Amine Elrayes, David Jun, Yilan Guo

Political Parties, Policy and the State of the Union

The State of the Union (SOTU) address discusses the current issues facing the country. With the SOTU address, the President has the opportunity to acknowledge the country's accomplishments, current events, the economy, and their legislative proposals. Although the SOTU is directed at Congress and the American public, it is not an entirely bipartisan address. Considering this fact, our research question is "Does the President's political affiliation have an effect on the policies that they put emphasis on in the SOTU address?"

This research question is significant because the SOTU address informs the general public. It is important to analyze the issues that are discussed because they affect voters' decisions and public sentiment. "Merely mentioning an issue in the State of the Union has the power to heighten the public's awareness of it." (Shogan) This determines future laws and policies that are enacted. As a result, we want to determine if a particular political party tends to push universalistic or particularistic policy. In other words, we want to determine whether or not a particular party tends to address the needs of every American, or if they tend to focus on issues affecting their constituents. With this research question, we also want to analyze the particular topics that Democrats find important and the topics that Republicans find important.

From a quick search, there are studies that answer our research question. However, these studies speak in generalities. These studies describe the tendencies of a general SOTU address but they do not analyze the individual speeches. While these studies take a more human approach to research, we take an algorithmic approach to answering our research question. In the paper written by Goedken, the author conducts a case study that compares the SOTU addresses of Donald Trump and Barack Obama. Goedken compares the way each president conveys their

policies but mainly focuses on the rhetoric of each speaker and its effects. With the paper by Shogan, there is research that implies presidents place emphasis on certain policies to advance their agenda and favorability. Shogan supports this with the fact that the public is more inclined to vote on a policy when the President urges them to. (See the works cited page to find the references to these studies.)

To examine the topic relationship between Democrats and Republicans, we adopted the Structural Topic Model approach and used a few libraries such as *tidyverse*, *tokenizers*, *quanteda*, and *stm*. Through the data cleaning process, we used a common bag of words pre-processing to create a document feature matrix including removing punctuation, numbers, stop words and word stemming. And similar to what we have done in previous classes, we trimmed the document feature matrix to only include words that appear in at least 5% of the documents. Because our interest focused on Democrats and Republicans, we filtered out other parties such as Nonpartisan and Democratic-Republican and only selected the speeches that belonged to the two parties. It reduced our data size from 236 to 188 but helped us to stay in the scope of the project's research question.

The Structural Topic Model (STM) is one of the unsupervised topic models in NLP that automatically generalizes the topic/cluster of the given text. To explore and decide the topic number best suited for the State of the Union dataset, we tried topic numbers of 10, 15, 20, and even 30. Although the more topics indicate a more fine and detailed group of clusters, most of the words become more similar when we increase the topic number, and the main topics we are interested in stay the same when we increase the number of topics. After increasing the number of topics to 20 and 30, we get topics that we are not interested in such as current events. Our research question hones in on strictly policy, and that appears with the specific number of topics

we chose. Given the relatively small dataset and the scope of our research question, we decided to adopt only 15 topic numbers to both avoid overfitting and focus on the events we are truly interested in.

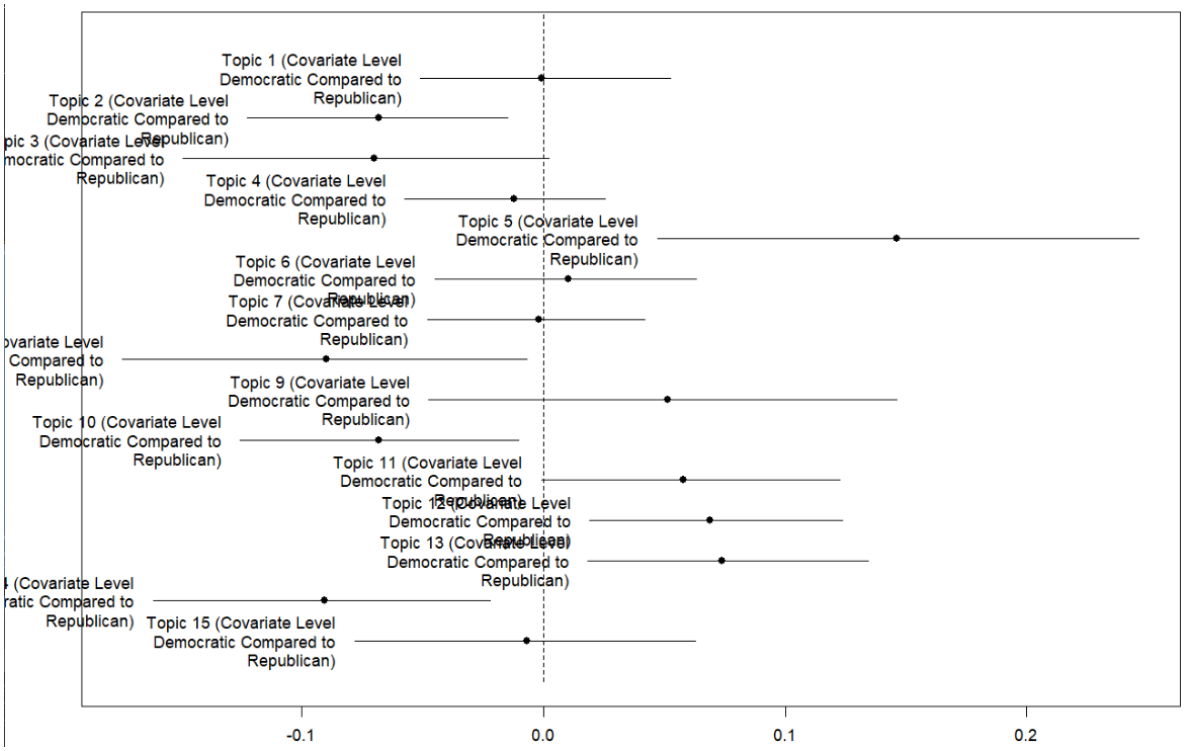
Finally, within the different numbers of topics, we then used the library *stm* to build our structural topic model, estimate using function *estimateEffect (method = difference)* and visualize the relationship between the two parties. In addition, we incorporated the smoothed time indicator in our model, which treated the year as a dynamic time trend instead of a linear one. The results of STM with 15 topic numbers are shown in Table 1. Surprisingly, there is no obvious difference between STM with and without smoothed time indicator. Each of the 15 topics has its unique label that is manually generated by us. Because of the special characteristics of the State of the Union speech, a few words such as *will, govern, and state* repeatedly appear in the highest probability category, meaning those words are likely to appear in every topic. To better understand the meaning of each topic, FREX is mainly analyzed in our project as it represents the words that are both frequent and exclusive. By combining the Highest Probability and FREX categories, each topic is paired with a label; for instance, topic 7 is about energy since it contains key words of *program, continu, oil, nuclear, and energ*.

Table 1: Comparison of STM with and without smoothed time indicator (topic number = 15)

	Highest Prob (no smoothed time indicator)	FREX (no smoothed time indicator)	Highest Prob (with smoothed time indicator)	FREX (with smoothed time indicator)
Topic 1: GOVERNMENT SPENDING	will, year, govern, program, nation, dollar, war	dollar, fiscal, expenditur, veteran, estim, billion, million	will, year, govern, program, nation, dollar, war	dollar, fiscal, expenditur, veteran, estim, billion, million
Topic 2: TRADE TARIFF	govern, state, unit, will, depart, congress, year	canal, panama, court, statut, tariff, appoint, arbitr	overn, state, unit, will, congress, depart, year	canal, panama, court, statut, tariff, appoint, arbitr
Topic 3: CRIME	will, year, can, american, america, govern, new	spend, percent, america, crime, let, ago, tonight	will, year, can, american, america, peopl, new	spend, america, crime, tonight, percent, let, ago
Topic 4:	state, govern, unit,	spain, cuba, spanish,	state, govern, unit,	spain, spanish, cuba,

SPANISH	island, will, author, congress	puerto, island, manila, hon	island, will, author, congress	puerto, manila, island, hon
Topic 5: MEXICO/SOUTHERN BORDER	state, govern, will, unit, upon, congress, public	mexico, texa, commenc, mexican, deem, constitut, whilst	state, govern, will, unit, upon, congress, public	mexico, texa, commenc, mexican, deem, constitut, whilst
Topic 6: ENVIRONMENT	will, year, program, feder, nation, develop, administr	environment, rural, reform, ensur, signific, administr, energi	will, year, program, feder, administr, nation, develop	rural, environment, reform, ensur, administr, signific, implement
Topic 7: ENERGY	will, nation, year, congress, state, program, continu	refuge, oil, afghanistan, soviet, nuclear, energi, deregul	nation, will, year, congress, state, continu, energi	refuge, oil, afghanistan, soviet, nuclear, persian, deregul
Topic 8: MONETARY SYSTEM	state, govern, year, will, upon, unit, congress	silver, circul, indian, pension, coinag, coin, commission	govern, state, year, will, upon, unit, congress	silver, circul, indian, pension, coinag, coin, commission
Topic 9: TERRORISM	will, american, year, america, work, peopl, can	that, terrorist, theyr, weve, kid, iraqi, dont	will, american, year, america, peopl, work, can	terrorist, theyr, that, kid, weve, iraqi, dont
Topic 10: CORPORATE JOBS	nation, law, will, govern, state, can, great	interst, corpor, man, forest, wage-work, supervis, standpoint	nation, law, will, govern, state, can, great	interst, corpor, forest, man, wage-work, supervis, standpoint
Topic 11: SLAVERY	will, nation, upon, govern, must, peopl, great	thought, manifest, passion, emancip, plain, gentlemen, seem	will, nation, upon, govern, must, peopl, great	hought, passion, manifest, emancip, plain, seem, hundr
Topic 12: COMMUNISM	will, nation, year, must, new, american, world	vietnam, communist, poverti, allianc, tonight, berlin, latin	will, nation, year, must, new, american, world	vietnam, poverti, communist, allianc, berlin, billion, south
Topic 13: WWII	nation, war, will, world, must, peopl, peac	japanes, victori, nurs, enemi, democraci, fight, german	nation, war, will, world, must, peopl, peac	japanes, victori, nurs, enemi, democraci, fight, german
Topic 14: AGRICULTURE	govern, will, nation, countri, congress, can, year	agricultur, relief, depress, board, consolid, method, bank	govern, will, nation, congress, countri, year, can	agricultur, relief, board, depress, consolid, method, bank
Topic 15: FREEDOM+DEMOCRACY	nation, world, will, must, free, peopl, can	communist, atom, free, freedom, strength, aggress, korea	nation, world, will, must, free, peopl, can	communist, atom, free, freedom, strength, aggress, korea

Figure 1: Topic difference graph (topic number = 15)



To better understand the content of each topic, we applied the *findThoughts* function on a few different topics, which outputs the most representative documents for a given topic. For example, Topic 5, as one of the most diverging topics shown in Figure 1, is calculated with a positive value, indicating that Democrats tend to favor it. Using the *findThoughts* function, the presidents of the top 10 representative documents are James K. Polk, Andrew Jackson, Martin Van Buren, and James Buchanan. Similar to what we predicted, those five presidents all belong to the Democratic party. In addition, the top 15 most representative documents' years are around the 1930s to 1940s, which is around when the Mexican-American War occurred. This further confirmed that the label (Mexico/Southern Border) we generated for topic 5 in Table 1 is accurate.

In opposition to topic 5, topic 14 has a rather negative value, meaning this topic is more likely to be discussed by Republicans. The result of the *findThoughts* function again proved our

prediction and label. The presidents of the top 10 representative documents are Calvin Coolidge and Herbert Hoover, who are both Republicans. Interestingly, the top 10 documents' years are between the late 1920s and early 1930s, which is the period of the farming problem of the 1920s after WWI (*The Farming Problem)*. Our hand-generated label for topic 14 is consistent with the historical events that happened during that time period and proved our label's credibility.

By creating topic models to answer and analyze our research question, we come to several conclusions based on the resulting data at the end of our algorithmic approach. After analyzing the plot, we can see that over nine of the topics are skewed towards the left of the dotted line, illustrating a topic preference that coincides more with Republican policy. In contrast, there are six topics that remain on the right of the dotted line, signifying topics that coincide more with Democratic policy preferences. Additionally, the table created shows all the highest probability words that would be said throughout each State of the Union address.

A topical model analysis helped display clear and concise information that substantially benefited our study of our research question. With this model, we were able to come to several conclusions based on the results shown. The results determined that political party does indeed have an effect on the topics mentioned within the address, though not a large one. The reason behind this is that presidents will often touch on non-partisan issues, or issues that don't cater to the preferences of a certain political party, in order to appeal to the general public. An example would be topics referencing the economy. Even so, there are results that determine whether a party does come into play. For example, speeches that favor issues such as climate change signify a higher appeal for Democratic support.

In order to designate the significance of our research question, we assess the validity of the process and its results. In terms of internal validity, there is a below average degree of

confidence due to several limitations that will be discussed later on. There however is high external validity as our sample size is large and cohesive while maintaining a wide distribution of dates that accommodate for different time periods. This information can be used to make credible predictions on the political preferences of the addresses of Presidents based on their political party in the distant past and future.

Through this approach, there exist a series of limitations when undergoing our analysis. Firstly, we only examined the addresses of presidents who identify with the Republican and Democratic parties. Third party presidents and presidents whose party affiliation may identify with non-polarized versions of American political parties are not represented within our analysis. Secondly, our dataset only contains State of the Union addresses up to the year 2016. Addresses that have taken place after 2016 are not accounted for. In addition, topical model analysis harbors a general limitation of being unable to wholly explain and represent a document like the State of the Union address due to the fact that it is more extreme than individual topics such as text analysis and political science. By incorporating both topics together in one analysis, it runs into issues such as repetitive distribution of high probability words that exist among every topic and difficulty in a more complete and accurate description of each.

Works Cited

Goedken, Danielle, "Presidential Agenda Rhetoric in the State of the Union Address: Obama and Trump" (2022). INSPIRE Student Research and Engagement Conference. 64.

Shogan, Colleen J. The President's State of the Union Address: Tradition, Function, and Policy Implications. , 2018. Internet resource.

“The Farming Problem.” *Ushistory.org*, Independence Hall Association, https://www.ushistory.org/us/49c.asp.