Just Jargon or Policy Priorities?
Text Analysis of Secretary of State Remarks for the Biden and Trump Administrations

Maggie Sullivan
Georgetown University, McCourt School of Public Policy
PPOL 5205: Data Science 3
Professor Brodnax
December 13, 2023

Summary

Text analysis of remarks attributed to the current U.S. Secretary of State (SoS) Antony Blinken (Biden Administration) and his predecessor, Secretary Michael Pompeo (Trump Administration) reveal both diplomatic jargon and policy priorities under their administrations.

Research Objective

For this research, I conducted both exploratory and predictive analysis on samples of remarks attributed to the current U.S. Secretary of State Antony Blinken (Biden Administration) and his predecessor, Secretary Michael Pompeo (Trump Administration). Public affairs materials point to both policy priorities and how we talk about those policy priorities. This is particularly relevant when comparing the priorities of presidential administrations from different political parties. The Department of State is the foremost agency responsible for making and executing U.S. foreign policy and the Secretary of State is the politically-appointed head of the agency. Through this research, I aimed to understand how government public affairs materials do or do not reveal key facts about U.S. foreign policy, particularly for changes between administrations of different political parties.

Data

My dataset consists of 1,973 documents I scraped from the webpage for the Office of the Secretary of State for Blinken (“Remarks: Secretary Blinken”) and Pompeo (“Remarks: Secretary Pompeo”). This includes 1,019 documents for Blinken and 954 documents for Pompeo. I selected one full calendar year for each Secretary because press announcements are often driven by recurring recognition days (such as World Day Against Trafficking in Persons or the national holiday of an ally). See Appendix I for a more detailed description of the web scraping process. Additional information in the dataset includes document title, date, and type of release (Remarks, Press Statement, Interview, Video Remarks, Remarks to the Press, Speech, FPC Briefing, Op-Ed, Readout, Special Briefing, or Other Release).

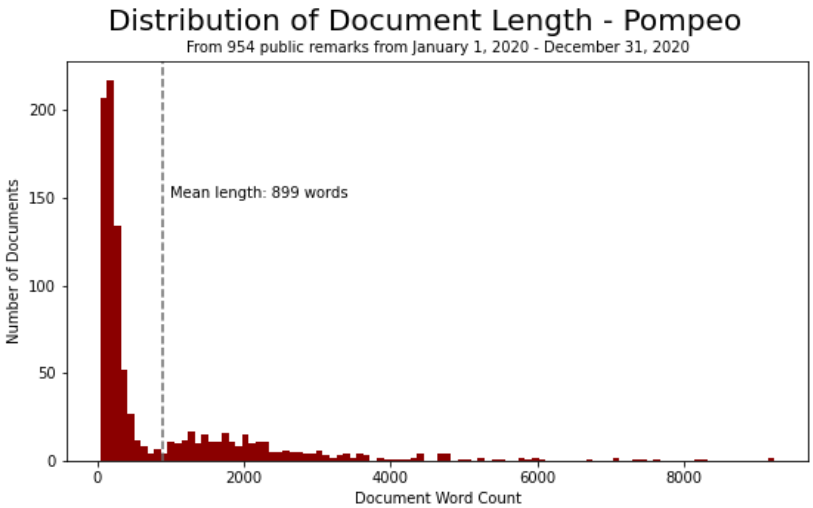
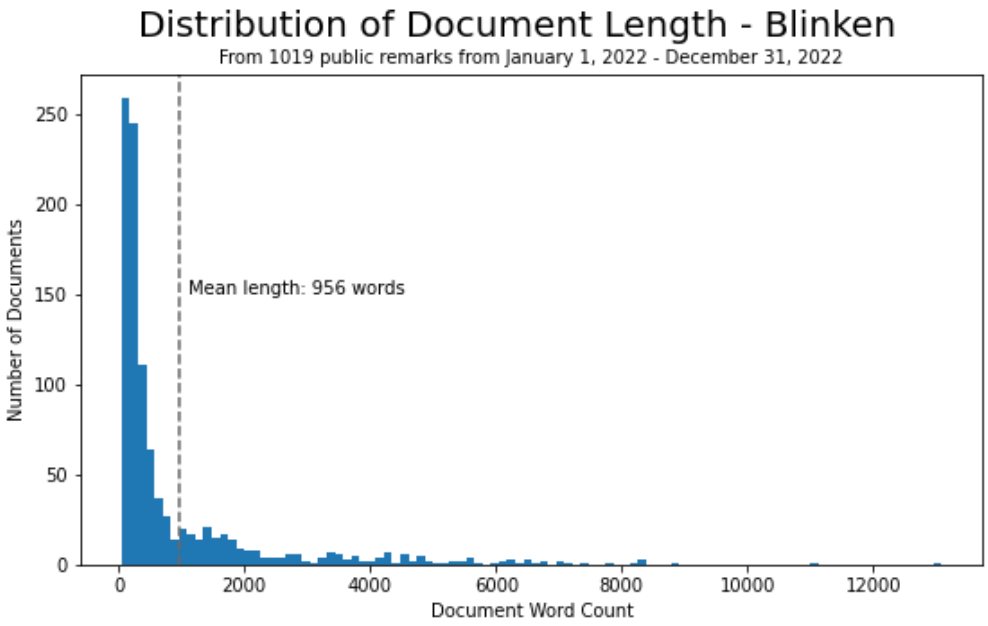


Figure 1: Distribution of Document Length (Pompeo)

Figure 2: Distribution of Document Length (Blinken)



Techniques Applied

Because my goal was to both understand the features in the dataset as they relate to the SoS in question as well as using these features to predict the SoS, I assessed the dataset both separately and merged with multiple techniques: 1) Term Frequency - Inverse Document Frequency or TF-IDF, 2) Principal Component Analysis or PCA, 3) K-Nearest Neighbors or KNN, 4) Tree-Based Methods (Decision Trees/Random Forest), and 5) Naive Bayes.

TF-IDF

As the basis for all my additional analysis, I constructed a TF-IDF for each dataset and the merged full dataset. I conducted hyperparameter tuning, including normalization (using the default L2 parameter), a max_df of , a min_df, lemmatization, limiting to unigrams and bigrams, and removing numbers, punctuation, and stop words (including typical english stopwords plus terms like United States, Pompeo, and Blinken).

PCA

Principal Component Analysis is a useful unsupervised learning technique that reduces dimensionality by maximizing the variance in the data. The first principal component explains the highest amount of variance in the data (James et al., 2021). This is useful for text analysis since our resulting TF-IDF has a large number of terms, with 1019 for Blinken and 954 for Pompeo.

Naive Bayes

As is common with text analysis, I used the supervised learning technique Naive Bayes to build a predictive model. This method works well in high-dimensional settings with sparse datasets. Naive Bayes uses Bayes' theorem of probability to classify values. For this research, I specifically used Multinomial Naive Bayes (James et al., 2021).

KNN

I used supervised learning technique K-Nearest Neighbors to classify my data. KNN, unlike Naive Bayes, does not assume the distribution of the data. KNN classifies based on the highest probability based on the class of n-number of neighbors (James et al., 2021). The Findings section includes more information about hyperparameter tuning.

Tree-Based Methods

Random forests are a non-parametric, supervised learning technique based on decision trees, which predict the outcome variable by splitting the data based on the independent variable values (resulting in an output resembling a family tree). Although the researcher can set hyperparameters such as max-depth or node purity to reduce overfitting, a useful technique is to employ random forests by building multiple, decorrelated trees on multiple training samples. (James et al., 2021). Using a random forest approach typically improves prediction accuracy while maintaining interpretability (James et al., 2021). I conducted hyper parameter tuning by establishing a max_depth value of 5 and a minimum sample for splitting value of 50 to maintain interpretability.

Findings

Individual Data Set Analysis

TF-IDF Top Terms

I constructed separate TF-IDF matrices for each of the two SoS datasets and isolated the top 100 terms based on the highest TF-IDF score. I then compared the top terms for each secretary to see which terms they shared and which terms were unique to the secretary. Both secretaries shared 74 of the same terms with each having 26 unique terms. This suggests that much of the time, press releases and materials from the Secretary of State contain similar diplomatic speak and pleasantries, regardless of the secretary or the administration. However, key regional or topical priorities are still evident based on unique terms. For example, the list of unique words for Pompeo include "china" and "iran" while Blinken's unique terms include "ukraine" and "russia." This may, however, reflect the current crises or state of global affairs at the time rather than their region of focus. However, funding and resources often are directed in this manner.

Another key insight is that Blinken's top terms include "climate," which demonstrates the emphasis on climate issues under his administration. See Appendix 2 for a full list of terms.

PCA

The first 5 Principal Components (PC) of each set explain less than 10% of the variance. PC1-5 of Blinken's data set explains 7.9% of the variance. PC1-5 of Pompeo's data set explains 8.34% of the variance. PCA essentially helps find latent patterns in the data to group terms into broader categories. Although the amount of variance explained by the first 5 components for each corpus was not meaningful, the top 10 terms for each PC for each Secretary of State do give significant insight into major priorities, such as China, Ukraine, international drug trafficking sanctions, or general military assistance. The similarity between the first PC for each SoS shows that niceties or diplomatic formalities exist in each administration. To focus more on policy priorities, a customized stopword list of diplomatic terms could be developed and utilized for future research. See Tables 1 and 2 for an analysis of the top 10 terms for the first 5 PCs for each Secretary. The italicized categories are my own interpretation.

Table 1: Top 10 Terms for first 5 Principle Components for Pompeo

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
	Interview terms	*General sanctions*	*China and human rights abuses*	*International drug trafficking*	*Narcotic reward program*
1	question	designated	hong	designation	million
2	secretary	designation	hong kong	narcotic	reward
3	get	action	kong	corruption	narcotic
4	think	pursuant	human right	section	assistance
5	know	department	chinese	reward	transnational
6	thing	regime	right	involvement	humanitarian
7	going	individual	freedom	significant corruption	bureau
8	want	sanction	human	immediate family	reward program
9	sure	treasury	beijing	department	transnational organized
10	well	entity	china	information	organized crime

Table 2: Top 10 Terms for first 5 Principle Components for Blinken

	PC1	PC2	PC3	PC4	PC5
	Interview terms	*Sanctions against Russia*	*Assistance to Ukraine in defense against Russia*	*Diplomatic formalities*	*Military and humanitarian assistance funding*
1	secretary	treasury	ukraine	minister	assistance
2	thank	department treasury	russia	thank	drawdown
3	question	ukraine	drawdown	foreign minister	million
4	think	action	defense	secretary	humanitarian
5	going	designating	equipment	good	billion
6	much	russia	assistance	inaudible	arm equipment
7	say	entity	defend	pleasure	pursuant delegation
8	said	pursuant	military assistance	much	billion since
9	want	department	ukrainian	foreign	military assistance
10	make	sanction	assistance ukraine	prime	department defense

Combined Data Set Analysis

Finally, I also explored Naive Bayes, KNN, and tree-based models (Decision Tree and Random Forest) to predict the SoS based on the combined TF-IDF matrix and, when possible, analyze feature importance. Table 3 includes a side-by-side comparison of the resulting scores from these models, with Random Forest performing the best prediction.

Table 3: Model Accuracy Scores

	Naive Bayes	KNN	Decision Tree	Random Forest
Score	0.815	0.818	0.744	**0.837**

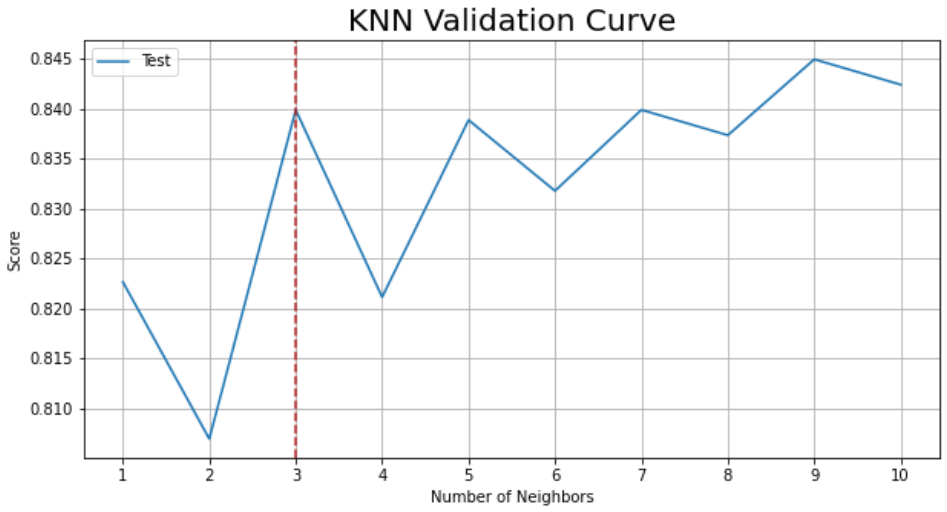
Naive Bayes

The third highest performing model was the Naive Bayes. I conducted this method with a 5-fold cross-validation and calculated the mean test accuracy. As Naive Bayes is commonly used in text analysis, I was surprised that this performed third out of the four techniques tested.

KNN

Using a five-fold cross-validation to test the number of nearest neighbors from 1-10, I determined that the k=3 would best maximize the score while keeping the model simple to avoid overfitting . Figure 3 shows the KNN validation curve.

Figure 3: KNN Validation Curve



Tree-based Methods (Decision Tree and Random Forest)

I conducted a Decision Tree classifier model with a max depth of 5 and a min_Sample_range of 50 (based on validation curves) to maximize test score and keep relatively simple to avoid overfitting the model and maintain interpretability. This resulted in a score of 0.744. See Appendix 3 for the full decision tree. Figure 4 shows the variables of importance (11 with values above 0). As previously mentioned, building random forests can help increase performance. I then built a random forest using 750 iterations and a maximum depth of 5, this time with no minimum sample range as I did not intend to review individual trees. I also used a five-fold cross-validation and calculated the mean R2 score 0.837 (the best out of all models attempted). Figure 5 shows the top 15 variables of importance. It is notable that China-related terms were among the top 15 variables of importance in the random forest, but were not present in the top factors of the initial Decision Tree. Overall, this model had the highest score for prediction.

Figure 4: Variable Importance (Decision Tree)

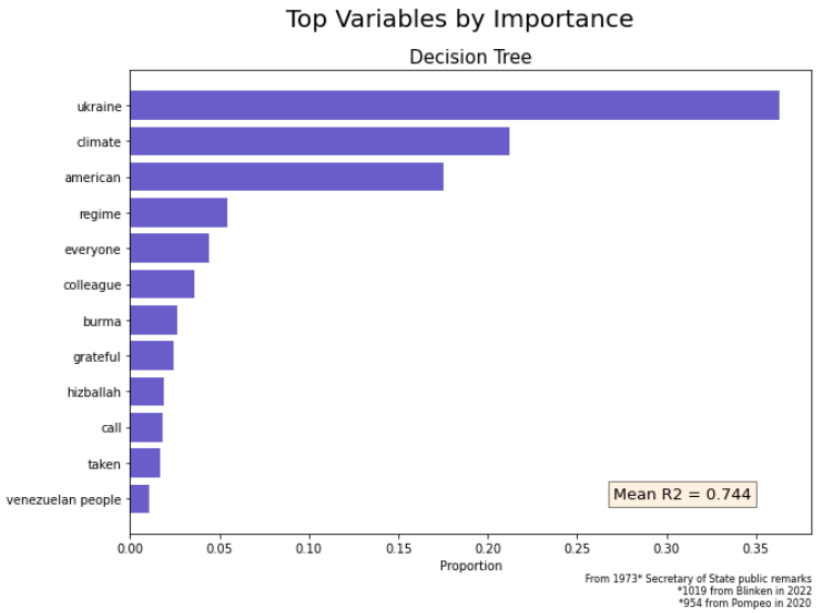
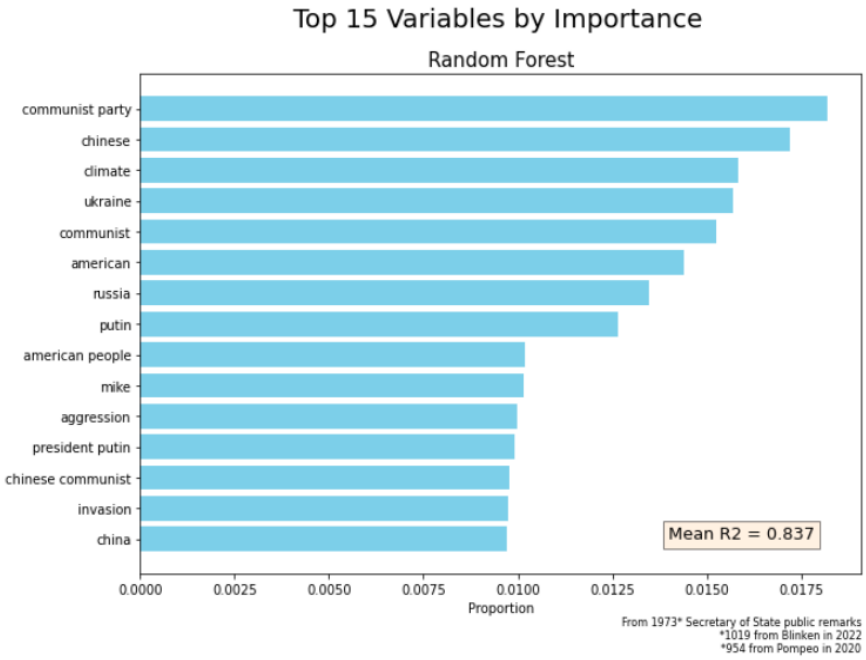


Figure 5: Variable Importance (Random Forest)



Conclusions and Future Considerations

In conclusion, my research found that while there are significant similarities between administrations on how they speak on public policy (think the diplomatic jargon and general language reiterating ongoing relationships), analyzing the corpus of each Secretary of State does allow us to isolate key regional or topical priorities. However, because so much of public remarks are time-sensitive, the data and success of the models can be highly influenced by major events (such as the Hong Kong protests and onset of the COVID-19 pandemic in 2019/2020 in China or Russia's invasion of Ukraine in 2022). Future research could explore the extent of this effect by pulling documents from multiple administrations to attempt to understand and classify by the administration's political party, rather than the individual Secretary of State.

Bibliography

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

"Remarks: Secretary Blinken." *U.S. Department of State*, www.state.gov/remarks-secretary-blinken/. Accessed 22 Oct. 2023.

"Remarks: Secretary Pompeo." *U.S. Department of State*, https://2017-2021.state.gov/remarks-secretary-pompeo/. Accessed 22 Oct. 2023.

Robots.txt. *U.S. Department of State,* https://www.state.gov/robots.txt. Accessed 22 Oct. 2023.

Appendix I: Implementation Appendix

After reviewing the *robots.txt* for the Department of State webpage (screenshots below), I scraped the documents using the Python *requests* and *bs4* packages with a time-delay of 5 seconds, first scraping all links, then iterating through the list of links to scrape the content itself.

Figure 1: Robots.txt for state.gov

```
crawl-delay: 5
# START YOAST BLOCK
# ---------------------------
User-agent: *
Disallow:

Sitemap: https://www.state.gov/sitemap_index.xml
# ---------------------------
# END YOAST BLOCK
```

Figure 2: Interface for the Secretary of State Speeches and Remarks for Blinken

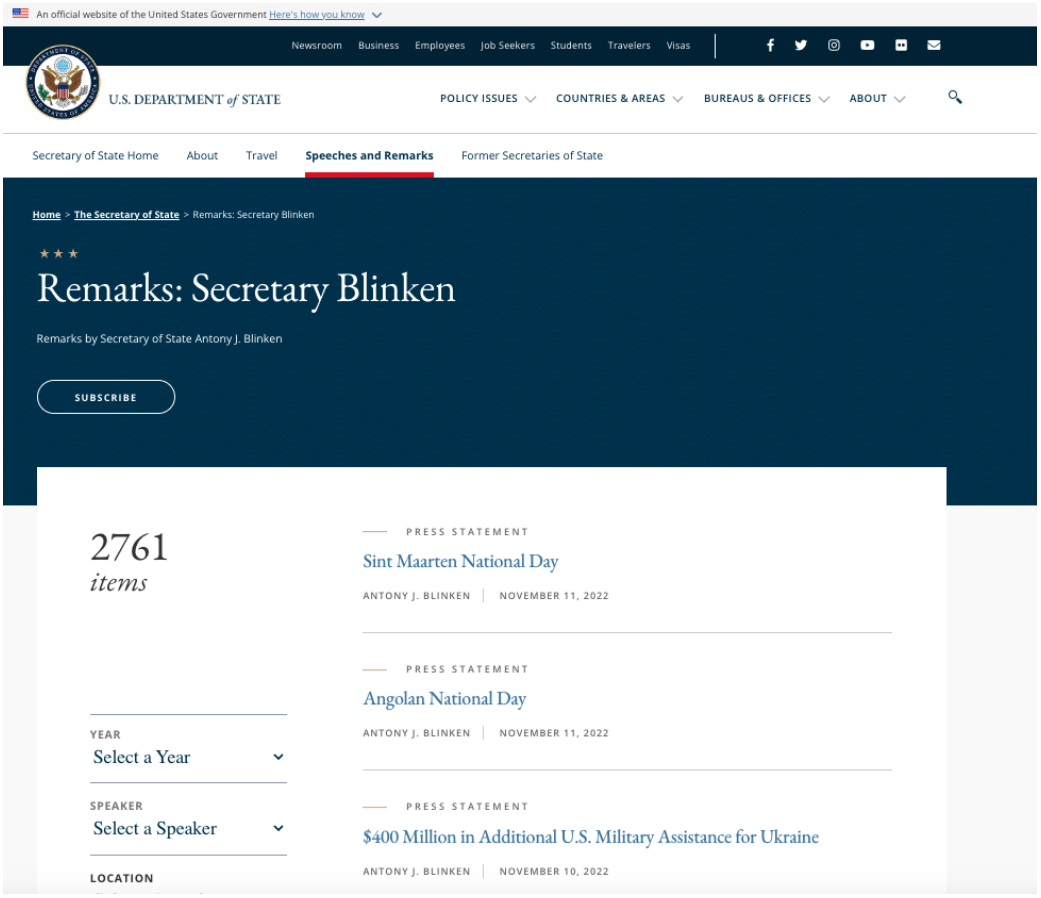
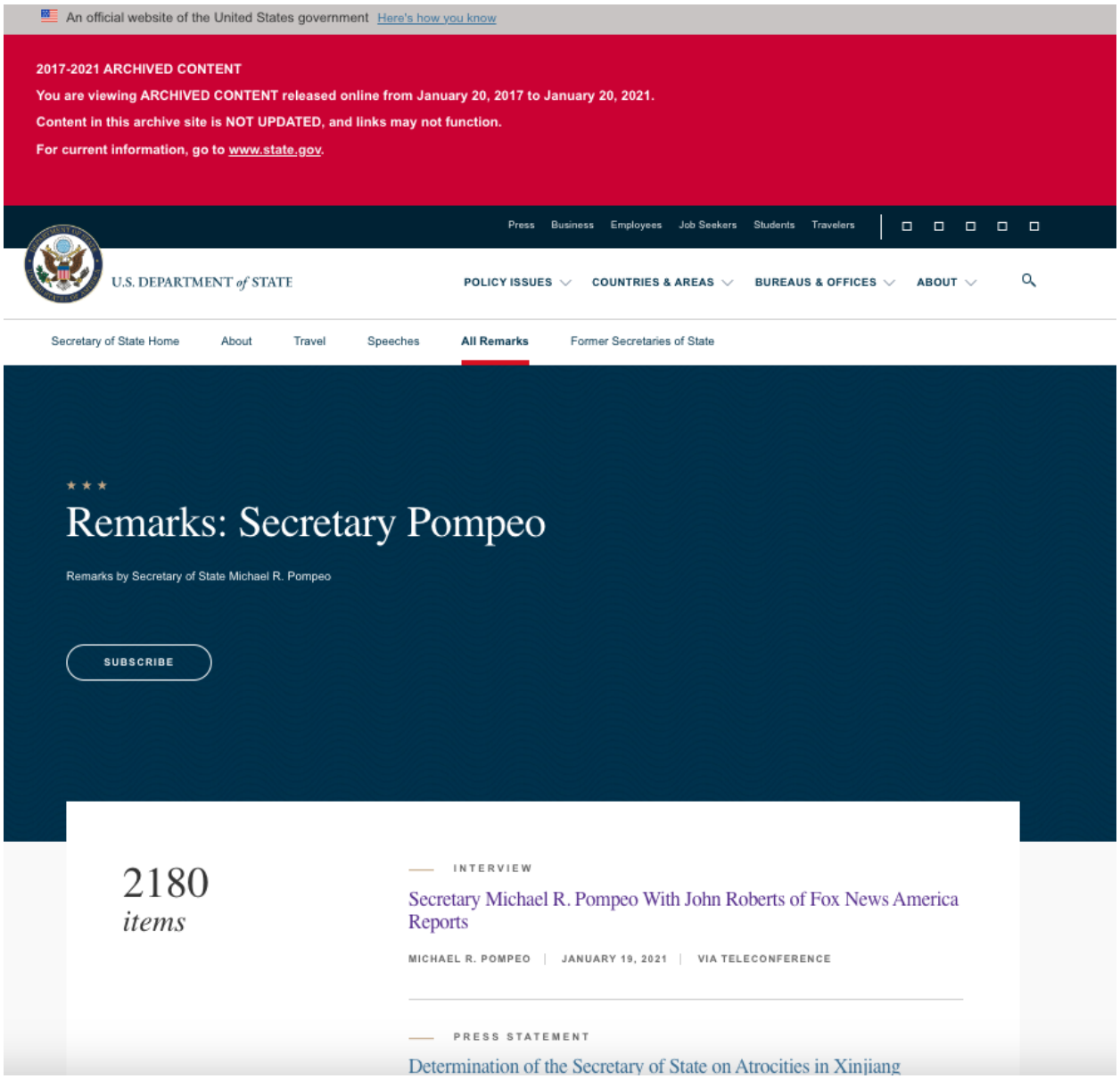


Figure 3: Archived interface for the Secretary of State Speeches and Remarks for Pompeo



Appendix II: Top Terms

	Both	Pompeo Only	Blinken Only
1	year	regime	ukraine
2	country	iran	russia
3	security	freedom	minister
4	secretary	china	much
5	thank	prosperity	russian
6	also	chinese	war
7	world	american people	future
8	together	behalf government	opportunity
9	support	party	ally
10	today	pandemic	welcome
11	day	made	issue
12	work	citizen	strong
13	government	free	climate
14	continue	anniversary	committed
15	president	get	community
16	right	place	aggression
17	partner	law	assistance
18	forward	congratulate	million
19	partnership	take	foreign minister
20	including	communist	stand
21	one	election	everyone
22	many	political	every
23	well	republic	say
24	time	iranian	behalf america

25	effort	back	part
26	look	communist party	help
27	department		
28	good		
29	international		
30	behalf		
31	america		
32	working		
33	foreign		
34	human		
35	challenge		
36	independence		
37	commitment		
38	global		
39	economic		
40	around		
41	look forward		
42	think		
43	important		
44	see		
45	wish		
46	shared		
47	human right		
48	democratic		
49	action		
50	relationship		
51	peace		

52	democracy		
53	make		
54	value		
55	new		
56	way		
57	region		
58	around world		
59	cooperation		
60	first		
61	like		
62	two		
63	celebrate		
64	come		
65	nation		
66	going		
67	american		
68	national		
69	want		
70	thing		
71	know		
72	question		
73	best		
74	health		

Appendix III: Decision Tree