Investigating the rise of partisanship in the globalisation era of the US, 1989 – Present

Kush Kapur
Department of Computer Science
Durham University
ccbd24@durham.ac.uk

Submitted as a part of COMP3517 - Computational Modelling in the Humanities and Social Science

I. Introduction

One of the defining features of the globalization era in the US, has been the increasingly partisan behavior of the American electorate. The most recent public display of this being the 2020 presidential elections, where a mob of far-right republicans attempted a coup of the government to prevent the ratification of Joe Biden's election: 16 out of the 21 government shutdowns [1] occurring after 1980. Hence, it can be argued that rise in partisanship has led to dysfunctional government, evident from the recent government shutdown estimated to cost at least $11 billion as estimated by the congressional budget office [2]. This paper aims to investigate the rise of partisanship with the use of computational techniques on the US president speeches between 1989-present. President Speeches are used since the dataset since it does not only represent the president's ideology, but also their affiliated party. Speeches utilized as a vehicle to model the difference in ideology of the party in power across time. Finally, the reason of using speeches from varying time periods as comparisons is that a real time analysis between opposition parties would potentially yield biased results, because regardless of country the aim of the opposition party is to oppose the party in power, therefore by analyzing the rhetoric of a party when its in power generates a more accurate representation of their ideology and sentiments towards key issues facing the country.

II. Tools

I will be using python as the programming language, alongside libraries genism [11], scikit learn [13] and nltk [15] for purposes of data cleaning, topic modelling and feature extraction. For the purposes of sentiment analysis and document similarity I will use Vader [16], text-blob [12] and pre trained models from huggingface [14].

III. Data

A. Sourcing

The Miller centre [5] provides access to all US president speeches dating back to George Washington inaugural address. Due to their non-partisanship affiliation with the University of Virginia and reliability, the speech transcripts, summaries, dates, and titles are scraped for analysis. For web scraping operations beautifulsoup is used, and to handle automatic scrolling selenium.. The reason for using Beautifulsoup and selenium are because of their efficiency and functionality of these tools to be able extract large amounts of specific information in a very structured format, also the detail of documentation available makes it very easy to use.

For further information wiki data is scraped to get party affiliation and term time for each president.

Source Name	Information	Start date	End date
Miller centre	Transcripts, Summaries, Date, Titles	1789-04-30	2022-09-21
Wiki data	Party affiliation, Term time	1789	2022

Figure 1: Data sources

B. Preprocessing

Speech transcripts and Summaries (textual data), are tokenized by word, each token is assigned Part-of-speech tag, to prepare for lemmatization with the use of nltk Part-of-speech tagging and wordnet lemmatizer. The reason for Part-of-speech is that words are lemmatized differently based on their pos tagging (Fig 2). To make crucial information more prominent, I delete low-level information from our text by removing stopwords (sourced nltk) from the corpus [15]. Each speech is tagged by associated time eras as provided by the Miler centre [5].

IV. Exploratory Data Analysis

Bag-of-words analysis is performed by splitting the speeches by party. The document was then encoded into unigrams and bigrams to a document-term matrix using scikit learn count vectorizer. Any words appearing more than 90% of the documents during encoding were removed due to high frequency; this would not generate an accurate representation. To gain insight into the phrases that best reflected each era, I categorised each speech by time period. I then problem modelled this using a Naive Bayes classification problem. By investigating the feature coefficients within the model, I was able to identify the terms for each era that had the greatest

probability of term given era. Figure 3 represents best terms representative for the globalisation period.

```
nltk.stem.WordNetLemmatizer().lemmatize('loving')
✓ 0.6s
'loving'

nltk.stem.WordNetLemmatizer().lemmatize('loving', 'v')
✓ 0.5s
'love'
```

Figure 2: Difference in Lemmatisation with position of Tagging

Most frequent Republican terms	Most frequent Democrats terms
Security	Together
Law	Child
Law Enforcement	Health Care
Billion dollars	Small businesses
Armed Force	Middle class
American workers	Global Economy

Figure 3: Most popular unigrams and bigrams for Republicans and Democratic presidential speeches in the globalisation era

Figure 3 displays some key differences between the rhetoric of republican vs democratic presidents. For example, the emphasis on law enforcement, and especially American workers displays the key difference between recent anti-globalist populist republican policies such as America First [6] and Democrats pro-globalist views represented by trade deals such as the TPP [7]. Furthermore, the fact that health care emerges as a key term for democrats, but not for republicans could be linked to Michael Henderson's paper exploring the partisanship in healthcare [8], suggesting that the reason for the appearance of health care as a popular bigram for democrats is because of the disagreement between the parties with regards to universal health care, and the fact that health care is considered higher priority for democrats, figure 6

With several publications fact checking Trump speeches, I wanted to examine for myself with the use of text blob subjectivity and polarity scores. As seen from [9], Trump speeches are found to be the most polar and subjective out of all presidents in the globalisation era. This polarity could be an indication of the rise in hyper partisanship observed especially in the alt-right voter base of the US [10].

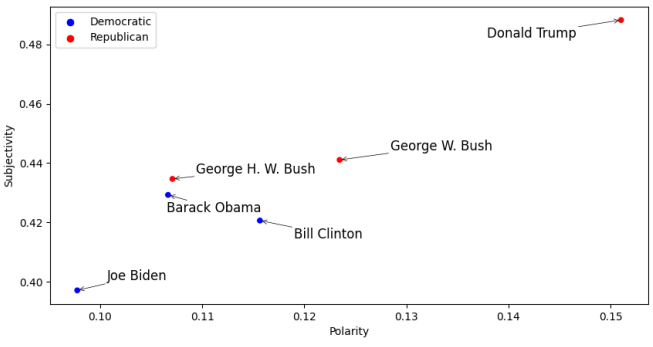

Figure 4: Subjectivity (Strength of opinion vs factual) vs Polarity of speeches for every president in the globalisation era

V. Model

Due to the number of speeches, and the time frame spanning two decades, it would be inefficient to find speeches covering similar topics, to make comparisons between both parties. In this paper I designed a filtering pipeline which employs several computational models to extract pairs of speeches from opposition parties covering the same range of issues. The aim of this is to use these filtered speeches to make a direct comparison of the rhetoric each party displays regarding key issues when in power, hence this research aims to evaluate if there is a case for negative partisanship present with regards to the opposition party.

A. Topic Modelling

After dividing speeches for opposition parties, a range of topic models were applied: Latent Dirichlet Allocation (LDA) [13], Non-negative Matrix Factorization (NMF) [13] and BertTopic [16]. LDA is implemented with genism [11] because of its Bayesian approach by using Dirichlet priors to estimate the document-topic and term topic-topic estimation. Due to this it can be applied to long documents which is the case for speeches, LDA also relies on the assumption that a document is a mixture of topics, suitable for our speeches. Similarly, NMF was implemented in genism [11] because of the same reasons as LDA, however it was implemented in addition of LDA because of its deterministic nature unlike LDA. One of the advantages of NMF over LDA is the fewer parameter choices involved in the modelling process, furthermore it has been proven capable of identifying niche topics that tend to be underrepresented in LDA [3]. From [4] it is evident NMF performs well on political speeches and text. BertTopic is also tried because LDA and NMF although great for longer documents, employ a bag of words model disregarding semantics, whereas BertTopic employs semantic embeddings for potentially more coherent topics. The fact that each document is assigned to one topic only meant that Bertopic was not suitable for my pipeline despite displaying good results, since the majority of speeches comprise of more than one topic. Overall, NMF performs better [4] determined by the topics being more interpretable and the initial coherence score for NMF being 0.45 is greater than LDA 0.39. This is potentially because NMF scales better than LDA, with the length of documents in this dataset that was key. The NMF model is tuned to maximize the word2vec coherence score, which trains a word2vec model on the corpus and measures how similar words belonging to a topic are with each other. Therefore, by adjusting the number of topics parameter we maximize the coherence score as shown in fig 8 and 9. I also noticed that TF-IDF vectorizer produced more diverse and semantically coherent topics in comparison to the count vectorizer when producing the document term matrix. By analyzing the topics in a temporal sense, I noticed immigration

is a major topic for republicans, especially referring to the southern border. From Fig 11, we deduce Immigration and tax cuts seemed to have become more popular topics with republicans since 2010s, possibly due to the Tea party movement in 2009 consisting of right-wing populist and conservative activists. As explored in [17] the rise of tea party supporters who are mainly republicans, showed a significant increase especially in the 2012 general election. It explores the hyper partisanship displayed by tea party supporters towards democrat policies. Looking at the federal budget topic, the republican policy of tax cuts and trickledown economics [18] can be inferred from the topic words. Topic 2 has also seen an increase since 2010s clear from fig 11 which consists of condescending words the current democrat president Joe Biden potentially displaying the rise of negative partisanship especially since 2010s [21].

This could refer to the increase in negative rhetoric towards the opposition since the election of Donald Trump.

In comparison, Democrat topic modelling suggests a focus on issues such as discrimination, and social welfare policies such as universal healthcare. From the difference in most popular topics for each party, we can infer the divide between republicans and democrats regarding what the nations top priorities should be. From the perspective of democrats, Race relations (Topic 3) take precedent, whereas republicans would argue towards Immigration. This is evident from research done by the pew research center as shown in Fig 6 [22], hence proving some social evidence for my topic modelling results.

Figure 5: Topic words and inferred overall topic for Democrats

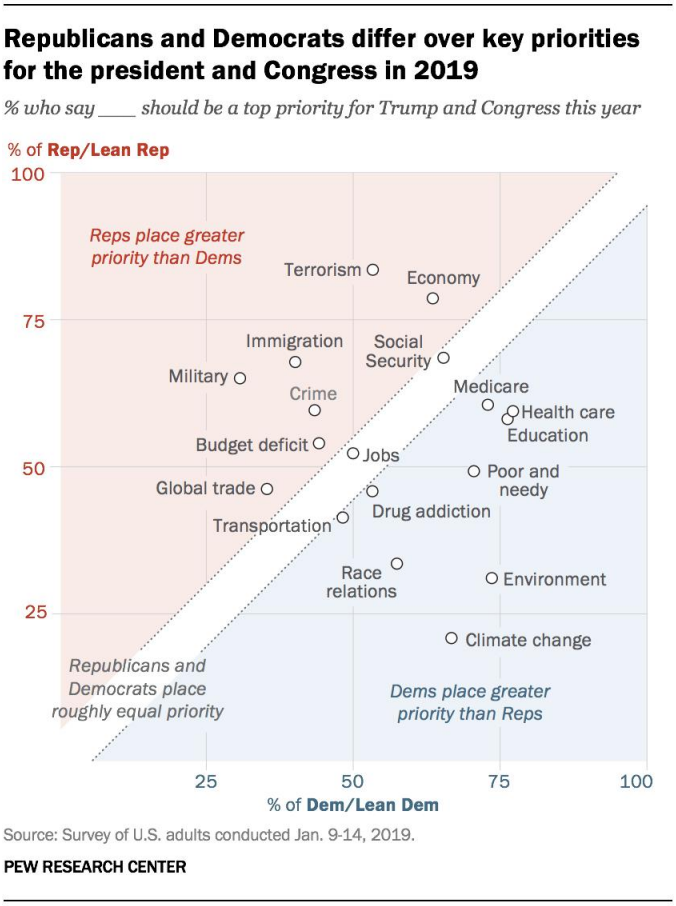

Figure 6: Pew research centre comparison of policy priorities between Democrats and republicans [22]

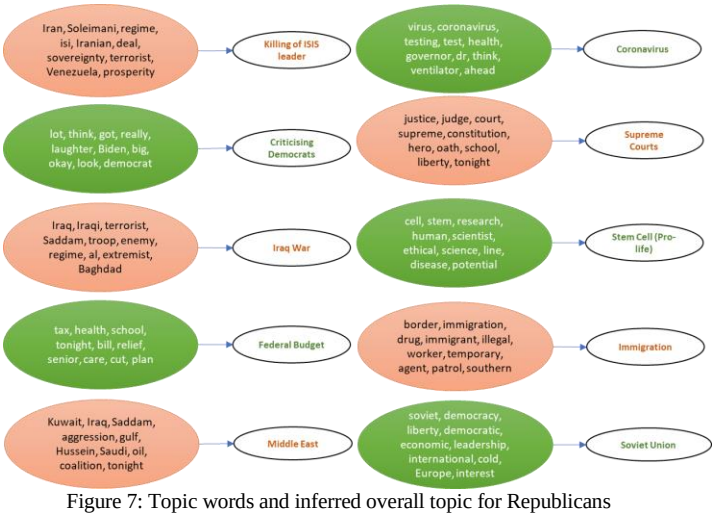

Figure 7: Topic words and inferred overall topic for Republicans

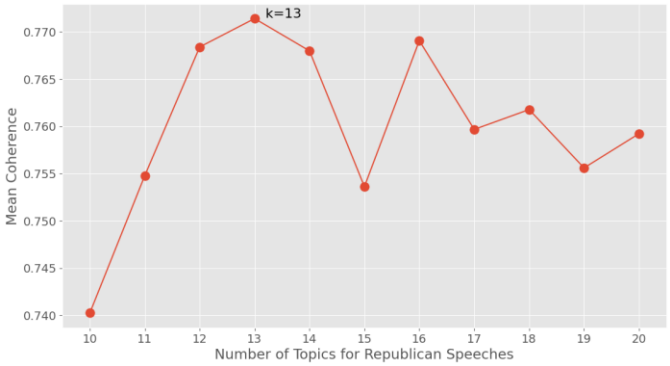

Figure 8: Fine tuning the Republican NMF model by altering the number of topics

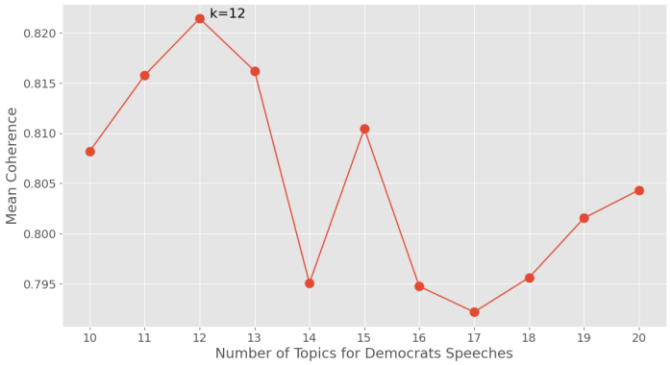

Figure 9: Fine tuning the Republican NMF model by altering the number of topics

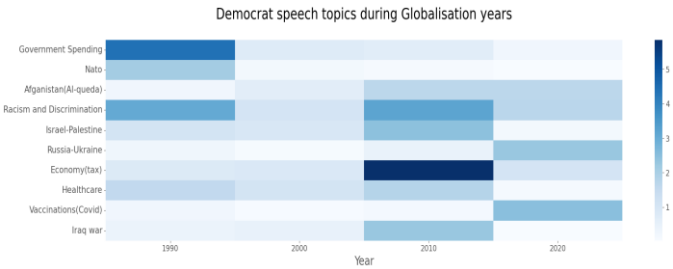

Figure 10: Distribution of inferred topics over time for Democrat Presidents

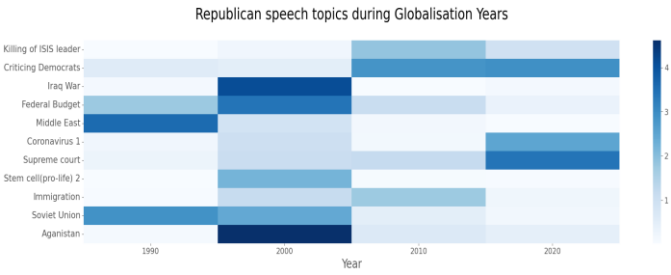

Figure 11: Distribution of inferred topics over time for Republican Presidents

B. Word embeddings

After topic modelling, we identify the most popular topics for both parties, and the words associated with those topics. The speeches are filtered by identifying the most similar topics between the opposition parties. To identify which democratic and republican topics are, the most similar I look at the topic words. The words associated with each topic is concatenated into a sentence and I generate a word embedding for each word. The cosine similarity is then used to identify how similar two topics are to each other. For word embeddings, I train a word2vec model on the corpus of republican and democratic speeches. The reason for this is that each word in the topic is standalone and doesn't have any attached context; this means that to generate a similarity metric the embeddings don't need to consider semantics, hence a word2vec model is the best approach as it generates a static word embedding

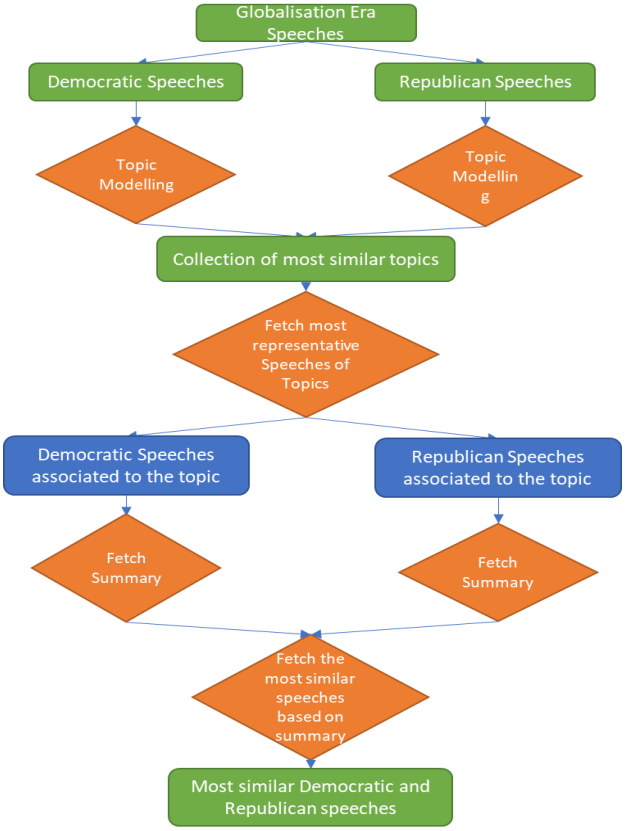

Figure 12: Filtering Pipeline architecture

for each unique word in the corpus regardless of the context. By training the word2vec model on our dataset it also means the model is fine tuned to be problem specific. For similarity the cosine similarity is used because it generates how similar words from two topics are likely to be based on their subject matter regardless of the length. Through this we find that Topic 1 from democrats and Topic 3 from republicans are very similar, as expected due to both topic words referring to government spending. Furthermore, the topic 1 from republicans and topic 9 from democrats are very similar as expected from their discussion about coronavirus.

C. Fetching speeches belonging to the most similar topics

Amongst the most similar topics, I take the eight most similar tuple of republican and democrat topics, where I create a mapping of topic-to-documents. This also includes the score for the association of a document with a topic where I return a list of ten documents with highest association with the most similar topics for each opposition party.

C. Fetching speeches belonging to the most similar topics

As a last filtering step, Miler center [5] provides a brief summary for each speech, which is used as a further filter the most similar speeches from opposition parties. Through this I construct a similarity matrix. which contains the score of speeches from opposition parties that are closest based on their summaries. To calculate this similarity a pretrained Bert based model all-MiniLM-L6-v2 [20] sentence transformer is used to generate an embedding for summaries of speeches belonging to the opposition parties. The reason for using this model, unlike Word2vec in the previous filtering stage, the summaries are full sentences, unlike standalone topic words. Hence, for similarity of the two documents the semantic information needs to be considered, for each word to ensure the right contextual embedding is generated. Furthermore, this model has been trained on variety of datasets, on more than 1.17 billion training tuples making it a very reliable model to generate contextual embeddings. Finally, for the similarity metric the cosine distance is used because it generates how two documents are likely to be based on their subject matter regardless of the length, which is relevant as some speech summaries are longer than the others.

D. Filtered speeches

From the filtering, as expected topic 1 from democrats, and topic 3 from republicans is found to be the most similar topics, which when applied through the pipeline generate the two most similar speeches as Donald Trump speech on the 2020-08-08 and Joe Biden's speech on 2021-04-28. Trump's speech refers to the disagreements with democrats regarding the covid-19 stimulus package, and Biden's speech is his first

Donald Trump Speech: 2020-08-08

Negative rhetoric towards democrats: In the current negotiations, we have repeatedly stated our willingness to immediately sign legislation providing expanded unemployment benefits, protecting Americans from eviction, and providing additional relief payments to families. Democrats have refused these offers; they want to negotiate. What they really want is bailout money for **states that are run by Democrat governors and mayors, and that have been run very badly for many, many years—and many decades.**

Nancy Pelosi and Chuck Schumer have chosen to hold this vital assistance hostage on **behalf of very extreme partisan demands and the radical-left Democrats, and we just can't do that**.

Coronavirus: But what the Democrats primarily want is bailout money. It has nothing to do with the **China virus**

Immigration: The Democrat bill includes stimulus checks for illegal aliens.

Democrat bill: includes stimulus checks for illegal aliens. They require the mass release of **illegal aliens** from detention

Tax cuts: Joe Biden and the Democrats may not want that. **They** don't want that because they're **adding $3 trillion in taxes.** We're further looking at **additional tax cuts, including income tax relief, income tax cuts, and capital gains tax cuts**

Figure 13: One of the filtered speeches obtained for republican

Figure 14: Corresponding filtered result for Democrat

address to congress and covers a range of topics from Immigration to climate change. Figs 13 and 14 display the key differences between the two speeches. Trumps use of negative rhetoric 'Illegal aliens' vs Bidens portrayal 'Undocumented folks'. By using a pretrained model Bert based model trained on 40k tweets, the term 'illegal aliens' can be classified as negative with 0.77 probability. The reason for using this model is that it is a fine-tuned version of distil-Bert on tweets, portraying a variety of training data. Similarly, by analyzing the difference between terminology used by consecutive presidents, the difference of opinion is clear between the two parties. Finally, the portrayal of negative partisanship is evident from the criticism of democratic run states and policies. This can be linked to [20] which explores the dysfunction in the US government due to partisanship. This is very evident from our filtered speech, where the covid relief stimulus was delayed because of partisanship. Additionally, this displays governmental

Joe Biden Speech: 2021-04-28

Climate Change: Look, the climate crisis is not our fight alone; it's a global fight. The United States accounts, as all of you know, less than 15 percent of carbon emissions. The rest of the world accounts for 85 percent. That's why I kept my commitment to **re-join the Paris Accord**

Gun control: **gun violence** is becoming **an epidemic** in America

Covid-19: Look, I also want to thank the United States Senate for voting 94 to 1 to pass the **COVID-19 Hate Crimes Act** to **protect Asian Americans and Pacific Islanders.**

Immigration: Immigration has always been essential to America. Let's end our exhausting war over immigration. If you believe in a pathway to citizenship, pass it so over 11 million **undocumented folks**

Tax Increase: I'm not looking to punish anybody. **But I will not add a tax burden—an additional tax burden to the middle class in this country**. They're already paying enough. I believe what **I propose is fair—(applause)—fiscally responsible, and it raises revenue to pay for the plans I have proposed**

dysfunction at its worst because this
was the peak covid period, where delaying a stimulus is harmful not only to the economy but several American citizens. Despite this partisanship restricted the rollout of this stimulus and displayed the inefficiency in ratification of key legislation due to negative partisanship.

VI. Validity of Results

Some weaknesses in the method may undermine the results of the study. NMF being a bag-of-words model, does not consider semantic relationships. Secondly, with regards to measuring similarity of summaries as the model is not trained on political text, its possible the embeddings generated by [20] aren't completely accurate. Lastly, the fact that NMF does not add a Dirichlet prior on top of the data generation process implies that it does not permit variation in themes. Finally, none of these topic models take metadata such as date into account which limits the ability to model the change of topics accurately over longer periods of time.

VII. Improvements and future works

As mentioned before NMF does not take metadata into account, hence in the future its possible to explore models such as Structured topic modelling which can take metadata for a document into account. Also, its possible to follow the two layered model outlined in [4] they define a method for combining the results of topic modelling from different time periods in order to identify a set of dynamic topics that span all or part of the corpus' duration.

VIII. Conclusion

This study explored a total of 206 speeches from US presidents between 1989 to present-day, to determine the rise of negative partisanship in US politics and its contribution towards government dysfunction. A collection of computational methods is used such as LDA, and word embeddings to create a filtering pipeline. The purpose of which is to output speeches from opposing parties spanning similar topics, in order to make a valid comparison regarding the ideologies of the two opposition parties. Finally, as explored throughout this study with the reference of previous social papers the success of the filtering pipeline, apparent from the two generated speeches which accurately define the polarized landscape of US politics.

References

[1] All 21 Government shutdowns in U.S. history https://www.thoughtco.com/government-shutdown-history-3368274J. History of Government shutdowns.

[2] CNBC reporting on the cost of last government shutdown https://www.cnbc.com/2019/01/28/government-shutdown-cost-the-economy-11-billion-cbo.html

[3] An analysis of the coherence of descriptors in topic modeling, Derek O'Callaghan, Derek Greene, Joe Carthy, Pádraig Cunningham http://derekgreene.com/papers/ocallaghan15eswa.pdf

[4] Greene, D., & Cross, J. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis, 25*(1), 77-94. doi:10.1017/pan.2016.7

[5] Miller Center

[6] Paul K. Macdonald, America First? Explaining Continuity and Change in Trump's Foreign Policy, *Political Science Quarterly*, Volume 133, Issue 3, Fall 2018, Pages 401–434, https://doi.org/10.1002/polq.12804

[7] Dade, Carlo and Ciuriak, Dan and Dadkhah, Ali and Xiao, Jingliang, The Art of the Trade Deal: Quantifying the Benefits of A TPP Without the United States (June 13, 2017). Canada West Foundation Trade and Investment Centre, June 2017, Available at SSRN: https://ssrn.com/abstract=2985355

[8] Michael Henderson, D. Sunshine Hillygus; The Dynamics of Health Care Opinion, 2008–2010: Partisanship, Self-Interest, and Racial Resentment. *J Health Polit Policy Law* 1 December 2011; 36 (6): 945–960. doi: https://doi.org/10.1215/03616878-1460533

[9] Beyond Fact-Checking: Lexical Patterns as Lie Detectors in Donald Trump's Tweets, Dorian Hunter Davies, Aram Sinnrei https://ijoc.org/index.php/ijoc/article/view/15397/3245

[10] Summers, Ryan T., "The Rise of the Alt-Right Movement" (2017). *Media and Communication Studies Summer Fellows. 11*. https://digitalcommons.ursinus.edu/media_com_sum/

[11] Gensim Documentation — gensim (radimrehurek.com)

[12] Textblob TextBlob: Simplified Text Processing — TextBlob 0.16.0 documentation

[13] Scikit learn scikit-learn: machine learning in Python — scikit-learn 1.2.0 documentation

[14] Hugging face Hugging Face – The AI community building the future.

[15] Nltk NLTK :: Natural Language Toolkit

[16] Bertopic BERTopic (maartengr.github.io)

[17] Partisan Polarization and the Rise of the Tea Party Movement, Alan Abramowitz Emory University https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1903153

[18] The Republican Devolution: Partisanship and the Decline of American Governance Hacker, Jacob S., Pierson, Paul Page 42 https://heinonline.org/HOL/Page?handle=hein.journals/fora98&div=73&g_sent=1&casa_token=

[19] All-Mini-LM-Bert based pre trained model for sentence embeddings https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[20] Alan I. Abramowitz, Steven Webster, The rise of negative partisanship and the nationalization of U.S. elections in the 21st century,Electoral Studies,Volume41,2016,Pages122,ISSN02613794, https://doi.org/10.1016/j.electstud.2015.11.001. (https://www.sciencedirect.com/science/article/pii/

[21] Partisan Conflict and Congressional Outreach https://www.pewresearch.org/politics/2017/02/23/partisan-conflict-and-congressional-outreach/