# E1282 Project Proposal: To what extent does Wikipedia accurately portray the political positions of the US American 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

## 1. Research Paper Summary

**Outline:** Since its launch in 2001, Wikipedia has become not only an important online source of information, but also an indirect tool of political communication for members of parliament. Although politicians do not have full control over the way they are portrayed on Wikipedia, they can influence the content of their profile. However and in contrast to other online communication tools such as Twitter and Facebook, Wikipedia has so far received only limited attention among political science scholars [1]. While Adam Brown [1] finds the information provided on technical issues to be mostly accurate, political issues have been less studied so far and seem to be biased in coverage. This points to a research gap in Political Science on both the engagement of politicians and citizens with Wikipedia and the Wikipedia community as a political actor.

Sascha Göbel and Simon Munzert [2] as well as Michael Herrmann & Holger Döring [4] thus introduce a new and relevant data source to the studies of politics using Wikipedia entries of political parties to measure their political ideology as a complement to traditional survey research. Focusing on the framing of policy positions, this project will explore how Wikipedia entries of US members of parliament (MPs) are biased with regards to political ideology (see also section 2). This is especially relevant because the platform is perceived as an overall neutral source of information by the users (see also Göbel & Munzert [2]).

### 1.1. Measuring political ideology using politicians' Wikipedia profiles

Michael Laver, Kenneth Benoit & John Garry (2003). Extracting Policy Positions from Political Texts Using Words as Data. The American Political Science Review, 97(2), pp. 311-331, http://www.jstor.org/stable/3118211.

**Summary:** Michael Laver, Kenneth Benoit & John Garry [5] use the frequencies of words occurring in a text to place the authors on a scale of political ideology. Instead of using a predefined dictionary with a limited set of words, they use a test set of documents with known positioning on the scale of political ideology to analyse the frequencies at which words occur in this specific set of documents. They use the relative observed frequencies of words to infer the probability of the positioning of other texts on this scale by attributing a political score to individual words and based on the occurrence of these word scores in the text under consideration.

Applying this model to a larger amount of party manifestos, the so-called virgin texts, they analyze how well the model predicts the ideological orientation of a text indicating a confidence interval. They validate their positioning of reference texts using external measures of political ideology. According to the authors, the appropriate choice of reference texts is key to the quality of the wordscores. The authors use British party manifestos of the Labour, Liberal Democrats and Conservative parties from 1992 to create the word scores, which are then used to predict the policy positions of the party manifestos from 1997.

The methodology adopted by the authors is of particular interest to us since the word scoring technique avoids many of the problems of traditional content analysis techniques. First, this technique provides quantitative measures of uncertainty thus allowing researchers to make informed judgements, specially when comparing relatively close or similar policy positions. Second, this technique does not require the meaning of words or context, and only needs a reference text. This makes the technique versatile, and text from any source and in any language can be analyzed. Lastly, because of a simple measure of scores of words, this technique is fast and computationally cheap.

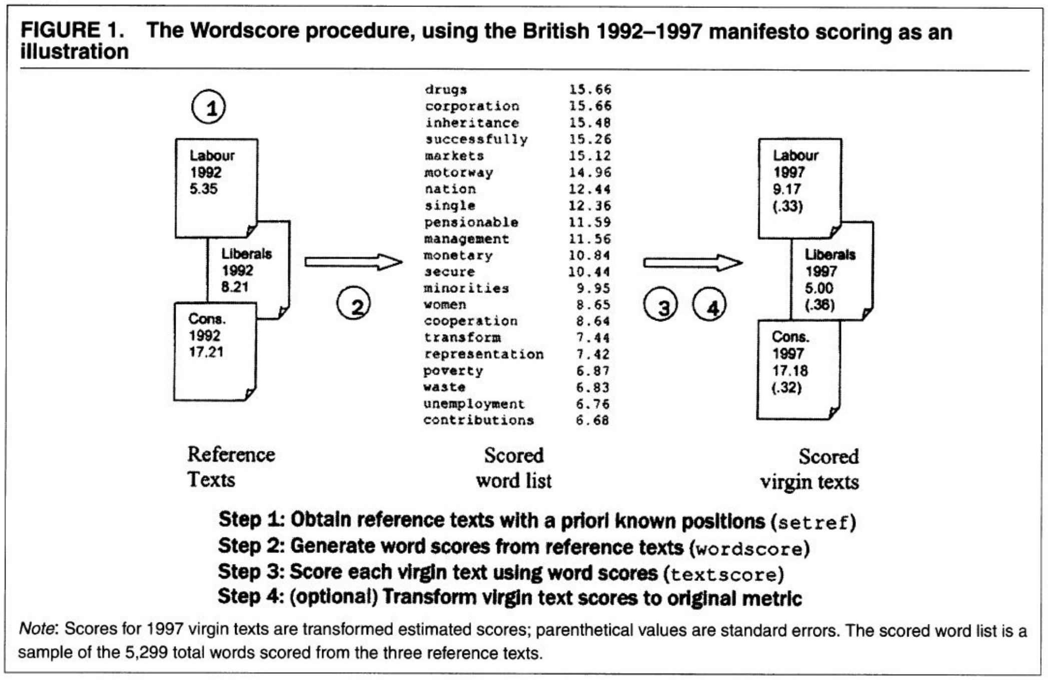
FIGURE 1. The Wordscore procedure, using the British 1992–1997 manifesto scoring as an illustration

Step 1: Obtain reference texts with a priori known positions (setref)
Step 2: Generate word scores from reference texts (wordscore)
Step 3: Score each virgin text using word scores (textscore)
Step 4: (optional) Transform virgin text scores to original metric

*Note*: Scores for 1997 virgin texts are transformed estimated scores; parenthetical values are standard errors. The scored word list is a sample of the 5,299 total words scored from the three reference texts.

Figure 1. The wordscore procedure, using the British 1992-1997 manifesto scoring as an example. Source: Laver, Benoit & Garry [5]

The authors show the flexibility and adaptability of their approach by testing the techniques for different datasets and measuring scores along various political dimensions. Starting with British party positions on economic and social policy, the authors do the same for Irish party positions on the two dimensions. They display the language versatility of the model by then applying it to German-language text. Successful application of the technique and favorable results confirm that the technique is language-blind and can migrate effectively into non-English language texts.

**Relevance:** The authors acknowledge that additional research on other data sources is required in order to validate the broader applicability of their method to measure policy positions. In combination with the approach developed by Herrmann & Döring [4] to measure political ideology of political parties, this project will apply the methodology of Laver, Benoit & Garry [5] to US members of Congress based on their Wikipedia personal profiles. The aim is to test the applicability of the measure of political ideology to individual politicians. This project is innovative in combining a new measure of political ideology and a new data source, the personal Wikipedia entries of politicians, for Political Science research.

Herrmann & Döring [4] use the labels provided in the infoboxes on Wikipedia entries of political parties to measure the ideological positioning of the parties. They use a probabilistic model to predict the ideology tags assigned to a party Wikipedia profile given the left-right party positioning using a broad dataset of about 4500 parties worldwide. External validation with expert surveys on party political ideology confirms that the Wikipedia tags are a qualitatively good and useful measure of ideological positioning of political parties across political systems. The authors further claim their measurement technique to be applicable beyond Wikipedia. Although a similarly thorough external validation might not be feasible within the scope of this project, these findings give reason to assume that the Wikipedia ideology tags for politicians are a similarly good approximation of politicians' political ideology. By extending this approach to US Congress members, this project demonstrates how the scaling effects of text as data methods can help explore new research objects in political science that have so far been neglected in research due to traditionally labor-intense human coding approaches. The positioning of individual politicians could inform further research on the composition and representation of political parties, electoral campaigning or political polarisation for instance.

## 2. Project Description

This project intends to take all the Wikipedia entries for US Congress members and attempt to create a typology of profiles that details multiple possible characteristics. Literature confirms, for example, that different perceptions based on the gender of American politicians apply regardless of party lines[3]. The literature also speaks about women's need to counter said narratives, and so, it stands to reason, that women would have more coverage when running for office. We will use a topic model to evaluate the text clusters and look at what denominators can be found across the different types of profiles and whether they indicate bias on the part of Wikipedia writers. We envision Wikipedia as a semi-mediated communication tool on the part of politicians, given that congressional staff make use of it but also that there is heavy input by Wikipedia editors and its community at large. Afterwards we will measure the political ideology on Wikipedia pages to place US Congress members on a left-right scale and compare it to the classification in the Wikipedia infoboxes.

### 2.1. Motivation

Wikipedia is often the people's first stop when researching an unknown topic or person. Studies show this is especially true of politicians as a search object. Given its importance, controversies over who edits the profiles and how objective the editors are abound, with some incidents casting a doubt on how much control there should be over certain pages in the website. As such our question is: To what extent does Wikipedia accurately portray the political positions of the US American Congress members?

With this project, we aim to see how neutral and accurate American politician's pages on the website are. Should our findings showcase significant bias, there may be room for Wikipedia to tackle this problem and re-think some of its features or contributor network features, namely diversity. We are also keen to lay out the different identities that shape a politician's information and to check them for discerning characteristics. We use Laver, Benoit & Garry's [5] work as a way to extrapolate from the description of policy positions (or lack thereof, in the case of Wikipedia) and test against the stated orientation in the Wikipedia infoboxes.

### 2.2. Task

We will analyse data obtained from Wikipedia pages of US legislators. We aim to understand if there are biases in terms of writing about these legislators based on political (party they belong to, leanings on the left/right and liberal/conservative spectra) and demographic traits (like gender, race, age, seniority in office). Ideological positions are also scaled using coded words and phrases from the content available on the respective legislator's Wikipedia page.

### 2.3. Data

For the purpose of this paper, we use data from Wikipedia pages of United States legislators in office since 2012. For all legislators who were in office in the selected period, content was collected using the WikipediR package, which makes Wikipedia page content available using page IDs from the legislatoR package. This data is then parsed to remove HTML tags and obtain page content with categories. Additionally, demographic and political data, which appears in infoboxes on Wikipedia pages, is obtained from the legislatoR package.

### 2.4. Method

We use a bag-of-words approach for descriptive analysis of the text. We perform dictionary analysis of words to understand what they represent for each legislator and party. For scaling positions on policy dimensions, like economic or social leaning, we use Wordscores, which is a commonly used scaling model for estimating political positions. This method generates scores for each word, for a particular *a priori* political dimension. We can thus measure how - and if - Wikipedia describes the multi-dimensional policy positions of US legislators in the selected period.

### 2.5. Baseline

The reference points for political dimensions are obtained from the labels provided in infoboxes for each legislator. Wikipedia provides labels like ”conservative” or ”liberal” for each politician and we aim to test scores for each entity against a numerical value for these labels.

## 3. Evaluation

The success of the project depends on how close the entity specific score predicted by the Wordscore model are to the reference scores. For quantitative evaluation, we will calculate the mean deviation of scores between the model and reference labels. A small or statistically insignificant deviation indicates a reliable model for predicting policy positions of politicians from their Wikipedia page entries. For qualitative evaluation, we will test our hypothesis against the Wikipedia content with respect to any bias found with respect to gender and race.

## 4. Contributions

We expect to divide all tasks across the team and have team members check each other's work. That said, Abhas Tripathi is focusing more on the coding of visual representations. Anna Hundehege is taking the lead on methodological research. Sofia Diogo Mateus will be offering a proposal for content to include in the dictionary. Maria Camila Garcia will be researching American political scaling.

## References

[1] A. Brown. Wikipedia as a data source for political scientists: Accuracy and completeness of coverage. *Political Science and Politics*, 44:339–343, 2011.

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

[3] D. Hayes. When gender and party collide: Stereotyping in candidate trait attribution. *Politics & Gender*, 7:133–165, 2011.

[4] M. Herrmann and H. Döring. Party positions from wikipedia classifications of party ideology. *Paper presented at the 2019 Annual MPSA Conference*, 2019.

[5] 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.

### Github Repository

abhast/TADA-Group-Assignment/