Politics on Wikipedia

Joseph Del Val | Iakov Vasilyev | Cameron Thomas
jdelval@ucsd.edu | ivasilie@ucsd.edu | cat028@ucsd.edu

ABSTRACT

This paper seeks to analyze the degree and prevalence of political bias and controversy in Wikipedia. Using pre-trained models from Rheault and Cochrane (2019) and Shapiro and Gentzkow (2019) we validate our methods for generalizability on the ideological books corpus (Sim et al., 2013) with sub-sentential annotations (Iyyer et al., 2014) and attempt to apply these methods to receive insight into political bias in Wikipedia. We attempt to combat overlap in political slants and avoid labeling political bias whose detection is unavoidable due to the topic of the article in question. With insight into political bias on Wikipedia gained we hope it will be able to prove useful in combating counterproductive activity on Wikipedia and allow for more precise and targeted activity by Wikipedia monitors.

INTRODUCTION

As established by Wikipedia itself, edit-warring is remarkably counterproductive and only makes consensus harder to reach. In *Edit Wars in Wikipedia*, Robert Sumi et al devised an M-Statistic which can grant any Wikipedia article a value representing its level of controversy; while it can quickly and effectively identify highly controversial articles, it is generalized to take into account any type of edit war (among other limitations), with an accuracy that is far from perfect. In general, this project seeks to address two key deficiencies in this method of conflict detection: scope of controversy and limitation in methods. While the aforementioned method was generalized for any and all edit wars across all topics, this project will focus on political controversy; additionally, our method will detect bias using page content and not just meta-data like the M-statistic.

The rationale behind focussing on political controversy is twofold. Firstly, unproductive political controversy and the resulting potential lack of accurate information is known to have severe consequences, and these consequences are particularly salient in these current times. As seen in Greenstein and Zhu's paper in 2018, bias in Wikipedia is indeed present, and it is both in Wikipedia's interest and in the interest of the general public for it to be as close as possible to a state of political neutrality and factuality. As a result, lowering controversy in this area becomes particularly salient. Political bias could be a particular method of targeting this— politically charged language is for one unhelpful, but additionally can provoke the other side and lead to additional controversy. Finding a way to neutralize politically charged language could then be helpful in efforts to quell political controversy and focus on neutral, factual information. As for the second rationale, politically biased language has excellent tools available and has already been a topic of study. Identifying ideological language is something that has already been done before; for example Rheault and Cochrane in *Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora* successfully uncovered ideology within digitized parliamentary debates.

There is, however, a lack of Wikipedia-focused bias research, which is unacceptable considering the importance and popularity of the website. Wikipedia itself only mentions three major papers written on the subject of ideological bias: Gentzkow and Shapiro (2012), Greenstein, Zhu, and Gu (2016), Greenstein and Zhu (2018). Upon further examination, the models used for those papers were trained on non-Wikipedia data, which made us question the validity of their findings. Wikipedia is completely unlike any other data source when it comes to its data generation process, and therefore it is hard to tell whether a model trained on newspapers or congressional speeches would produce valid results when applied to Wikipedia articles. The above mentioned papers address this problem in their own ways, for example, Greenstein and Zhu apply their model to both Wikipedia and Encyclopedia Britannica and perform comparative analysis, which reduces the potential harm of model overfitting. However, since we wanted to focus our research on Wikipedia, we attempted to mitigate the problem by using two different models and comparing their performances on the same set of data.

We apply the models to a subset of articles and measure the ideological bias of the current versions as well as revision histories of those articles in order to gauge the level of ideological slant across topics and throughout time.

DATA

For this project we had three main sources of data. The two models we employed were both trained using data from transcripts of congressional speeches. Most studies we found on measuring political/ideological bias use congressional data: there is a lot of it and it is easy to get a political slant label on every speech by identifying the political affiliation of each speaker. This data was then transformed once by Shapiro & Gentzkow (2019) with available two-word phrases and their political association, and second by Rheault and Cochrane (2019) with pre-trained models available on GitHub https://github.com/lrheault/partyembed for download. As mentioned above, there is a real risk of the models being overfit to congressional speeches, so we had to acquire a validation dataset that was generated by a different process than the congressional dataset. Ideally, our validation set would come from Wikipedia itself, however, we could not come up with a way to algorithmically extract labels from Wikipedia data. Our best attempt was to look for article edits tagged with comments containing the word "bias", under the assumption that such edits point out and replace ideologically slanted phrases with more neutral language. However, this approach turned out to be too inconsistent, so we had to find some other dataset with bias labels.

While looking into previous ideological slant research we found a rigorously compiled dataset called the Ideological Books Corpus (IBC) (Sim et al., 2013) with sub-sentential annotations (Iyyer et al., 2014). IBC is a collection of sentences labelled with left/right/neutral political ideology compiled from books and magazine articles by authors with well-known political leanings. Initially, we intended to use this dataset to train our own bias-detection model, however, due to the careful compilation process, IBC contains only around 4300 total sentences (2025 liberal sentences, 1701 conservative sentences, 600 neutral sentences) which is too little data for a new model, so instead we used it to validate the generalizability of the models trained on congress speeches. While there is no reason to assume that the phrases from IBC are in any way more representative of the "Wikipedian dialect" than the congress speeches, the data generation process is still different enough for us to be able to spot overfitting. This dataset was downloaded from https://people.cs.umass.edu/~miyyer/ibc/index.html, with sample data available publicly for download and the full dataset available via email. If one were to reproduce our experiments, they must first email the address posted on the website and request access to the full dataset.

The third dataset was extracted from our main source of interest, Wikipedia. At first we wanted to analyze the entirety of Wikipedia, however, with over 6 million articles to consider we were risking infeasible runtime lengths for our timeline. Besides, our models were trained on data pertaining to U.S. politics specifically, so the results on unrelated articles would have been even less trustworthy. Therefore, we decided to only focus on U.S. politics-related articles. Wikipedia's category system is inconsistent and none of the previous approaches to this task were usable for different reasons, so we had to find some other method or resource to help with the selection process. We ended up settling on the list we found on a U.S. Politics "task force" page. Task forces on Wikipedia are voluntary collaborations focused on improving different parts of the website, and the task force we got the list from specializes in Wikipedia's coverage of the U.S. Politics. The list we got from their dedicated page contained the top 700 most-viewed U.S. politics-related Wikipedia articles, which served our purpose well for two reasons. First, it was small enough so that we could be flexible about which models we used and which statistics we generated since our code did not take too long to run. Second, the more popular pages are usually more developed, so we were able to study the results of more active collaborations, which is what Wikipedia was intended for. We scraped the current versions of almost all the articles from the list (with few exceptions such as list articles), and then manually picked out 9 of them for further examination of their revision histories based on length (short-medium-long) and current bias scores (left-neutral-right). We used article length because we found it to be the best available estimator for the number of edits, or how well the article is developed, and got edit counts ranging from 800 to 8000.

One of the aforementioned models we're using, the pre-trained model by Rheault and Cochrane, is available on github as partyembed. A key function of this model that we are able to use for our task is the "issue" function- this provides us with data created by their trained model. The vocabulary of this model is associated with different weights of positive or negative democratic and republican slants, from congressional speeches from every two years from 1873 to 2015.

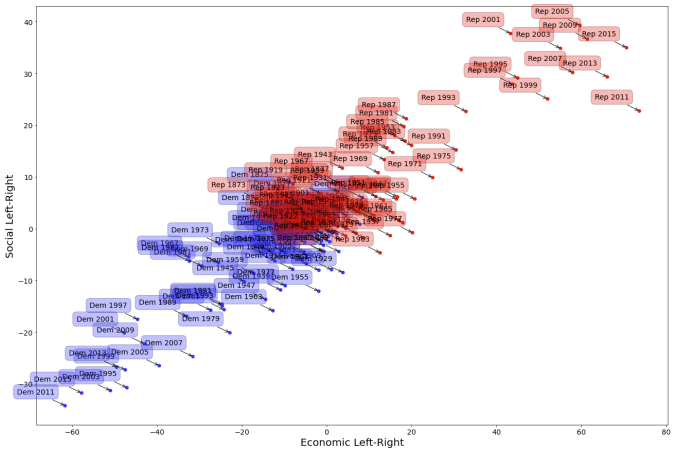


Fig 1. Overall Slants of congressional speech data from each party for each year.

Ultimately, belief in the efficacy of this data is relatively certain, as the model performs excellently at distinguishing the ideological placement of its corpora. However, what we intend to do is analyze if this data can be generalized, which will be explored further in Methods.

METHODS

Wikipedia is one of the most unique data sources out there just by the nature of the data generation process. The articles are written by the combined efforts of tens and hundreds of different editors, each with their own writing style and ideology, which automatically brings up the question: are models trained on data from congress speeches, or books, or magazines general enough to be applied to Wikipedia data? As mentioned above, validating the performance of the models on Wikipedia data is not a trivial task, and we did not find the previous research to address this problem sufficiently. Therefore, we are employing two different methods of validation. Firstly, we use the IBC to assert both models' generalizability by checking how accurate they are at labelling the sentences from the dataset. Secondly, since we are using two different models, we are able to compare their performances on the Wikipedia data, which allows us to make sure the methodology does not affect the results too drastically.

The first model we use was developed by Gentzkow, Shapiro and Taddy for their paper *Measuring Group Differences in High-Dimensional Choices.* The model uses a neural net to assign bias scores to two-word phrases, or bigrams. We then use the resulting score dictionary to assign bias scores to bigrams in the selected Wikipedia articles. While there are many different ways to map the resulting array of numbers to a single representative value, we decided to go with summing all the bias values together, effectively getting the formula:

$$score(a) = \sum_{x \in S(a)} freq(x) * bias(x)$$

Where x is a unique word/bigram in the list of words/bigrams S derived from a Wikipedia article a. The article text was pre-processed the same way Gentzkow, Shapiro, and Taddy pre-processed the congressional speeches in order to ensure consistency. That includes removing punctuation and stopwords, lowercasing, and porter-stemming the whole text.

At this stage we already spot the first signs of unreliability: about 22% of the articles have one of the top 10 frequent phrases in the title. That means that there is a possibility of the results being skewed by the topics of the articles. So, for example, if the phrase "San Francisco" is considered left-biased by the model, an article about San Francisco will receive a high bias score even if the language used is not biased. Upon further investigation, we found that the top 10 most frequent phrases constitute around 42% of the article's absolute bias score on average, while also being thematically connected to the article's topic. A telling example of this phenomenon is the article for Martin Luther King (fig. 2).

	phrase	count	bias	abs_score
0	civil right	80	-1.326589	106.127090
1	right movement	25	-28.909571	722.739267
2	nativ american	14	-14.807184	207.300581
3	right act	14	-89.486734	1252.814270
4	american co	10	5.745237	57.452370
5	high school	8	7.639545	61.116360
6	year old	7	-24.968011	174.776074
7	poor peopl	6	-12.152022	72.912134
8	vote right	5	-148.596026	742.980132
9	african american	5	-150.267981	751.339904

Fig. 2. Table of the most frequent phrases for the MLK article with counts, bias scores and absolute total scores.

As can be seen from the figure, the top 10 most common words are connected to MLK's biography. In this particular case their combined score constitutes over 56% of the whole article's absolute score. We call this propensity of an article's score to be skewed by topically connected words and phrases "topic bias". In order to combat topic bias we decided to ignore the top 10 most common phrases while calculating the articles' scores, the formula effectively becoming:

$$score(a) = \sum_{x \in S(a),\, x \in T(a)} freq(x) * bias(x)$$

Where T is the top 10 most common phrases for an article a. This method of dealing with the topic bias is only one of many possible ways, the problem is deep enough to warrant a whole another investigation. However, even with this solution about 27% of the articles ended up reclassified, which could mean that this method is at least somewhat effective.

Additionally, we decided to normalize the articles around their length by dividing the total score by the number of words the article contains. Although the data does not show a strong correlation between the article's score and its length (fig.3), we still decided to make that adjustment since it might help to further combat topic bias: the longer articles may still contain more topically-skewing words than the shorter articles. The final formula we went with is:

$$score(a) = \frac{\sum_{x \in S(a),\, x \in T(a)} freq(x) * bias(x)}{len(S(a))}$$

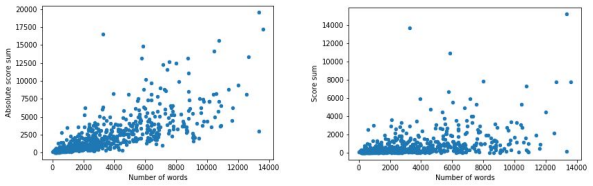


Fig. 3. Scatterplot of absolute sum over number of words (left) and scatterplot of non-absolute sum over number of words (right).

The second model we analyzed uses the issue() function from Rheault and Cochrane's *Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora* this model was created by, in using each word in each sentence in our validation set (The Ideological Books Corpus (Sim et al., 2013) with sub-sentential annotations (Iyyer et al., 2014)), extracting the weights from house corpora from 2007 onwards from the pre-trained model. Ergo, for each word in each item in the ideological books corpus, if this word existed in the vocabulary of Rheault and Cochrane's pre-trained model, we received the democratic and republican total leanings for each year. After applying this to one particular sentence, we are then left with an array of values. With these, we were then able to explore different aggregate functions for each sentence.

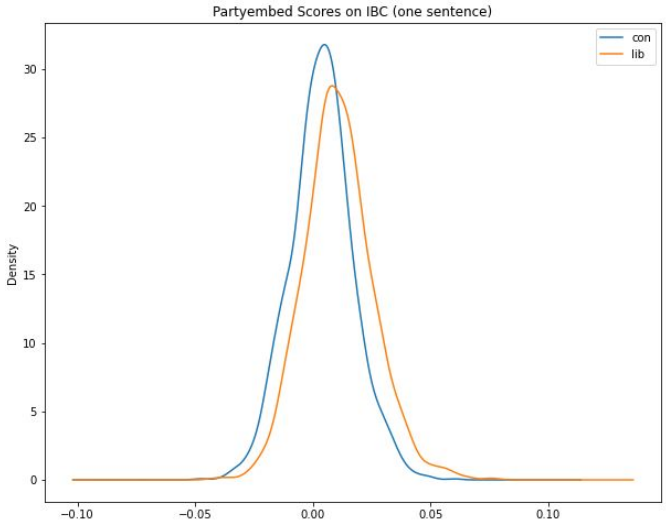


Fig 4. Mean aggregate function performed on R&S Slants from 2015 house corpora on ideological books corpus data, 1 sentence

Overall, when we looked at individual sentences, we found that there was a concerning amount of overlap. While the scores did indeed differ in the correct directions, the overlap itself was quite worrisome. However, under the justification that most Wikipedia articles have multiple sentences, we tried the same model with samples of multiple sentences.

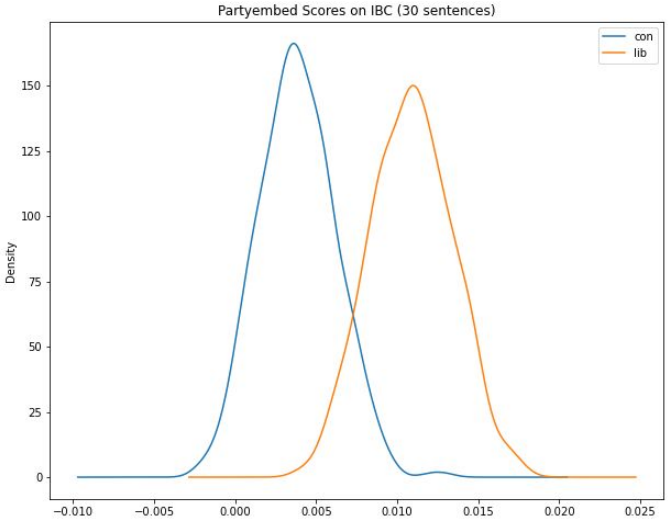
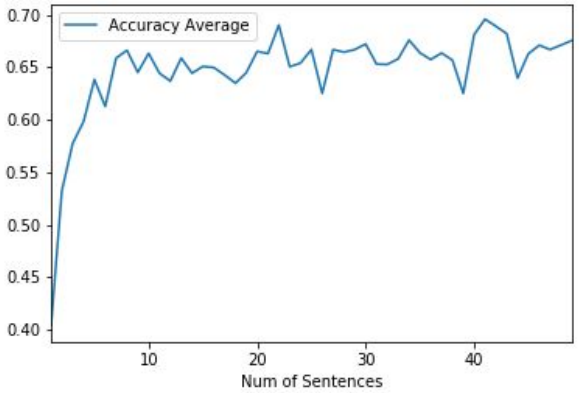


Fig 5. Mean aggregate function performed on R&S Slants from 2015 house corpora on ideological books corpus data, 30 sentences

The Gentzkow/Shapiro/Taddy model performed similarly when applied to the IBC: with only 1 sentence inputs we got an average classification accuracy of 40% (worse than even picking at random), however, as we fed more sentences into the model we started getting way better results, the average spiking up to almost 70%. It could be argued that 70% accuracy is not very reliable, but that is why we have a second layer of validation with inter-model comparisons.



Overall, we chose the mean as our aggregate function, as other aggregate methods, such as the Maximum, often had strange formations in the distributions

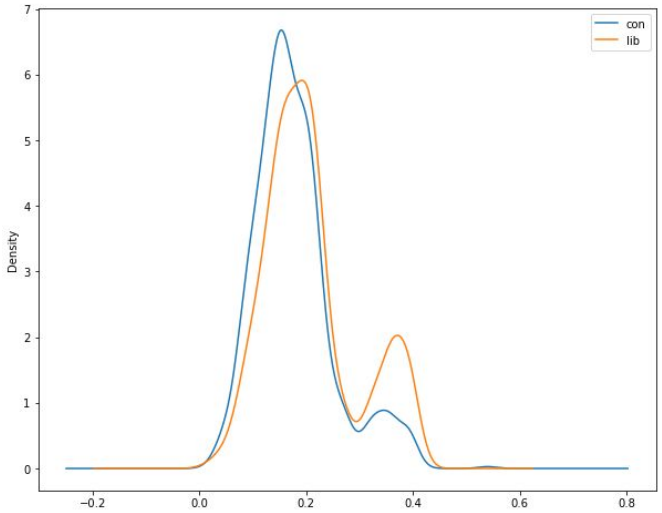


Fig 6. Max aggregate function performed on R&S Slants from 2015 house corpora on ideological books corpus data, 1 sentence.

RESULTS

When analyzing the results of the two models on current page articles, our first objective was to explore the models' similarities and differences. We first noticed a moderate correlation between the output scores of the two models, that is, 0.285213. When looking further into the differences, we found that overall the differences were somewhat normally distributed, with Partyembed overall giving somewhat more liberal scores, resulting in a right skew.

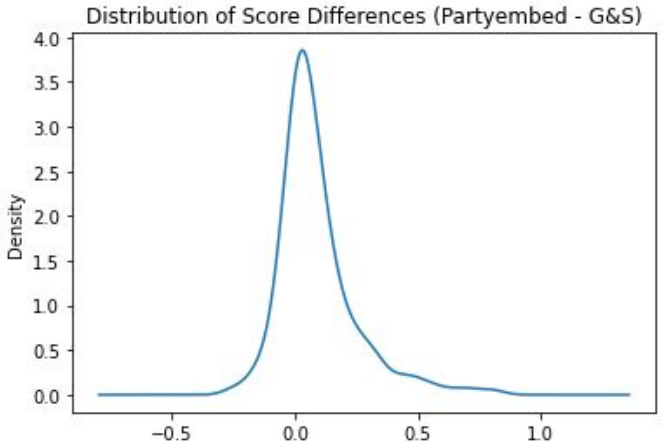


Fig 7. Distribution of score differences (Gentzkow and Shapiro score subtracted from Partyembed score)

Generally, there seemed to be a not-insignificant amount of disagreement between the two models, which made itself particularly clearer when plotting the scores side-by-side.

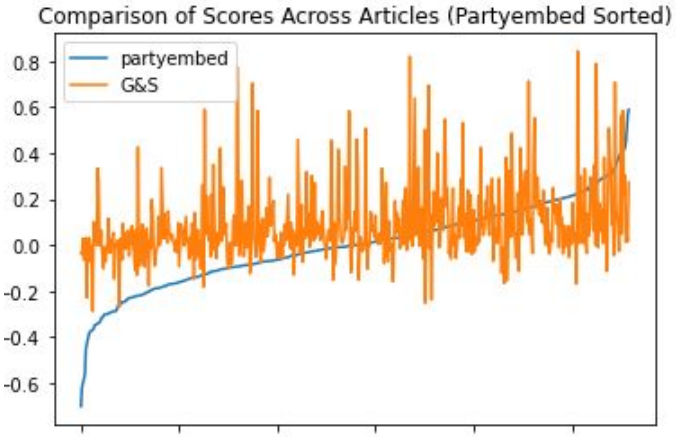


Fig 8. Comparison of scores across articles (with partyembed sorted)

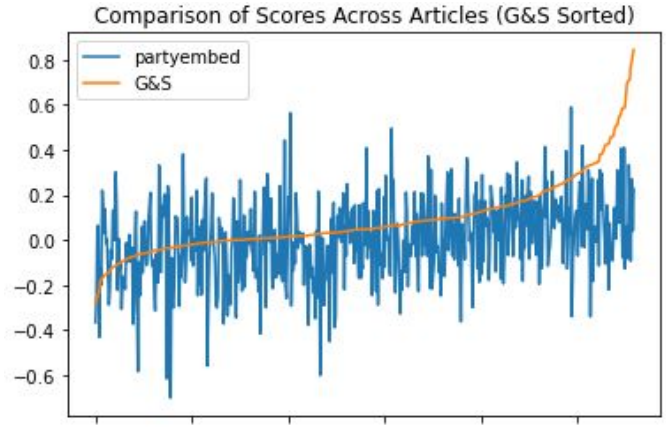


Fig 9. Comparison of scores across articles (with G&S sorted)

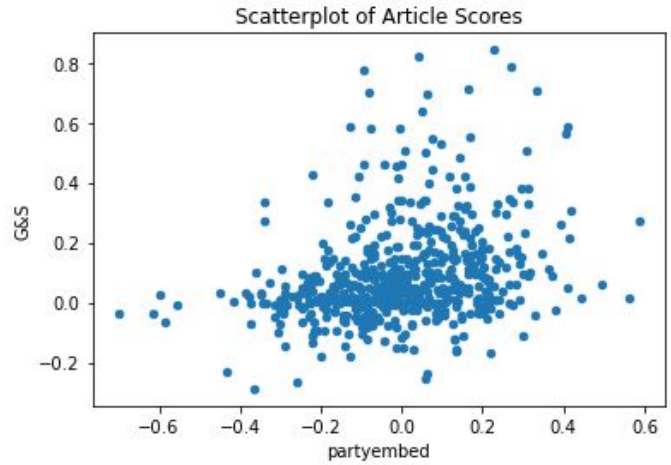


Fig 10. Comparison of scores, scatterplot

Fortunately, the differences in scores were not entirely random, and there appeared to be patterns within the types of articles with large or small differences. The largest differences has a large number of individuals mentioned, particularly representatives and politicians.

	article	absdiff
0	Lauren Underwood	0.819861
1	Brian Kemp	0.815293
2	Claire McCaskill	0.784592
3	Tammy Baldwin	0.761509
4	Dan Crenshaw	0.712092
5	Justice Democrats	0.695768
6	Maeve Kennedy McKean	0.687376
7	Political positions of the Democratic Party	0.673666
8	Freedom Caucus	0.632639
9	Georgia House of Representatives	0.601890
10	Emily W. Murphy	0.591379
11	Steve Bullock (American politician)	0.583115
12	Democratic Party (United States)	0.543645
13	Doug Jones (politician)	0.539928
14	Blue Lives Matter	0.535505
15	Ritchie Torres	0.521769
16	Texas House of Representatives	0.521481
17	Cori Bush	0.506001
18	Political appointments by Donald Trump	0.495688
19	Charlie Baker	0.477034

Fig 11. Articles with largest difference between models

Meanwhile, the smallest differences had a noticeable number of articles concerning presidential elections featured.

	article	absdiff
0	Operation Inherent Resolve	0.000992
1	1852 United States presidential election	0.001012
2	Gerald Ford	0.001054
3	1860 United States presidential election	0.001194
4	1924 United States presidential election	0.001359
5	P. T. Barnum	0.001616
6	James Buchanan	0.002017
7	Jeffrey A. Rosen	0.003102
8	Anti-Americanism	0.003490
9	Maria Butina	0.003589
10	Jo Jorgensen	0.003761
11	1848 United States presidential election	0.003987
12	John Conyers	0.004349
13	Federal Reserve	0.004512
14	Jim Acosta	0.004769
15	Clinton–Lewinsky scandal	0.004834
16	Thomas Jefferson	0.004939
17	1812 United States presidential election	0.005000
18	1944 United States presidential election	0.005116
19	Democratic-Republican Party	0.005445

Fig 12. Articles with smallest difference between models

However, these patterns aside, there were many articles whose large differences can be areas for concern, such as “Political positions of the Democratic Party” and “Blue Lives Matter.” Moreover, figures and politicians are of course not excluded from the articles with the lowest differences.

Regardless, an area we were particularly interested in was the most politically slanted articles. Here, the results were quite different by each model.

For the most left-leaning articles identified by the partyembed model, it made us quite concerned over whether or not our aforementioned strategy was effective in avoiding our models acting as a topic detector.

These articles were heavily centered around left-leaning topics, such as the article for “Civil rights movement” and “New Deal coalition”. While this could possibly mean that these articles were in fact written with a left-leaning slant, it could also mean our article could not avoid being a topic detector when it is applied to these particular kinds of Wikipedia articles.

	article	score
0	Great Society	0.058814
1	Civil rights movement	0.056364
2	117th United States Congress	0.049442
3	Government shutdowns in the United States	0.044246
4	Barbara Lee	0.041793
5	New Deal coalition	0.041295
6	Democratic Party (United States)	0.040954
7	1836 United States presidential election	0.040813
8	Ritchie Torres	0.040658
9	History of the Democratic Party (United States)	0.039406
10	Franklin D. Roosevelt	0.038063
11	Presidency of Lyndon B. Johnson	0.037090
12	1968 United States presidential election	0.036271
13	Southern strategy	0.034569
14	Political positions of the Democratic Party	0.033311
15	116th United States Congress	0.033094
16	Equal Rights Amendment	0.031499
17	1972 United States presidential election	0.031168
18	History of the Republican Party (United States)	0.031048

Fig 13. Most left-leaning articles identified by partyembed model

As for the right-leaning articles, the partyembed model identified interesting choices. For instance, although our method for applying the partyembed model to Wikipedia articles used data from 2007 onwards, it consistently identified articles relating to Trump and Russia, data only relevant for the very end of the data selection. What’s more, it is quite likely that democrats too were talking about the Mueller investigation, etc. So these results were particularly perplexing.

Another interesting result within the right-leaning articles (though not the most right-leaning articles) was the prevalence of conspiracy-related articles. All but one of the articles relating to conspiracies that we found (the exception being the article pertaining to conspiracy theories relating to the assassination of Robert F. Kennedy) each had a right-leaning slant. What’s more is that the majority of these right-leaning slants were not mild, but many were quite noticeably towards the right-wing end. Furthermore, these articles were not all about right-wing conspiracy theories, but included relatively non-partisan theories such as the New World Order.

While we are hesitant to make drastic claims about the world from these results, what we could possibly conclude is that the language being used in these conspiracy Wikipedia articles is similar to that being used by republican congresspeople.

	article	score
0	Steele dossier	-0.070050
1	Taxation in the United States	-0.061557
2	Special Counsel investigation (2017–2019)	-0.060005
3	Crossfire Hurricane (FBI investigation)	-0.058383
4	Contiguous United States	-0.055772
5	Bible Belt	-0.045021
6	Mueller report	-0.043092
7	Flags of the Confederate States of America	-0.041506
8	Barack Obama religion conspiracy theories	-0.038836
9	Georgia General Assembly	-0.037774
10	Insurgency in Khyber Pakhtunkhwa	-0.037309
11	Gettysburg Address	-0.037190
12	Separation of church and state in the United S...	-0.036648
13	United States Department of Defense	-0.036208
14	William Henry Harrison	-0.034829
15	Jefferson Davis	-0.034639
16	Black Lives Matter Plaza	-0.034417
17	Russian interference in the 2016 United States...	-0.033993
18	Bowling Green massacre	-0.033935

Fig 14. Most right-leaning articles identified by partyembed model

	article	score
8	Barack Obama religion conspiracy theories	-0.038836
34	Burr conspiracy	-0.029171
57	Epstein didn't kill himself	-0.022827
73	New World Order (conspiracy theory)	-0.021400
95	QAnon	-0.017914
105	Pizzagate conspiracy theory	-0.016815
115	Biden–Ukraine conspiracy theory	-0.015848
199	Cultural Marxism conspiracy theory	-0.008204
430	Robert F. Kennedy assassination conspiracy the...	0.007654

Fig 15. Scores of conspiracy-related articles, partyembed

For the left-leaning articles identified by the Shapiro & Gentzkow model, the current pages revealed many articles centered around individuals, often senators and representatives. As there seems to be less (though still present) representation of left-wing topics within these topmost articles, one could perhaps take away that the S&G approach is more able to avoid topic detection within the most left-wing articles. However, more research would be needed to make sure of this.

	index	scbn
0	Joe Neguse	1.247925
1	Catherine Cortez Masto	0.974850
2	Fair Fight Action	0.970201
3	Lauren Underwood	0.842551
4	Brian Kemp	0.819379
5	Tammy Baldwin	0.788409
6	Claire McCaskill	0.775247
7	Maggie Hassan	0.747780
8	Ben Sasse	0.733562
9	Justice Democrats	0.712180
10	Political positions of the Democratic Party	0.706976
11	Dan Crenshaw	0.703825
12	Chris Murphy	0.703029
13	Patriot movement	0.701884
14	Maeve Kennedy McKean	0.693417
15	John E. James	0.680519
16	Freedom Caucus	0.637483
17	Lou Correa	0.605684
18	Chris Van Hollen	0.599967

Fig 16. Topmost left-wing articles, G&S

On the other hand, the most right-wing articles were more of a mix. Like partyembed, the topmost right-wing articles did include a number of articles pertaining to Trump and Russia, such as the articles for Mueller report and Michael Flynn. However, there was also a lower representation of articles pertaining to right-wing topics. While the former included articles such as “Bible belt,” “Flags of the Confederate States of America,” “Taxation in the United States,”, “Jefferson Davis,”, and “Barack Obama Religion Conspiracy Theories”, the latter included far fewer articles such as “Newsmax,” “Donald Trump 2016 presidential campaign.” With this common trend in both left and right wing, one could possibly take away the message that the Gentzkow & Shapiro approach performs slightly better at avoiding topic detection overall.

As to why this is, it is still difficult to say. Both had their original sources in generally the same data-congressional corpora. And both of these methods tended to have the same approach in differentiating the ideology behind the two different documents. One of the main differences, as small as it is, is that the Gentzkow & Shapiro approach uses bigrams, whereas we used unigrams for the issue() function in our application of the partyembed model.

	index	scbn
0	Separation of church and state in the United S...	-0.287615
1	Iran–Contra affair	-0.266390
2	American Enterprise Institute	-0.252642
3	Michael Dukakis	-0.237237
4	Mueller report	-0.228100
5	Newsmax	-0.180973
6	United States Militia	-0.176879
7	1932 United States presidential election	-0.169072
8	Adlai Stevenson II	-0.165130
9	Rationale for the Iraq War	-0.159015
10	Black nationalism	-0.158779
11	Anti-Defamation League	-0.154739
12	Watergate scandal	-0.150001
13	Donald Trump 2016 presidential campaign	-0.149953
14	Jeffersonian democracy	-0.147519
15	Michael Flynn	-0.146907
16	Central Intelligence Agency	-0.131289
17	Roe v. Wade	-0.129793
18	Abscam	-0.123639

Fig 16. Topmost right-wing articles, G&S

Time-series analysis

An area in which there was a noticeable amount of disagreement was that of the time series analysis. When analyzing the plots side-by-side, some articles featured drastically different interpretations of the change of our selected articles. While many articles had modest positive correlations with each other and generally looked quite similar, other articles such as "Separation of Church and State" seemed to have opposite interpretations of the lifespan of an article, with overall ratings of bias moving in entirely opposite directions, as seen in Figure 13 below.

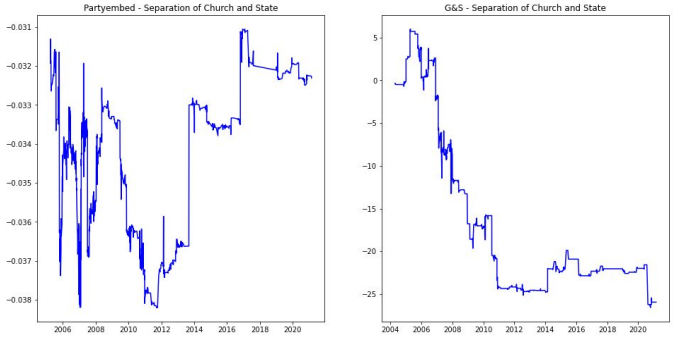


Fig 17. Article with high disagreement - Separation of Church and State time series slant plot (Partyembed on left, G&S on right)

However, one noticeable trend was that for the majority of articles, (Jim Acosta and Democratic Party being the only two exceptions) the majority of the variation in an article's bias rating was typically found at the very beginning of its lifespan. One of the more drastic examples of these (Mueller Report) can be found in Figure 14

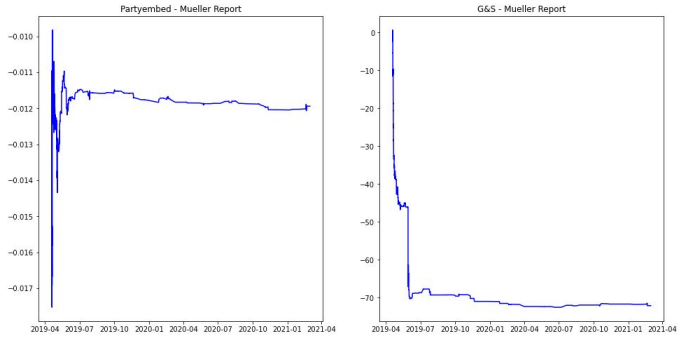


Fig 18. Mueller Report time-series slant plot (Partyembed on left, G&S on right)

This makes sense intuitively, if article lengths generally increase over their lifespan, slanted edits of around the same size will typically have less and less of an impact as time goes on.

Another finding we found was the "stair step" movements of the various plots. We assume that these are the results of large chunks of the article being added or taken away.

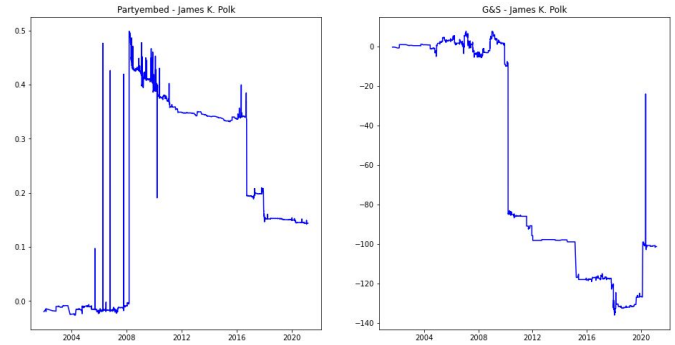


Fig 19. Large stair step motions on the article for James K. Polk (partyembed on left, G&S on right)

Another object of interest to us was the varying ranges of scores over time— if these ranges were similar, or different in some pattern, etc. In general, we found that the range of scores given to an article over its span varied wildly.

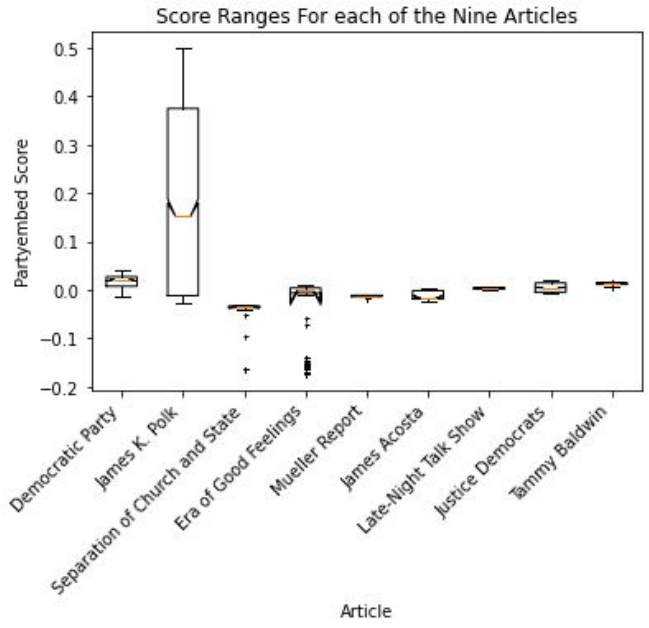


Fig 20. Score Ranges for the nine time-series articles (partyembed)

When looking at the ranges for the scores of each of the articles in question, the articles with the largest variation in scores were in fact so large as to obscure the patterns in any of the smaller ranges (Figure 16).

When analyzing the score ranges that seemed closer together (that is, after dropping the three articles with the highest variation, "James K. Polk," "Separation of Church and State," and "The Era of Good Feelings"), we found that even then, the ranges in scores varied substantially.

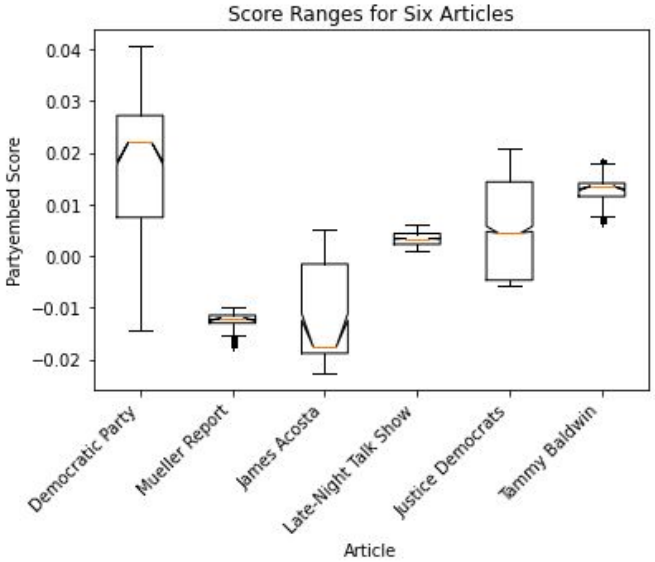


Fig 21. Score ranges for six articles (partyembed)

Overall, the variation does not seem to be related to the overall slant or the size of the article. The largest variations were with a long-neutral article, medium-republican article, and short-neutral article. The smallest variations were in a long-republican article, medium-neutral article, and medium-democratic article.

CONCLUSION

The two political parties in the U.S. utilize diverging vocabularies. Democrats are likely to use left-leaning term "undocumented immigrants" while Republicans are likely to use the right-leaning term "illegal aliens". Ideological divisions are increasing and pervasive. There is a distinct political polarization of language used in congressional speeches as preliminarily explored by Rheault and Cochrane. Since congressional speeches feed media and public discourse, this growing partisanship of language can then find its way onto open-source resources such as Wikipedia. Extreme political opinions can have reverberating consequences and Wikipedia has expressed its desire to remain factually neutral. Our work expands upon the papers by Rheault and Cochrane and Gentzkow, Shapiro, and Taddy by applying their methods to Wikipedia, one of the most visited websites in the world for free public information.

In this paper, we develop and compare two models to detect and measure ideological slant on Wikipedia across article topics and throughout time. Both models are trained on congressional speech data, validated using the Ideological Books Corpus, and applied to the 700 most-viewed articles related to U.S. politics. Model validity increases with respect to the number of sentences and both models produced similar results. Over 62 percent of article scores had an absolute difference less than 0.1. The only area where the models differed was on articles regarding political figures, such as congresspeople. Generally, the models agree on abstract political articles. A particularly surprising result is that both the models consistently identified articles related to Trump and Russia as right-leaning.

Our main issue of concern is that the Partyembed model, originally designed by Rheault and Cochrane, acts more like a topic detector. Topic bias is more prevalent in the Partyembed model than in the Gentzkow & Shapiro model. For example, the most left-leaning articles identified by Partyembed are heavily centered around left-leaning topics. On the other hand, the most left-leaning and right-leaning articles identified by the Gentzkow & Shapiro model feature a variety of topics. In order to determine if the Gentzkow & Shapiro model performs better than Partyembed we would need to do more research.

Measuring ideological slant is a core topic in political science and a daunting task for data scientists. Not only can Wikipedia use these two models to determine the degree of political bias in their articles, but political scientists can use the quantitative finding of this paper when examining shifts in public opinion. While we were able to apply two predeveloped models to Wikipedia data, they can be improved upon. In the future, we would like to implement a bias detector that does not get swayed by the topic at hand by possibly altering the training/validation data.

REFERENCES

[1] Sumi, Robert, et al. **Edit Wars in Wikipedia**. Budapest University of Technology and Economics, 9 Feb. 2012, arxiv.org/pdf/1107.3689.pdf.

[2] Greenstein, Shane, and Feng Zhu. 2012. **Is Wikipedia Biased?** American Economic Review, 102 (3): 343-48. DOI: 10.1257/aer.102.3.343

[3] Rheault, Ludovic & Cochrane, Christopher. (2019). **Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora**. Political Analysis. 28. 1-22. 10.1017/pan.2019.26.

[4] Gentzkow, Matthew, and Jesse M Shapiro. **What Drives Media Slant? Evidence from U.S. Daily Newspapers**. Econometrica, Jan. 2010, web.stanford.edu/~gentzkow/research/biasmeas.pdf.

[5] Gentzkow, Matthew, et al. **Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech**. July 2016, siepr.stanford.edu/sites/default/files/publications/16-028.pdf.

APPENDIX

Project Proposal

As established by Wikipedia itself, edit-warring is remarkably counterproductive and only makes consensus harder to reach. In Edit Wars in Wikipedia, Robert Sumi et al devised an M-Statistic which can grant any Wikipedia article a value representing its level of controversy; while it can quickly and effectively identify highly controversial articles, it is generalized to take into account any type of edit war (among other limitations), with an accuracy that is far from perfect. In general, this project seeks to address two key deficiencies in this method of conflict detection: scope of controversy and limitation in methods. While the aforementioned method was generalized for any and all edit wars across all topics, this project will focus on political controversy; while the aforementioned method solely focussed on edit wars, this will combine that with sentiment analysis.

The rationale behind focussing on political controversy is twofold. Firstly, unproductive political controversy and the resulting potential lack of accurate information is known to have severe consequences, and these consequences are particularly salient in these current times. As seen in Greenstein and Zhu's paper in 2018, bias in Wikipedia is indeed present, and it is both in Wikipedia's interest and in the interest of the general public for it to be as close as possible to a state of political neutrality and factuality. As a result, lowering controversy in this area becomes particularly salient. Political bias could be a particular method of targeting this— politically charged language is for one unhelpful, but additionally can provoke the other side and lead to additional controversy. Finding a way to neutralize politically charged language could then be helpful in efforts to quell political controversy and focus on neutral, factual information. As for the second rationale, politically biased language has excellent tools available and has already been a topic of study. Identifying ideological language is something that has already been done before; for example Rheault and Cochrane in Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora successfully uncovered ideology within digitized parliamentary debates. As for what tool we will use, we will have to do further research as to which tool will be most effective (see Schedule, Week 1), however for now we are planning to train a model on the ideological books corpus (Sim et al, 2013) and attempt to generalize this to Wikipedia articles, validating it on edit comments which explicitly mention reverting bias.

In order to confirm the relationship between politically charged language and controversy, we could run the chosen model on full article text and talk pages. In order to get the data for these pages, full data of all of Wikipedia is regularly uploaded to Wikimedia downloads. From here we can download full revision history in order to perform analysis of controversy if we decide to use a similar reversion analysis as we did with the M-Statistic. This data contains the full text of each revision of each article, as well as information concerning the time of the edit and the user behind this edit. From this, we can hash the text and take note of the time and the user. As for the talk pages, these are available in the "meta-current" rendition. We can match the titles with those of the full history in order to join these sets together. From here, we can perform sentiment analysis on the current article as well as the talk page. As for what this data looks like, it is in a similar format to the standard article XML data, but topics and comments are all denoted with textual symbols (topics starting and ending with "==" and comments with ":"), and the title of these pages connect to the current page in the format "Talk: Original_Article_Title". Joining all of this data, then, should be quite simple, as is downloading it all. The majority of the work, therefore, will be in manipulating and transforming this data.

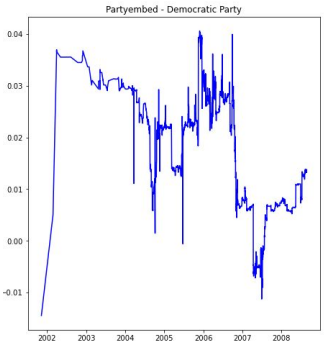
We are also interested in exploring the relation of clickstream data with political bias. If we can find an association with traffic to and from a particular article and the political bias of this article, this can lead to more efficient detection of politically biased articles. This clickstream data is freely available as well. It consists of (referrer, referee) pairs, in addition to the number of times this pair appears in the data. While we cannot make larger chains beyond this with absolute certainty, we can still gain an idea of from where people are arriving to these articles, and where they are going. From these clicks, only about 62% are internal, but it should still provide us with plenty of information about the nature of the users. As for the number of times these pairs appear, there is a median of 24 clicks, a mean of 92.6, skew of 126, 90% quantile of 143 and a max of 220000 clicks; the data is very much right-skewed and with very serious outliers.

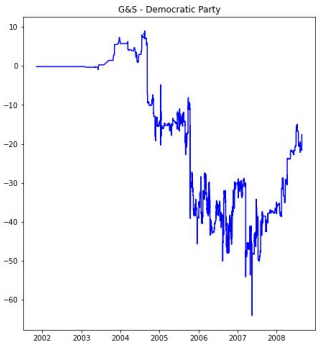
Overall, after gathering all of our insights from this research, we intend to create a bot that could effectively find politically charged Wikipedia articles and notify editors of the issue. We intend to create it such that it is able to point out specific lines that are particularly biased, or perhaps even suggest possible corrections that are less politically charged. This bot could then be set to run periodically in order to identify these problematic phrases. And regardless of the

creation of this bot, all insights will be compiled into a paper in order to communicate our findings.

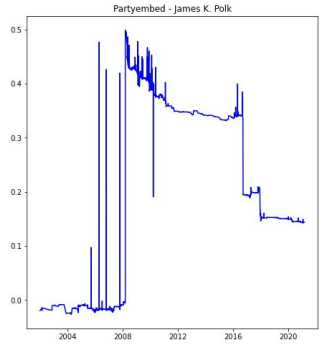
<u>All Time series graphs:</u>

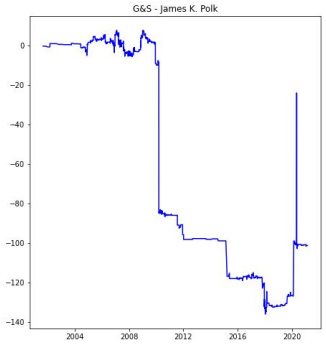
Democratic Party:



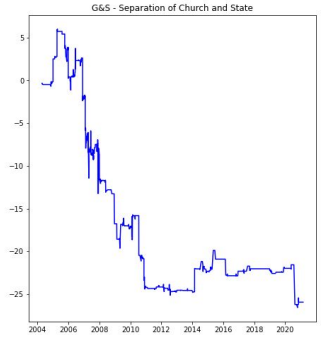


James K. Polk:

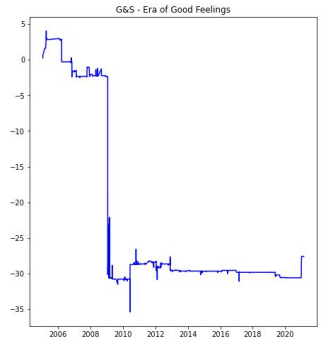




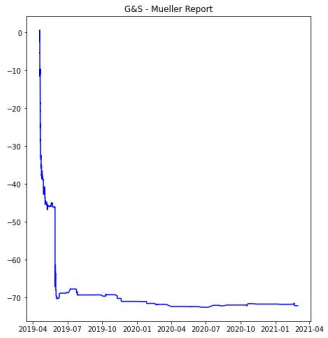
Separation of Church and States:



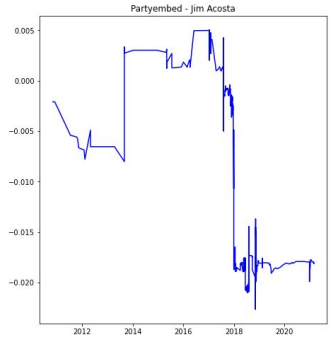
The Era of Good Feelings:

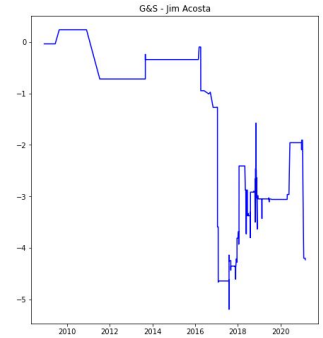


Mueller Report:

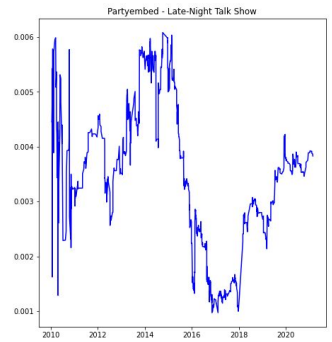


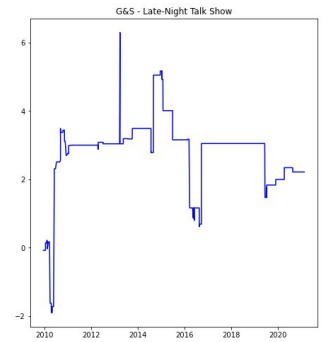
Jim Acosta:



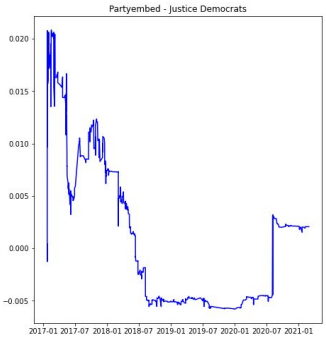


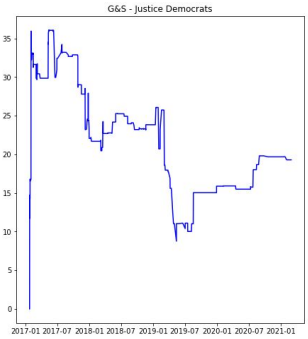
Late-Night Talk Show:

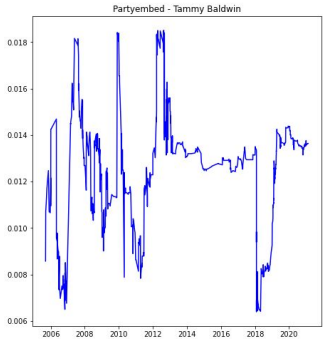


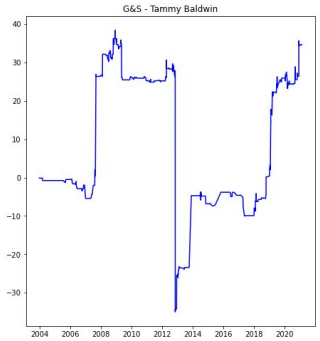


Justice Democrats:









Tammy Baldwin: