Replication Materials for
"Racial Tropes in the Foreign Policy Bureaucracy: A Computational Text Analysis"

Austin Carson, Eric Min, and Maya Van Nuys

Last updated March 25, 2024

This document contains all information regarding the files necessary to replicate all results in the main manuscript and the appendix.

The document is divided into four main sections. In the first, we list and briefly describe the function of every single item in the replication file. In the second, we specify which R file produces a table, figure, or set of results that appears in the main text and appendix. In the third, we describe all objects in an R workspace that contains the results of our supervised learning and structural topic models. In the fourth, we describe every covariate in the CSV files that contain relevant data.

Technical Specifications **2**
Overview of Files **2**
RData Files 2
CSV Files 2
R Files 3
Files to Produce Tables, Figures, and Results **4**
Main manuscript 4
Appendix 5
Objects and Variables in R Workspace **6**
predictionsAndTopics.RData 6
Variables in CSV Files **8**
pdbEntries.csv 8
predictedData.csv 9
regionData.csv 10
ysiData.csv 10
tropeData.csv 11
stmLabelsData.csv 13
stmCoherenceValidation.csv 13
stmLabelValidation.csv 14

Technical Specifications

All analyses were performed on a PC with Windows 11 Pro, using R version 4.3.1.

Overview of Files

RData Files

File	Description
predictionsAndTopics.RData	R workspace containing data from all supervised learning models of racial tropes and main structural topic model.

CSV Files

File	Description
pdbEntries.csv	Dataset containing all entries from the President's Daily Brief. Note that this file is not explicitly used in any of the R files in this replication package; it is simply provided for informational purposes.
predictedData.csv	Dataset of racial tropes from supervised learning models.
regionData.csv	Dataset of countries and associated regions.
ysiData.csv	Dataset tracking the number of years since independence (post-WWI) for all relevant country-years.
tropeData.csv	Dataset containing all racial trope data and other necessary variables necessary for the main analysis.
stmLabels.csv	File containing hand-coded names for all 65 topics in the structural topic model used in the analysis.
stmCoherenceValidation.csv	File containing data to perform topic coherence validation (random 4-word set intrusion task) for structural topic model.
stmLabelValidation.csv	File containing data to perform label validation (optimal labor task) for structural topic model.

R Files

File	Description
loadPackages.R	Installs and loads all necessary packages.
doRegressions.R	Performs and exports results from regressions in "analyzeTropes.R."
getPredictedData.R	Extracts best data from all supervised learning models; produces core racial trope data. Data used: • /data/predictionsAndTopics.RData Data produced: • /data/predictedData.csv
modelMetrics.R	Gets predictive performance metrics from all supervised learning models. Data used: • predictionsAndTopics.RData
addFeatures.R	Adds necessary variables to the core racial trope data. Data used: • /outside_data/predictedData.csv • /outside_data/regionData.csv • /outside_data/icowcol.csv • /outside_data/AoW_v1.csv • /outside_data/P5v2018.xlsx • /outside_data/diplomatic_core.replication.dta • /outside_data/ucdp-prio-acd-201.csv Data produced: • /data/tropeData.csv
analyzeTropes.R	Conducts main statistical analysis. Data used: • /code/doRegressions.R • /data/tropeData.csv
appendix.R	Generates all results in the appendix. Data used: • /data/tropeData.csv • /data/predictionsAndTopics.RData • /data/stmLabels.xlsx • /data/stmCoherenceValidation.xlsx • /data/stmLabelValidation.xlsx • /code/doRegressions.R

Files to Produce Tables, Figures, and Results

Main manuscript

Item in Manuscript	Relevant File
Table 1: Original terms used with thesaurus to create dictionaries	getPredictedData.R
Table 2: Performance for highest-quality predictive models	modelMetrics.R
Table 3: Example PDB entries with high predicted probability of exhibiting key themes	analyzeTropes.R
Figure 1: Histograms of dependent variables	analyzeTropes.R
Figure 2(a): Map of the "Global South" (Group of 77 and China)	analyzeTropes.R
Figure 2(b): Years since independence in 1977. Countries in very light white were not independent or gained independence prior to World War I	analyzeTropes.R
Table 4: Results of regressions on relationship between racial tropes in PDB entries and measurements of the racialized Otherness	analyzeTropes.R
Table 5: Predicted values from models in Table 4	analyzeTropes.R
Figure 3: Coefficient plots of regressions in Table 4 and additional models that include topic prevalence measures from structural topic model	analyzeTropes.R
Figure 4: Map of geographical regions	analyzeTropes.R
Figure 5: Coefficient plots of regressions disaggregating countries by region	analyzeTropes.R
Figure 6: Coefficient estimates for the Global South variable, using a moving seven-year temporal window and full models accounting for topics	analyzeTropes.R

Appendix

Appendix Section	Relevant File
2.1: Method 1: Trope Dictionaries	appendix.R
2.2.4: Model Metrics	modelMetrics.R
3: Quantitative Descriptive Statistics	appendix.R
4.1: Full Results	appendix.R
4.2: Presidents and Time	appendix.R
4.3: Entry Topics with a Structural Topic Model	appendix.R
4.4: Including Leader Tenure	appendix.R
4.5: Only Decolonized States	appendix.R
4.6: More on Regional Variation	appendix.R
4.7: More on Temporal Variation	appendix.R

Objects and Variables in R Workspace

predictionsAndTopics.RData

R workspace which contains relevant data and output from supervised learning models and structural topic model.

- *all[METRIC]Matrix:* A list containing all measures of a given metric for all supervised learning models that only use document embeddings to reflect each PDB entry.
- *all[METRIC]MatrixDict:* A list containing all measures of a given metric for all supervised learning models that only use counts of trope-related terms to reflect each PDB entry.
- *all[METRIC]MatrixSyn:* A list containing all measures of a given metric for all supervised learning models that use both document embeddings and counts of trope-related terms to reflect each PDB entry.
- *allPredsList:* A list containing predicted levels of tropes for all supervised learning models that only use document embeddings to reflect each PDB entry.
- *allPredsListDict:* A list containing predicted levels of tropes for all supervised learning models that only use counts of trope-related terms to reflect each PDB entry.
- *allPredsListSyn:* A list containing predicted levels of tropes for all supervised learning models that use both document embeddings and counts of trope-related terms to reflect each PDB entry.
- *allTabsList:* A list containing all confusion matrices for all supervised learning models that only use document embeddings to reflect each PDB entry.
- *allTabsListDict:* A list containing all confusion matrices for all supervised learning models that only use counts of trope-related terms to reflect each PDB entry.
- *allTabsListSyn:* A list containing all confusion matrices for all supervised learning models that use both document embeddings and counts of trope-related terms to reflect each PDB entry.
- *dd[N]:* Document embeddings of dimension N (N could take the values 16, 32, 64, 128, 256, 512) for all PDB entries.
- *dtm:* Document-term matrix for all PDB entries.
- *leaderCount:* Data frame containing the number of times leaders were mentioned in an entry, as well as how many days of tenure a leader had in the country on the day that the PDB was distributed.
- *pdc:* The raw PDB entry data for all relevant entries, structured at the country-entry level. Note that the country-entry level is the main unit of analysis in the statistical results.
- *pde:* The raw PDB entry data for all relevant entries, structured at the entry level.
- *segstm:* STM object containing a 65-topic structural topic model.

- *synCounts:* A list containing the number of times synonyms for each of the four tropes were used in the PDB entry data. The list has four smaller objects inside, which correspond with the four tropes (infantilization, animal analogies, belligerence, and irrationality).
- *topicData:* Topic prevalences for all relevant entries according to the structural topic model contained in *segstm*.
- *tset:* The raw training data used in the supervised learning models.
- *all_[TROPE]_syn:* A vector containing all synonyms associated with each trope (infant = infantilization, animal = animal analogies, bellig = belligerence, irr = irrationality). Note that these vectors' lengths are higher than what is reported in the "Total stemmed synonyms" row on Table 1, as Table 1 only enumerates the number of synonyms that actually appeared in the PDB data. This vector contains all synonyms, regardless of whether they appeared in the PDB. To get the values reported in Table 1, calculate the number of columns in each object inside *synCounts*.
- *models:* A vector containing the names of all supervised learning models used to create predicted trope levels.
- *pdeVocab:* A vector containing the vocabulary of the PDB entries, with sparser terms removed. Terms were considered sparse if they occurred fewer than 200 times overall or appeared in fewer than 50 documents.
- *themes:* A vector containing abbreviated names for the four tropes (INFANT = infantilization, ANIM = animal analogies, BELLIG = belligerence, IRR = irrationality).

Variables in CSV Files

pdbEntries.csv

Dataset containing all entries from the President's Daily Brief (PDB). Note that this file is not explicitly used in any of the R files in this replication package; it is simply provided for informational purposes. This file has more rows than the *pde* object inside *predictionsAndTopics.RData* because this file is the entire untouched dataset, including rows of two words or fewer that simply say "Notes" or "Annex." *pde* only contains rows that had pertinent information for the study at hand.

- *pdbID:* The document number given by the CIA.
- *date:* The date on which the PDB was distributed.
- *admin:* The administration in office on the day the PDB was distributed.
- *year:* The year in which the PDB was distributed.
- *entryID:* A unique identifier for the specific entry. This identifier is constructed by combining the *pdbID* variable with the *entryNum* variable.
- *entryNum:* The sequential number of the entry in a PDB.
- *title:* The title of the PDB entry, if there is one.
- *text:* The text of the PDB entry.

predictedData.csv

Dataset of racial tropes from supervised learning models.

- *pdbID:* The document number given by the CIA.
- *date:* The date on which the PDB was distributed.
- *admin:* The administration in office on the day the PDB was distributed.
- *year:* The year in which the PDB was distributed.
- *entryID:* A unique identifier for the specific entry. This identifier is constructed by combining the *pdbID* variable with the *entryNum* variable.
- *entryNum:* The sequential number of the entry in a PDB.
- *title:* The title of the PDB entry, if there is one.
- *text:* The text of the PDB entry.
- *tt:* The combined text from the *title* and *text* variables.
- *nWords:* The number of words in the *tt* variable.
- *ccode:* The Correlates of War (COW) country code for the state mentioned in the entry.
- *leaderMention:* A binary variable indicating whether a head of state was mentioned in the entry (0=no, 1=yes).
- *leaderTenure:* The number of days that a head of state has been in office as of the date of the entry.
- *pred[TROPE]:* The predicted level of a given trope, according to the supervised learning model which performed best in terms of Cohen's Kappa.
- *count[TROPE]:* The number of times any words/phrases in the trope dictionaries occurred in the entry for a given trope. For example, *countBELLIG* reports the number of words/phrases that appear in an entry reflecting the belligerence trope.
- *TopicN:* The prevalence of topic N in the entry, according to the fitted structural topic model.

regionData.csv

Dataset of countries and associated regions.

- *ccode:* The COW country code.
- *country:* The name of the country.
- *mapname:* The name of the country, as recorded in the ggmap package which produces maps.
- *globalsouth:* A binary variable indicating whether the country is considered part of the Global South (the Group of 77 and the People's Republic of China) (0=no, 1=yes).
- *geo_region:* The geographical region in which the country exists. The variable sorts all countries into one of seven regions: Americas, Asia, Eastern Europe, Middle East/Northern Africa, Oceania, Sub-Saharan Africa, and Western Europe.
- *geo_region2*: The geographical region in which the country exists, but separating South Africa, the Soviet Union, and Vietnam from their associated regions. This variable is used in the appendix to show that the main results are not solely driven by these countries, which had great significance during the temporal coverage of the PDB data.

ysiData.csv

Dataset tracking the number of years since independence (post-WWI) for all relevant country-years.

- *ccyear:* A single variable that combines a COW country code with a relevant year.
- *ysi:* A running count of the number of years since independence for the country in the year. Note that this count only applies to countries that gained independence following World War I.
- *decol:* A binary variable indicating whether the country was formally decolonized, as coded in the COW Project's Colonial History Data Set (0=no, 1=yes).

tropeData.csv

Dataset of racial tropes from supervised learning models, with several additional variables added through the use of "addFeatures.R," which itself uses data from "regionData.csv" and "ysiData.csv."

- *pdbID:* The document number given by the CIA.
- *date:* The date on which the PDB was distributed.
- *admin:* The administration in office on the day the PDB was distributed.
- *year:* The year in which the PDB was distributed.
- *entryID:* A unique identifier for the specific entry. This identifier is constructed by combining the *pdbID* variable with the *entryNum* variable.
- *entryNum:* The sequential number of the entry in a PDB.
- *title:* The title of the PDB entry, if there is one.
- *text:* The text of the PDB entry.
- *tt:* The combined text from the *title* and *text* variables.
- *globalsouth:* A binary variable indicating whether the country is considered part of the Global South (the Group of 77 and the People's Republic of China) (0=no, 1=yes).
- *logYSI:* A running count of the (logged) number of years since independence for the country in the year. Note that this count only applies to countries that gained independence following World War I.
- *decol:* A binary variable indicating whether the country was formally decolonized, as coded in the COW Project's Colonial History Data Set (0=no, 1=yes).
- *geo_region:* The geographical region in which the country exists. The variable sorts all countries into one of seven regions: Americas, Asia, Eastern Europe, Middle East/Northern Africa, Oceania, Sub-Saharan Africa, and Western Europe.
- *geo_region2*: The geographical region in which the country exists, but separating South Africa, the Soviet Union, and Vietnam from their associated regions. This variable is used in the appendix to show that the main results are not solely driven by these countries, which had great significance during the temporal coverage of the PDB data.
- *conf_high*: The total number of conflicts with intensity level 2 (war) involving the country mentioned in the entry on that given day, according to the UCDP/PRIO Armed Conflict Dataset.
- *democracy*: A binary variable indicating whether the country mentioned in the entry was a democracy or not in that year, according to the Polity 5 dataset. The binary variable is created by classifying all country-years with a Polity score of 6 or higher as democracies (0=no, 1=yes).

- *personal*: A three-level variable indicating the degree of personalist rule in the country in the year mentioned in the entry, derived from the *Autocracies of the World* dataset (0=not personalist, 1=somewhat personalist, 2=highly personalist).
- *leaderMention:* A binary variable indicating whether a head of state was mentioned in the entry (0=no, 1=yes).
- *leaderTenure:* The number of days that a head of state has been in office as of the date of the entry.
- *UStrade*: Levels of US trade export dependence on the country mentioned in the entry in the relevant year. This measure is taken from the replication data for "The Diplomatic Core: The Determinants of High-Level US Diplomatic Visits, 1946-2010) by Lebovic and Saunders (2016).
- *USmilaid*: Levels of US military aid to the country mentioned in the entry in the relevant year. This measure is taken from the replication data for "The Diplomatic Core: The Determinants of High-Level US Diplomatic Visits, 1946-2010) by Lebovic and Saunders (2016).
- *USdefense*: A binary variable indicating whether the US had a defense pact with the country mentioned in the entry in the relevant year (0=no, 1=yes). This measure is taken from the replication data for "The Diplomatic Core: The Determinants of High-Level US Diplomatic Visits, 1946-2010) by Lebovic and Saunders (2016).
- *nWords:* The number of words in the *tt* variable.
- *logWords:*
- *country*: The name of the country.
- *ccode:* The Correlates of War (COW) country code for the state mentioned in the entry.
- *pred[TROPE]:* The predicted level of a given trope, according to the supervised learning model which performed best in terms of Cohen's Kappa.
- *count[TROPE]:* The number of times any words/phrases in the trope dictionaries occurred in the entry for a given trope. For example, *countBELLIG* reports the number of words/phrases that appear in an entry reflecting the belligerence trope.
- *TopicN:* The prevalence of topic N in the entry, according to the fitted structural topic model.

stmLabelsData.csv

This dataset contains all the information necessary to analyze when and how leaders seek counsel during deliberations, as reflected in Section 3.1 of the main text. Recall that this dataset reflects a sample of randomly selected meetings for which research assistants qualitatively coded whether speech acts reflected dissent and/or information search.

- *segID*: A unique ID for the speech act, created using the meeting number, part number, and speech act number.
- *meetNum*: Meeting number.
- *admin*: The administration during which the meeting took place.
- *firstlast*: The first and last name of the actor responsible for the row's speech act.
- *dissent*: Whether the speech act expresses dissent (0=no, 1=yes).
- *infosearch*: Whether the speech act indicates any form of information search (0=no, 1=yes).

stmCoherenceValidation.csv

File containing data to perform topic coherence validation (random 4-word set intrusion task) for structural topic model.

- *group:* The number that identifies a unique set of words involved in a single task. The numbers range from 1 to 500, indicating the 500 tasks involved.
- *words:* The four indicative words drawn from a random topic.
- *topic:* The topic number associated with the four indicative words. Note that in every group, a single row will have a number that is different from the others. This is the intruder set of words. Note further that the human coder did not see these numbers when performing the task of identifying the intruder set.
- *guess:* A variable indicating whether the human coder believed the set of four words to be the intruder (1=yes).

stmLabelValidation.csv

File containing data to perform label validation (optimal labor task) for structural topic model.

- *group:* The number that identifies a unique set of words involved in a single task. The numbers range from 1 to 500, indicating the 500 tasks involved.
- *text:* The raw text of a single randomly selected PDB entry. Note that the full text only appears in the first row associated with each group.
- *topic:* A list of four randomly selected topics, with one topic being the highly prevalent one in the PDB entry text.
- *rightLabel:* A binary variable indicating whether the topic labels shown in *topic* is the correct (most prevalent) label (0=no, 1=yes). Note that the human coder did not see these numbers when performing the task of identifying the intruder set.
- *guess:* A variable indicating whether the human coder believed the topic label to be the correct one for the PDB entry (1=yes).