E1282 Midterm Project Report: To what extent does Wikipedia accurately portray the political positions of US Congress members?

Anna Hundehege
a.hundehege@mpp.hertie-school.org

Sofia Diogo Mateus
s.diogo-mateus@mpp.hertie-school.org

Maria Camila Garcia
m.garcia@mpp.hertie-school.org

Abhas Tripathi
a.tripathi@exchange.hertie-school.org

Abstract

Wikipedia is often the primary source of knowledge of internet users on any given topic and specifically on politicians. But do the profiles on the platform represent the lawmakers accurately? We try to measure how accurately a politician's Wikipedia profile showcases their political stances and policy positions. We attempt to quantify and analyse how much information the profiles provide and rate them on a left-right scale, so as to accurately place US lawmakers in relation to another. Our first efforts show that Wikipedia does not allow us to accurately represent lawmakers based on the information provided in their profiles. In fact, it shows a picture that lacks nuance and that is quite informed by the visibility of the congress member, rather than their voting records.

1. Proposed Method

Wikipedia has become an important source of information on politics and is considered a relevant platform for political communication by politicians. Göbel and Munzert [2] found that editing behavior of Wikipedia profiles from German legislators reflects (re)election motifs. It is thus highly relevant how politicians and their policy positions are portrayed on Wikipedia as the profiles might inform voting preferences. At the same time the platform is often perceived as neutral[2], which would lead most to expect no politico-ideological bias in the Wikipedia profiles. To evaluate the level of accuracy at which policy positions are presented on Wikipedia, we drew a test set of 32 US politicians' profiles to explore the data and to develop our model. Every profile is a document in our corpus, which we transformed into a document-feature matrix removing English stopwords, punctuation and digits and stemming all the remaining tokens.

Although the profiles differ in total length of the profiles and number of tokens, the documents in the corpus are overall similar in terms of lexical diversity and readability. This corresponds to the crowdsourced editing of Wikipedia profiles and a generally similar structure and writing style across Wikipedia entries. Overall, all the profiles seem to be rather similar or different to the same extent with cosine similarity for example varying roughly between 0.3 and 0.5, but with no clear pattern of party membership or gender. Exploring the top features in the corpus and the keyness of individual profiles confirms the expectation of an overall neutral tone. While tokens relating to the person and the political system in general appear at high frequencies, tokens hinting at the policy positions of individual politicians are much less frequently used (see also Figure 1).

In order to measure the accuracy at which policy positions are presented on Wikipedia, our first approach was to replicate the Manifesto Project [6]. We used Wikipedia entries and coded the profile sentences to individual politicians rather than political parties, assuming that the basic ideological dimensions would be the same and overall reasonably transferable. However, we realized that coding the corpus at a sentence level would require substantial machine-learning expertise to train an algorithm to scale the manual coding to a larger amount of virgin texts. So we decided to analyse the corpus on a word level rather than at sentence level. This also leaves us with less methodological problems of transferability of the coding scheme from parties to individual politicians.

To analyse the ideological positioning of US members of Congress in their Wikipedia profiles, we then decided to use the wordscores methodology following Laver, Benoit and Garry [3] and adapting the code provided by https://uclspp.github.io/PUBLG088/wordscore.html to our corpus. First, we selected a set of US politicians that would be placed on the extremes of the political spectrum as our reference texts and assigned them scores based on our judgement of the US party system and informed by scores for party manifestos. Based on these reference texts we used the textmodel function of the quanteda package to predict

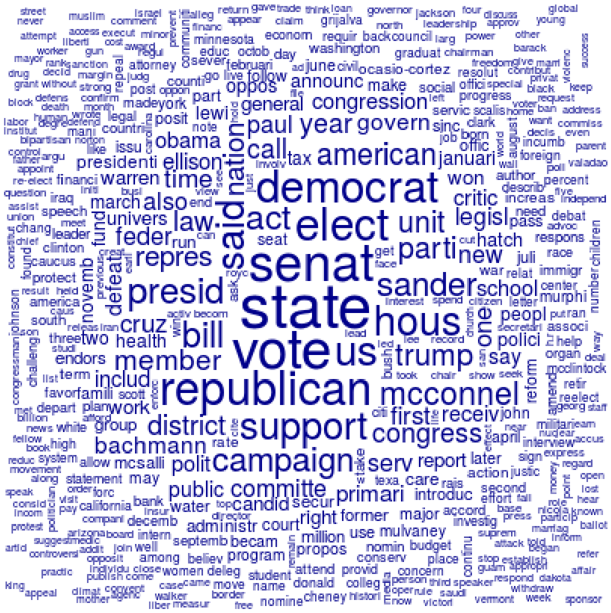


Figure 1. Top features in 32 selected Wikipedia profiles.

the scores of the remaining Wikipedia profiles, the so-called virgin texts. This allows us to classify the politicians as either Republican or Democrat based on the content of their respective Wikipedia profiles. Comparing our scores to the party membership and information on their ideological positioning provided by the VoteView project [1] allows us to quantify how accurately Wikipedia profiles portrait the policy positions of US politicians. For this midterm report, we first applied the methodology to a test set of Wikipedia profiles to evaluate the accuracy of our approach.

For a brief summary of the wordscores underlying methodology, we follow the approach of Lowe, [5] for the estimation of wordscores we have the following formula on computing the wordscore of a given document based on the average of scores obtained from it.

$$\hat{\theta}_d = \frac{1}{W} \sum_{w}^{W} \hat{\pi}_w$$

where W represents the number of words present in the document, π_w is the average of the wordscores, and θ stands for the estimated score of document "d". Based on this equation, as Lowe explains, each word token is considered as giving the same amount of information about the general score of the document.

Lowe also includes other specification that focuses on the effect of word types and the weigthed probability for the average of the wordscores:

$$\hat{\theta}_d = \sum_{j}^{V} \hat{\pi}_j \ \hat{P}(w_j|d)$$

where V accounts for the types of words that appear in the document. Finally, under the assumption that "frequency is assumed to be a direct reflection of a word's importance in determining a document's score" [5], the final true estimator of the document score is:

$$\theta_d = \sum_{j}^{V} \pi_j \ P(w_j|d)$$

Lowe [5] also mentions, that one weakness of the wordscore methodology is that in a context without an underlying statistical model we cannot foresee the assumptions made by the wordscores coding therefore complicating the analysis of the results.

2. Experiments

Data: Wikipedia page entries for all Congress members from 2013 to present. For our initial test set we used 12 politicians, 6 from each party. We attempted to use women and men but also fairly established and recognised names, assuming that would give us more text to work with.

Evaluation method 1: We experimented applying the coding structure from the "Manifesto Project" [6] on a sample of 12 congressmen/women and with it create a case for the extremes of Republicans and Democrats. However, we faced different challenges when trying to extrapolate this methodology to our database; (i) Wikipedia entries have extensive sections on personal life and voting behaviour of congressmen/women that could not be accounted for when using the coding structure of the ”Manifesto Project” [6] - this implied having quite a significant number of sentences uncoded with the risk of biasing our results; (ii) some of the political issues specific to the United States (i.e. gun control, climate change debate) did not properly fit into the coding system. One specific case was for those who oppose climate change as the code provided an option for ”Environmental friendly: Positive” but not a negative alternative. This implied we would have to modify the coding structure by adding new categories, again running into the risk of corrupting our results; (iii) finally, Wikipedia entries are not comparable to party manifestos - the latter have a specific format with clear political statements and repetition of main critical points. Therefore, assuming that Wikipedia entries can be analyzed under the same structure can provide misleading outcomes. Wikipedia entries for Congressmen/women provide information about voting behaviour and political career rather than focusing on clear political statements.

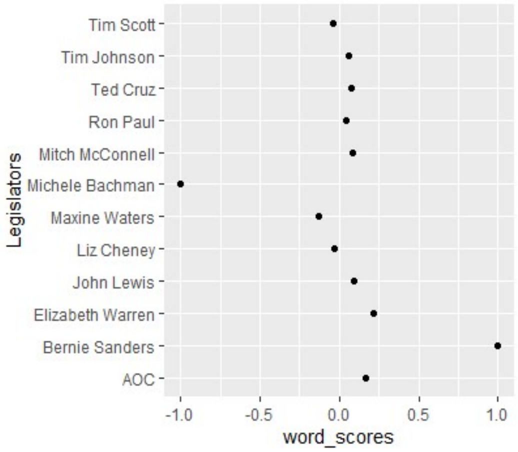
Evaluation method 2: We used wordscores, picking the most extreme member and placing them at each end of the scale. We then allow the algorithm to place all the other profiles in relation to the two extremes. The results are compared to VoteView [4], UCLA's tracker of US lawmakers' voting records and overall positioning on the left-right scale.

Experimental details: We attempted to use women and men but also fairly established and recognised names, assuming that would give us more text to work with. As such, we did not pick straight from the VoteView score of the most liberal or conservative but tried to restrict ourselves to legislators who feature highly on that score but also whose profiles are lengthy. Table 1 shows the combination of extremes we used to build the wordscores.

Models	Democratic	Republican
(1)	Bernie Sanders	Michele Bachmann
(2)	Bernie Sanders	Orrin Hatch
(3)	Bernie Sanders	Ted Cruz
(4)	Alexandria Ocasio-Cortez	Michele Bachmann
(5)	Alexandria Ocasio-Cortez	Orrin Hatch
(6)	Alexandria Ocasio-Cortez	Ted Cruz
(7)	Jerry Nadler	Michele Bachmann
(8)	Jerry Nadler	Orrin Hatch
(9)	Jerry Nadler	Ted Cruz

Table 1. Wordscore models created

Results: The wordscores result for the reference texts is very inaccurate: In model (1) - Bernie Sanders as most left (+1) and Michele Bachmann is the most right (-1) - names like Ron Paul and Ted Cruz have positive scores, indicating they are left to the center. In fact, in the VoteView, Ted Cruz is consistently ranked in the top 4 most conservative senators, 97 percent more than all other Senators in the 2013-2019 timeframe. The same holds true for Maxine Waters, consistently rated by VoteView as 99 percent more liberal than her colleagues - and yet, under the wordscores model, a conservative. In fact, all 12 figures are fairly far out on the political scale - and yet, with wordscores they congregate around the center, as can be seen in Figure 2.



ss

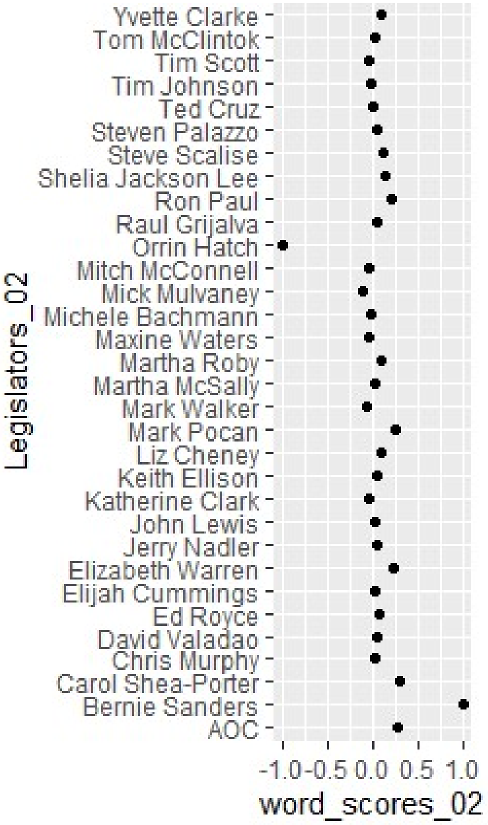
Figure 2. Sanders - Bachmann

To overcome this issue, we expanded the sample to 32, adding 20 more names - 10 Democrats and 10 Republicans. Again, we pick them based on profile length but also visibility and positioning. Again, the results congregate around the center, with scores that do not represent the positioning on a left-right scale even remotely well. We ran 9 models, trying to find matches for careers in terms of time - Bernie Sanders and Orrin Hatch, Alexandria Ocasio-Cortez and Michele Bachmann, Elizabeth Warren and Ted Cruz. The results continue to congregate around the center, as seen in Figure 3, for model (2) pitting Sanders against Orrin Hatch, an equally longstanding member of Congress whose profile is also very much rich in information, unlike most others.

In this model, Elizabeth Warren, consistently ranked the most liberal member of the Senate, is ranked less liberally than Mark Pocan, whose VoteView ranking has him, at best, in the top 10 most liberal. Pocan also is considered significantly more liberal than Maxine Waters, a longstanding fixture of the most radical Democrats, ranked 99 percent more liberal than the rest of the House.

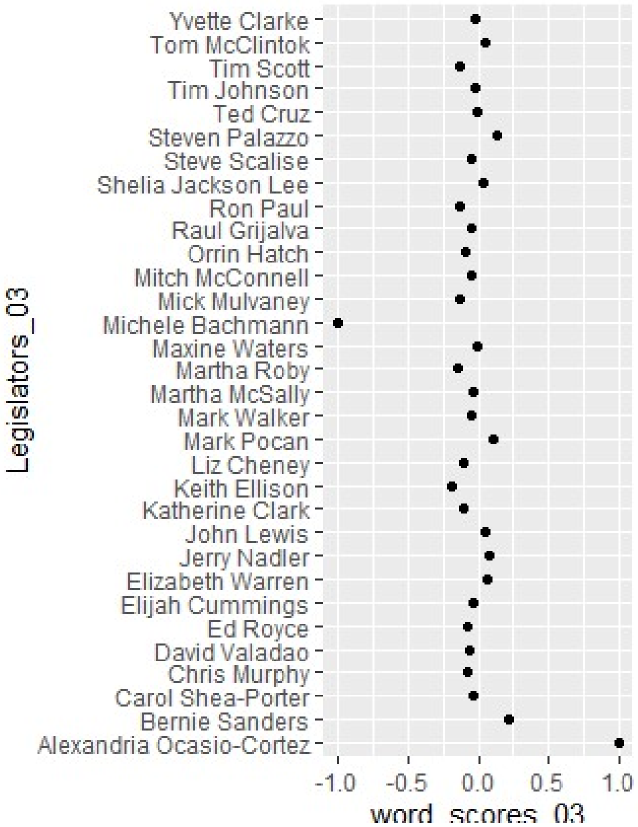
The disphasement continues in most other models, as Figure 4 showcases for our third specification.

Maxine Waters is ranked as conservative-leaning, with



ss

Figure 3. Sanders Hatch



ss

Figure 4. AOC - Bachmann

a negative ranking. This experiment places Alexandria Ocasio-Cortez against Michele Bachmann on the extremes but the algorithm places Yvette Clarke as more conservative than Ted Cruz, Sanders far away from Warren and Steven Palazzo, a longstanding Tea Party supporter, as twice as liberal as Warren (VoteView has Palazzo towards the 50 percent most conservative Republicans).

In many ways, the results are exactly what we expected: We posited that Wikipedia, given its nature, is not a good showcase of a politicians' policy position and of their political leanings. We think this is true, given the conditions of Wikipedia coverage but we are considering experimenting more with the model or possibly try different approaches, since wordscores was written explicitly for manifestos, a type of document that has a very different use of language than Wikipedia.

3. Future work

Following the methodological steps mentioned by Lowe [5], the next step would be to scale our model to the remaining virgin texts to try to predict the ideological positioning of the other ca. 800 members of Congress based on the wordscores. This would allow us to easier compare the results from the virgin texts with the reference text scores obtained from Wikipedia.

We expect to confirm that a direct comparison of Wikipedia profiles on a word level does not allow for an accurate classification of politicians. This could then be interpreted as a sign of rather neutral portrayal. However, our methodology has limitations because it does not account for differences of linguistic context and different sections of the text. An additional insightful analysis might be to develop a topic model at a sentence level to get a more nuanced picture of the policy positions presented in Wikipedia profiles. Furthermore, we would like to explore other factors that might explain differences in politicians' profiles such as gender, seniority in office, presidential runs or media attention.

References

[1] A. Boche, J. B. Lewis, A. Rudkin, and L. Sonnet. The new voteview.com: preserving and continuing keith poole's infrastructure for scholars, students and observers of congress. *Public Choice*, 176:17–32, 2018.

[2] S. Göbel and S. Munzert. Political advertising on the wikipedia marketplace of information. *Social Science Computer Review*, 36(2):157–175, 2018.

[3] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *The American Political Science Review*, 97(2):311–331, 2003.

[4] J. B. Lewis, K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet. Voteview: Congressional roll-call votes database. 2019.

[5] W. Lowe. Understanding wordscores. *Political Analysis*, 16:356–371, 2008.

[6] A. Werner, O. Lacewell, and A. Volkens. Manifesto coding instructions (5th revised edition), february 2015. *Wissenschaftszentrum Berlin*, 2014.