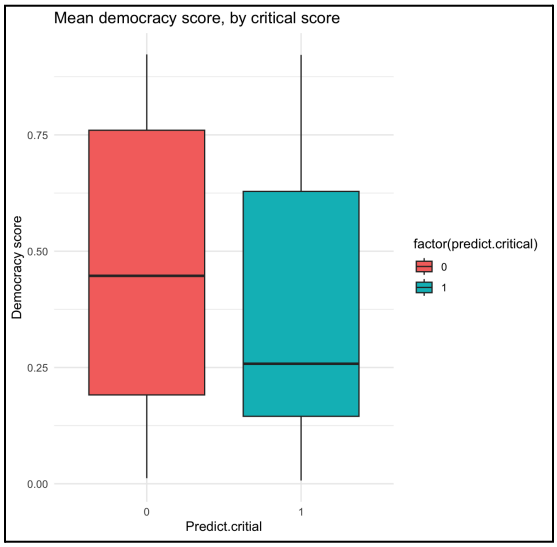


Fig 5. The percentage of predicted.critical speeches over time.



Figs 6. The democracy score grouped by country and predicted.critical value.

As shown in figures 1-4, the most heavily predicted critical regions were the Americas and Africa. When broken down by subregion, one can see that the most critical (predicted) subregions were North Africa and Latin America and the Caribbean. These insights more or less fit with my hypothesis, that historically colonized countries would criticize the structure of the UN with respect to the major powers' domination of the organization. The outliers for Latin America with the highest percentages of critical speeches were Belize and Cuba, showing that the model may pick up on any mention of "Cuba" as being indicative of an overall critical speech. This misattribution of the model– attributing mentions of certain polemic events to an overall critical tone and negativity, then labeling any mention of the event as "critical"-- was probably replicated in other regions as well with historic conflicts during the time span of the UNGD corpus. As seen in figure 5, predicted critical speeches peak in the 1970s and decrease dramatically as time goes on, especially after 2000. While this could be due to a true decrease in calls for UN reform, I do not feel confident inferring this from the model. This is probably a

result of the end of certain conflicts that, as previously discussed, the model always predicted as critical, like U.S. interventions in Central American politics and trade or the Cold War. As for figure 6, the same hesitancy applies. I do not feel confident in stating definitively whether the mean democracy index score for critical countries is lower than that of non-critical countries without a stronger model, but the visual itself raises interesting questions. Given a more rigorous model, I would have liked to see if this pattern would hold across democracies vs. non-democracies.

While the charts do provide interesting insight into potential patterns of criticism of the UN over time and across regions, the model applied does not allow me to confirm or reject the original hypotheses. This problem could potentially be addressed by changing my hand coded variable and codebook (as discussed on page 3), and splitting the documents into smaller chunks. Overall, however, the model still provides interesting insight into perceived criticism of the UN as dependent on time and certain specific global conflicts. The percentages of critical speeches did vary dramatically across regions, fitting in with my hypothesis that European countries and major global powers would be less critical of the UN. Moving forward, I hope to apply this method more rigorously to other text corpora, the UN General Debate Corpus included.

Appendix: Codebook

- **Doc_id:** UNGDC document id.
- **Text:** The content of the UNGD speech.
- **Country:** The ISO-alpha3 country code.
- **Session:** The UN session.
- **Year:** the year the session took place.
- **Id:** The index of the document within the corpus, for cross-validation purposes.
- **Country_name**: The full name of the country corresponding to the ISO-alpha3 country code, taken from the UN Standard Country Codes dataset.
- **Region:** The region (continent) of the country according to the UN, taken from the UN Standard Country Codes dataset.
- **Subregion:** The subregion of the country, taken from the UN Standard Country Codes dataset.
- **LDC_01**: Least Developed Country. Binary variable, taken from the UN Standard Country Codes dataset, turned into a binary variable for analysis purposes (this used an "x" for 1 before). 1 = LDC, 0 = not applicable.
- **LLDC_01**: Landlocked developing country. Binary variable, taken from the UN Standard Country Codes dataset, turned into a binary variable for analysis purposes (this used an "x" for 1 before). 1 = LLDC, 0 = not applicable.
- **SIDS_01**: Small Island Developing State. Binary variable, taken from the UN Standard Country Codes dataset, turned into a binary variable for analysis purposes (this used an "x" for 1 before). 1 = SIDS, 0 = not applicable.
- **V2x_polyarchy:** Scale, from 0-1, taken from the V-Dem dataset. From the V-Dem codebook: "Question: To what extent is the ideal of electoral democracy in its fullest sense achieved?" (Coppedge et al. 2025, 46). 1 = Electoral democracy is fully achieved, 0 = electoral democracy is not at all achieved. Expressed as a decimal value.
- **Critical** (HANDCODED):
    - 1 = Texts that directly criticize the major global powers and the structure of the UN itself with respect to said power imbalance. Implicit references alone do not count, but repeated criticism (esp. pushing for UN reform) does. References to the "North" vs. "South" may count as long as it is still critical of the UN's imbalance between the two.
    - 0 = otherwise.

Works Cited

Baturo, Alexander, Niheer Dasandi, and Slava J. Mikhaylov. 2017. "Understanding State Preferences with Text as Data: Introducing the UN General Debate Corpus." *Research & Politics* 4 (2): 2053168017712821. https://doi.org/10.1177/2053168017712821.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Fabio Angiolillo, Michael Bernhard, Agnes Cornell, M. Steven Fish, Linnea Fox, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Ana Good God, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Joshua Krusell, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Natalia Natsika, Anja Neundorf, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Marcus Tannenberg, Eitan Tzelgov, Yi-ting Wang, Felix Wiebrecht, Tore Wig, Steven Wilson and Daniel Ziblatt. 2025. "V-Dem [Country-Year/Country-Date] Dataset v15" Varieties of Democracy (V-Dem) Project. https://doi.org/10.23696/vdemds25.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Fabio Angiolillo, Michael Bernhard, Agnes Cornell, M. Steven Fish, Linnea Fox, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Ana Good God, Sandra Grahn, Allen Hicken, Katrin Kinzelbach, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Anja Neundorf, Pamela Paxton, Daniel Pemstein, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Marcus Tannenberg, Eitan Tzelgov, Yi-ting Wang, Felix Wiebrecht, Tore Wig, and Daniel Ziblatt. 2025. "V-Dem Codebook v15" Varieties of Democracy (V-Dem) Project.

Herre, Bastian, Lucas Rodés-Guirao, and Esteban Ortiz-Ospina. 2013. "Democracy." *Our World in Data*, March 15. https://ourworldindata.org/democracy.

"Palestine's Status at the UN Explained | UN News." 2024. April 18. https://news.un.org/en/story/2024/04/1148351.

Slaughter, Anne-Marie. 2005. "Security, Solidarity, and Sovereignty: The Grand Themes of UN Reform." *American Journal of International Law* 99 (3): 619–31. https://doi.org/10.2307/1602294.

"UNSD — Methodology." n.d. Accessed December 1, 2025. https://unstats.un.org/unsd/methodology/m49/overview/.

Voeten, Erik. 2000. "Clashes in the Assembly." *International Organization* 54 (2): 185–215.

Zhou, Yuan, and Kaoru Kurusu. 2021. "How Major Powers Diverge on Global Governance? Evidence from the United Nations General Debate." Preprint, 神戸大学大学院法学研究科. https://doi.org/10.24546/81013152.