E1282 Final 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 US lawmakers accurately? We try to measure how accurately a US 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 Democratic-Republican scale, so as to accurately place US lawmakers in relation to one another. Our analysis shows that using quantitative analysis methods that are meant for political texts on Wikipedia articles leads to quite mixed results. Most methods do not allow us to identify the lawmakers' party affiliation with a high degree of accuracy. Given the information provided in their profiles, our analysis lacks nuance and is unable to discern any specific topics across the board. We believe this could be fixed with further tweaking to the model or possibly by designing models for non-political corpus analysis.

1. Introduction

Wikipedia is often the people's first stop when researching an unknown topic or person. Studies show this is especially true when people search for information on politicians online. Given its importance, controversies arise 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 was: To what extent does Wikipedia accurately portray the political positions of the US Congress members?

With this project, we intended to estimate how accurate US politician's pages on the website were in terms of their political stand and policy positions. Our findings did not allow us to confirm our expectations (i.e. finding major biases in their profiles) but they allow us to consider that there may be room for designing a method that better explores Wikipedia data based on its nature as a non-political text. In addition, the online encyclopedia may also consider setting up some politician-specific guidelines regarding the content mix in politicians' profiles. We were also keen to lay out the different identities that shape a politician's information and to check them for discerning characteristics, for which we used Laver, Benoit & Garry's [9] 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.

For this project we collected all the Wikipedia entries for US Congress members from 2013 to 2019 and applied diverse text-as-data methods to attempt to create a typology of profiles that details multiple possible characteristics or clusters of relevant political topics. We attempted to understand if there are biases in terms of profiling these legislators based on political factors (party they belong to, leanings on the left/right and liberal/conservative spectra) and demographic traits (like gender, race, age) and further scale their positions.

Our assumptions from classic political science literature confirm, for example, that there are different perceptions based on the gender of American politicians which apply regardless of party lines[7]. This literature also speaks about women's need to counter those narratives; based on this we assumed that online profiles on congresswomen would attempt to do the same.

One of the methods explored in this project was Wordscores through which we identified extreme cases to represent a 'true democrat' or 'true republican' to test Wikipedia profile's power to rank the other profiles in accordance to the extreme scale. We also used topic model to evaluate the text clusters and looked at which common denomina-

tors could be found across the different types of profiles. We envisioned Wikipedia as a semi-mediated communication tool on the part of politicians, given that congressional staff make use of it, but also that there would be heavy input by Wikipedia editors and its community at large.

2. Related Work

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 elected parliamentary bodies. 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 [3]. While Adam Brown [3] finds the information provided on technical issues to be mostly accurate while 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 [6] as well as Michael Herrmann and Holger Döring [8] 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 Congress are biased with regards to political ideology. This is especially relevant because the platform is perceived as an overall neutral source of information by the users (see also Göbel and Munzert [6]).

Herrmann and Döring [8] 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.

3. 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 [6] 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[6], 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 conduct a series of analysis based on different NLP methods: i) Wordscores, ii) methods from supervised learning, and iii) methods of unsupervised learning.

Our first exercise consisted in exploring the data and understanding how Wikipedia text is analysed by using NLP methods. To do so, we conducted a series of statistical summaries to approach the data by creating a test set of 32 popular and known US politicians' profiles. With these profiles we had the first exploratory analysis of the data and developed our first model. The first review of the sample statistics was useful to get acquainted with how NLP packages deal with the information provided by the Wikipedia profiles. For this initial part, we considered every profile as 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 and number of tokens, the selected data set allowed us to determine that 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 article 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, we decided to analyse the corpus on a word level rather than at sentence level. This approach left us with less methodological problems of transferability of the coding scheme from parties to individual politicians, as compared to other initial ideas. We used word frequencies or bag-of-words approaches - Wordscores, Naive Bayes, Regularised Regression, a Latent Dirichlet allocation and a Structural Topic Model, all of which we detail in the following section.

4. Experiments

Data: Wikipedia page entries for all US Congress members from 2013 to present. As previously stated, for our initial test set with wordscores we used 12 politicians, 6 from each party, which we later extended to 32 profiles. We attempted to use women and men but also fairly established and recognised names, assuming that would give us more text to work with. For the other evaluation methods we used the entire dataset of 793 members of Congress without any further selection exercise.

Evaluation method 1: We used Wordscores, building a scale based on the selection of two lawmakers, each representing the "extreme" profile for both Republicans and Democrats to then compare the positioning of other selected lawmakers with clearly defined allegiance to either party. We then allow the algorithm to place all the other profiles in relation to the two extremes to identify whether such placement matched the expected positioning of the other candidates as compared to the chosen extremes. The results were compared to VoteView [10], UCLA's tracker of US lawmakers' voting records and overall positioning on the left-right scale. 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.

Evaluation method 2: To validate our findings from the Wordscores approach, we then proceed with the application of other supervised methodologies. We apply a Naive Bayes and Regularised Regression as classifier models for the classification of US Congress members. For these models we use the entire dataset of 793 Wikipedia profiles with the correspondent definition of a training and test set.

Evaluation method 3: Finally, we use Latent Dirichlet Allocation (LDA) and a Structural Topic Model (STM) to identify potential structural differences in the topics covered in Wikipedia profiles of Republicans/Democrats and women/men. For LDA we ran different experiments to identify the optimal number of topics for classification.

5. Results:

Wordscores: To analyse the ideological positioning of US Congress members in their Wikipedia profiles, we use the Wordscores methodology following Laver, Benoit and Garry [9]. 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 the scores of the remaining Wikipedia profiles, the so-called virgin texts. This allowed 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 [2] allowed us to quantify how accurately Wikipedia profiles portrait the policy positions of US politicians. We first applied the methodology to a test set of Wikipedia profiles to evaluate the accuracy of our approach. With regards to the power of this method, Lowe [11] mentions that one weakness of Wordscores is that without an underlying statistical model we cannot foresee the assumptions made by the Wordscores coding therefore complicating the analysis of the results.

The Wordscores results for the reference texts are 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. The same holds true for Maxine Waters - 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 (see Appendix).

To overcome this issue, we ran 9 models (see Appendix for a full list of the models), trying to find matches for careers in terms of time - Bernie Sanders and Orrin

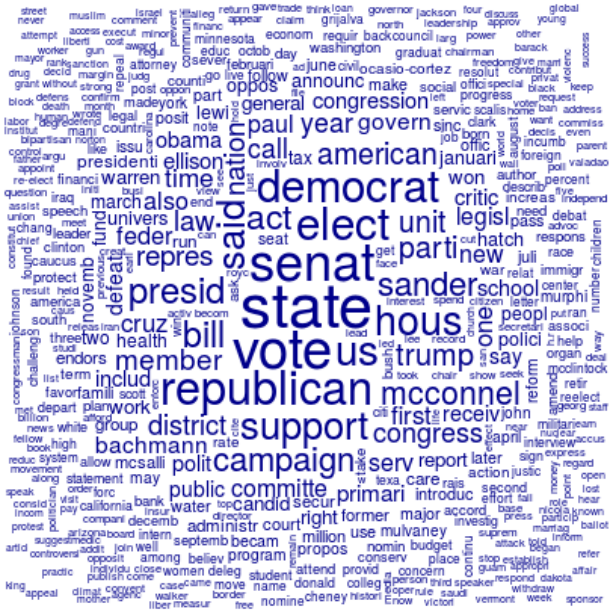


Figure 1. Top features in 32 selected Wikipedia profiles.

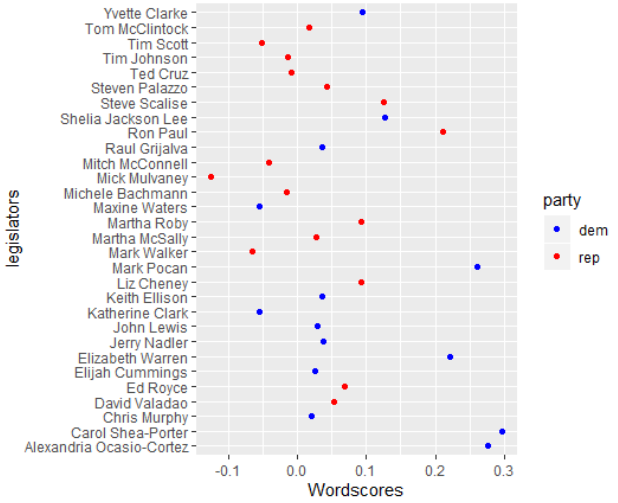


Figure 2. Wordscores comparison Sanders - Hatch (Note: Extremes removed)

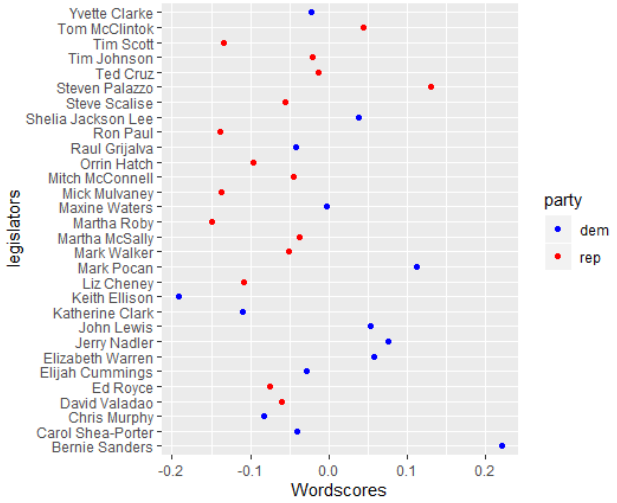


Figure 3. Wordscores comparison AOC - Bachmann (Note: Extremes removed)

Hatch, Alexandria Ocasio-Cortez and Michele Bachmann, Elizabeth Warren and Ted Cruz. The results continue to congregate around the center, as seen in Figure 2, for model (2), and the performance of the classification is not improved. The disphasement continues in most other models, as Figure 3 showcases for our third specification.

The results correspond to our theoretical expectations: Given its nature, Wikipedia is not a good showcase of a politicians' policy position and of their political leanings. The poor performance of the Wordscores approach in classifying Wikipedia profiles can be explained by the underlying model assumptions that every word is equally important to the classification and that the relevance of a word for the classification is directly proportional to its frequency in the reference texts. For Wikipedia profiles, however, the terms

appearing most frequently are related to the political system and personal biographies and are used across party lines or specific to the person. This is why the scores cluster around the center. Any words that appear in the virgin texts, but not in the reference texts do not add to the classification, which is problematic to classify policy positions because the terms used in these sections are usually specific to the person under consideration or at least to the respective policy area. This does not necessarily mean that policy positions are not accurately or sufficiently clearly displayed on Wikipedia, but rather that Wordscores is not a suitable method to measure politicians' policy positions based on Wikipedia profiles (for a discussion of the methodological limitations of the Wordscores approach see also Lowe [11]).

Naive Bayes: For the Naive Bayes model we randomly split our corpus of 793 documents into a test and a training set with a probability of 20:80. We removed stopwords, punctuation and digits for a party-based classification and kept the stopwords for a gender-based classification to not exclude terms such as "she/he" and "her/his". The party-based classification with the Naive Bayes model performs very well. In 83.3% of the cases, the party affiliation of US Congress members in the test set is accurately predicted based on the politician's Wikipedia profile. The performance of the gender-based classification is even higher: The gender is accurately predicted in 97.5% of the cases in the test set.

The high performance of the Naive Bayes classifier is in line with findings of comparative methodological studies on supervised methods for political text (see for example Mocherla, Danehy and Impey [13]). Despite the large number of descriptive terms that appear across party and gender lines, there seems to be significantly clear differences in the frequencies of certain terms that allow to distinguish profiles of Republicans and Democrats as well as male and female politicians based on the probability of words appearing in the test set. So Naive Bayes can be a useful classification method for Wikipedia entries, if the corpus and training set are sufficiently large because the algorithm can only classify the test documents based on the words appearing in the training documents.

Regularised Regression: We use Regularised Regression to better control the issues that arise from collinearity in a linear model. It is usually the case that traditional regression models produce higher correlation coefficients which are affected by overfitting and low accuracy [14]. The techniques of regularised regression aim at finding the best weighted combination of variables that predict an outcome; in the context of text analysis, it would be the best weighted combination of variables that best predicts the stated hypothesis (e.g. a legislator's party). In general practice, it has been noticed that Ridge Regression in particular performs well for text classification cases; this is due, in part, to the fact that while other methods (i.e. non-linear) might be more powerful but difficult to estimate due to their complexity. Although there may be a non-linear classifier that gives better generalisation performance than the best linear model, it is too difficult to estimate those parameters using the finite sample of training data that we have. In practice, the simpler the model, the less problems we face when estimating the parameters, with a lower tendency to over-fit and therefore obtaining better results.

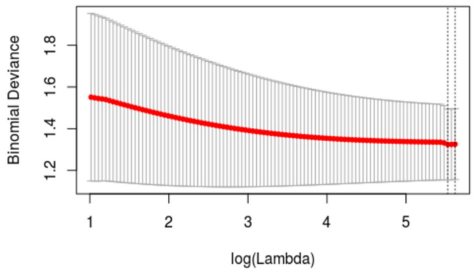


Figure 4. Ridge Regression

In our exercise, we tried the Ridge Regression on the data, using the same dataset, training set and test set as for NB, but limiting it to the general dfm (i.e. not with the two previously explained specifications of removing stopwords and gender-based classification-related words). The results however showed low accuracy power for this case, with an overall of just 60% accuracy. As compared to the Naive Bayes classifier, which exhibits levels above 80% of accuracy, the power of the Ridge Regression was considerably less. Indeed, from figure 4 we can observe that the plot showing the Mean Square Error has very broad intervals. Because of this, ridge regression was not the best classifier under supervised modelling. Additionally, in opposition to what happens with the Naive Bayes classifier, Regularised Regression didn't work as well in this case as the dataset was extremely large and demanding a large amount of processing.

LDA and STM: To identify potential structural biases in the linguistics of Wikipedia profiles across party lines and by gender we used unsupervised models. First, we applied a probabilistic LDA model (see also Blei[1]) to the entire corpus experimenting with different numbers of categories between 5 and 30. The topics as clustered by the LDA model are not clearly distinct, but include a variety of terms

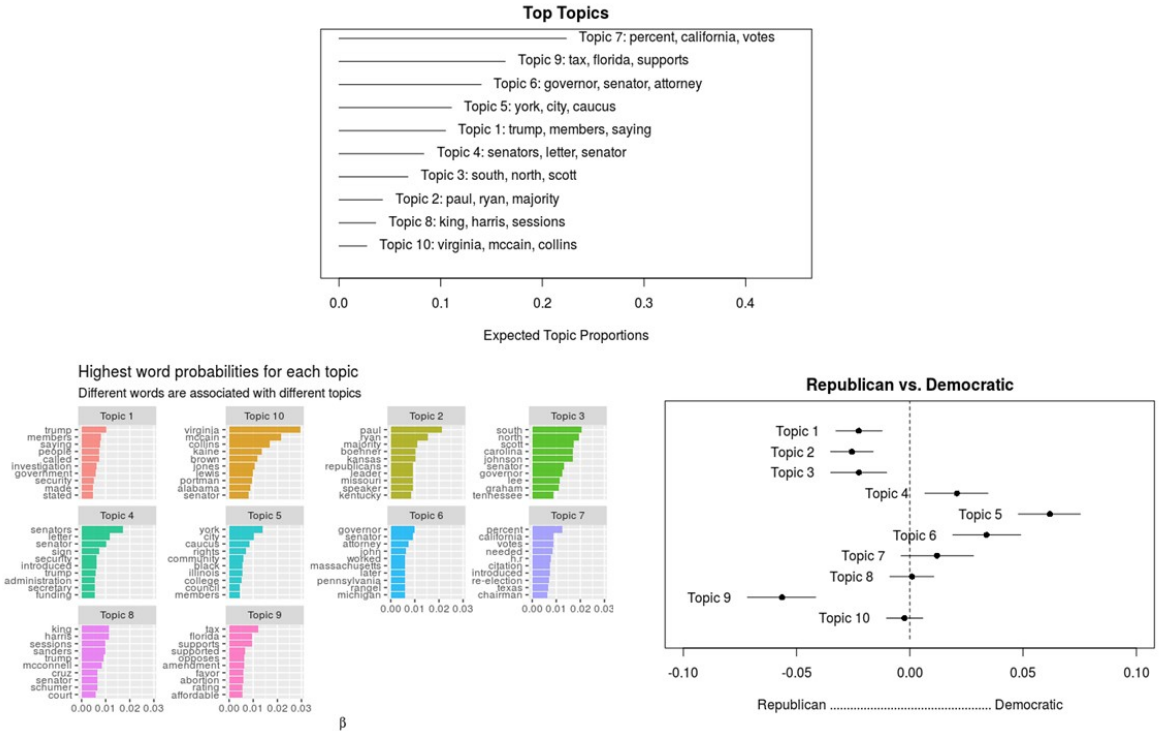


Figure 5. STM with K=10

that appear frequently in all Wikipedia profiles mainly related to the political system and personal biography, but not specifically to policy positions. There are no obvious hidden variables that might explain the structure of these topics. This finding corresponds to theoretical expectations about the linguistics of Wikipedia entries. Only a tiny fraction of the words relates to policy positions, consequently the underlying linguistic structure of Wikipedia profiles is similar across party lines and regardless of gender - at least in terms of word frequencies. Potential biases can thus not be identified using probabilistic topic models, if the bias might concern only a specific fraction of the corpus.

For the Structural Topic Model, we used the whole corpus and ran the model for 8, 10, 12 and 15 topics, in an attempt to discern if relevant topics appeared in the aggregate data of US Congress members. In addition to stopwords, we removed the most common terms we saw in our analysis: months of the year, numbers 1 to 9 and common phrases such as "said" and "us". We ran multiple iterations in each of the models, which seemed to improve the results slightly, an expected result given the generative logic of the STM model, indicating that there may be room for extended attempts with more processing power. We highlight the most relevant results in Figure 4, with a k=10 and 100 iterations but the complete results are featured in the annex.

Our results reinforced our assumptions: the contents of Wikipedia articles mean that the most common terms are fairly politically insignificant, as can be seen in the Figure 4: names of states, senators, positions occupied or common words, such as "people." The size of subsections containing biographical information can be as big as policy or voting history in many profiles, namely of less known senators, almost certainly skewing the results. However Topics 5 and 9 do seem to indicate some words that could be, in aggregate, seen as traditionally Democratic and Republican, respectively. Topic 5 mentions Illinois, black, community and caucus, all terms that would be used in reference to civil rights movements and in coordination with policies of the country's most prominent African-American politicians, namely in the Chicago area.

Topic 9 aggregates Florida, tax, abortion, and both terms for support and opposition, something that could be seen as topics around traditionally Republican issues and politicians.

However, that is the extent to which we saw any relevant results in five attempts we made with STM. In introduction of the model by Roberts et al [15], the authors say its goal

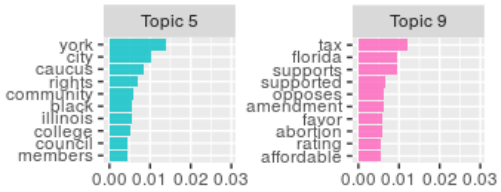


Figure 6. Topic 5 and Topic 9

is to "allow researchers to discover topics and estimate their relationship to document metadata," (p. 2), which we have been unable to do, thus stopping us from proceeding with hypotheses testing.

6. Analysis

Our findings correspond to what Cohen and Ruths [4] conclude for the classification of political attitudes on Twitter: "It's Not easy!" and standard classification methods fail to accurately classify average Twitter users because they "lack strongly differentiating language between political orientation classes" (p. 98). Whilst Twitter and Wikipedia are necessarily different mediums, both seem to not contain sufficiently large shares of politically distinctive language.

This does not necessarily mean that Wikipedia profiles are generally not suited as data for a political scientist to be able to classify political orientations or to measure policy positions of individual politicians. Rather, the existing classification methods need to be applied with caution and might benefit from adaptation to the specific data at hand when the underlying assumptions about the distribution of words and the data generating process do not apply. As the methods do not allow to weight terms by their relative importance to the classes, the classification follows the underlying structure of the text. Unsurprisingly, the dividing features of Wikipedia profiles are not party lines, but the content and language are to a large extent shared across these lines given the relative importance of sections on personal biography.

This is indeed in line with the findings of Ewonus et al. [5] for the classification of US presidential speeches. One of the main methodological problems with these models is the bag-of-words assumptions, i.e. the assumption that every word is equally informative, that the relevance is directly proportional to word frequency and that the order and context of words do not matter (see also Lowe [11]). For Naive Bayes the independence assumption is similarly problematic. From a linguistic perspective, the words that are most distinct and appear the least frequently are usually most informative and the data generating process of text is usually not such that a normal distribution of word frequencies and the independence of the probabilities of individual words appearing can reasonably be assumed. Same logic applies to Ridge Regression where all coefficients are considered and not equalized to zero (as is the case with Lasso Regression). The fact that there is some sort of penalty to control for the over-fitting still does not control for the importance of each differentiated word within the dataset. The Structural Topic Model could be a better fit given that it can test for non-linear relationships. However, our results are disappointing, given that the aggregation focuses on non-political terms and the weights do not seem to offset their lack of importance. Perhaps further iterations could perfect the model enough that results would be clearer but the literature does not indicate that more attempts would significantly improve results, given that Roberts et al. [15] and Wesslen [17] use between 50 and 75 iterations with significant success.

This is not to say that the methods can generally not be applied to political text, but rather that Wikipedia is not an inherently political source and that the results need to be interpreted carefully as there is still room for methodological improvement for this purpose. There is a lot of work evaluating how language has changed in the internet era that can be drawn from, both for methodological improvement and as a part of an analysis of factors that may influenced these models. In 'Because Internet', linguist Gretchen McCulloch provides an overview of how writing is changing due to the proliferation of computer-mediated mediums such as messaging services and social media. She calls its evolution online towards a more informal type of writing a "typographical tone of voice" and argues that technology has facilitated and accelerated a conceptualization of language that is less formal – less dictionary in the traditional sense, and more network and crowd-sourced thinking.

As an example of this, we examined the amount of edits and contributors to the 32 profiles we used for our initial Wordscores model. The sheer numbers of additions and removals - and its frequency - must be considered a major factor when analysing Wikipedia data. Political manifestos or speeches are highly coded language, full of significance and signifiers and they are ultimately put together by a relatively small group of people, especially in comparison to Wikipedia articles. Ron Paul's entry has been changed over 10,000 times, by over 3,300 people in a time span of 16 year [21] ; roughly over the same period, Bernie Sanders' article has been edited nearly 7,000 times by over 2,000 people [19]. The same is also true of prominent newer members of Congress, such as Michelle Bachmann and Alexandria Ocasio-Cortez, which have seen their article edited over

5,000 times by 1,600 people [20] and 3,000 times by 770 people [18], respectively. We posit that what the models see as lexical diversity is both true and misleading: There is coherence in a "networked" corpus but there are other inherent features, namely information density and selection, that we do not yet know enough about that could be possible hidden features influencing our models.

Lawmaker	Number of edits
Alexandria Ocasio-Cortez	3,006
Bernie Sanders	6,914
Elizabeth Warren	4,805
John Lewis	1,918
Liz Cheney	994
Maxine Waters	2,062
Michelle Bachmann	5,046
Mitch McConnell	3,825
Ron Paul	10,917
Ted Cruz	5,204
Tim Johnson	811
Tim Scott	1,227

Table 1. :Number of edits for the 32 selected Congress members

These factors are essential to any future model that uses Wikipedia, given that its main body of source is online content and its contributions based on model that derives from many different editors, as Matei and Brin state [16].

7. Conclusions

The main finding of this paper is that, at present, the literature and state of the art are lacking a methodology to accurately portray politicians' policy positions from their Wikipedia entries, as the the current NLP methods have proved not to have the expected predictive power. It is hard to accurately assess if these findings derive from using the wrong analytical tools for the style of text; if Wikipedia entries for politicians are incomplete or not representative of their stances; or a combination of both. Furthermore, it is difficult to extrapolate the results and conclusions of our analysis beyond our sample and the assumptions we have done here, given the visibility and interest US politicians have. Therefore, we cannot say the results would (or wouldn't) be the same with a different corpus or with a different language/edition of Wikipedia, or with a different approach which focuses on specific treats for the Wikipedia construct. Given the ongoing developments on NLP methodologies, it could be expected that new ideas on how to approach and analyse this sort of data will be introduced. An analysis of linguistic features of German Wikipedia by Beißwenger et al [12] reaches similar conclusions, stating that "off the shelf" tools for the linguistic annotation of written language data do not perform on CMC [ed. computer-mediated communication] data in a satisfying way", elaborating on extensive suggestions for significant adjustments to machine learning models for CMC.

As such, our biggest learning is that, in fact, text-as-data processing models can only be, in fact, be seeing as part of a "pipeline" of treatments and analysis and may not always, in and of themselves, be able to generate results that can be extrapolated from.

Most NLP tools available have been designed for highly structured text, namely political manifestos. The tone, information density and structure of Wikipedia – and its irregular nature – make it so that it is inherently different from political manifestos or speeches, which tend to be rich in political signifiers. However, there is a lack of formal academic research that evaluates exactly how Wikipedia editors use language and its organisational logic, and as such, an interdisciplinary project on how to develop a methodology for non-political text with an (internet) linguist could be conceived as a possible future area of research.

8. Acknowledgements

We would like to thank Simon Munzert for initial guidance on how to work with the Wikipedia corpus and the WikipediR and LegislatoR packages. For the wordscores approach we adapted the code provided by https://uclspp.github.io/PUBLG088/wordscore.html to our corpus. We would also like to thank Sorin Adam Matei for granting us access to his book, Structural Differentiation in Social Media.

9. Contributions

Abhas was responsible for initial setup of the corpus, as well as the coding of the STM model. Anna Hundehege took the lead on the methodological soundness of the overall paper, including further research into text classifiers on social networks, as well as overall coherence. Sofia Diogo Mateus designed the political approach and conducted the analysis for wordscore models, as well as the research of linguistic implications of the social models of Wikipedia article creation. Maria Camila Garcia focused on the theoretical background from wordscores and regularised regression, as well as supporting the coding for both of them in conjunction with the coding of visual representations.

10. References

Github Repository

abhast/TADA-Group-Assignment/

References

[1] D. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[2] 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.

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

[4] R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! *Seventh International AAAI Conference on Weblogs and Social Media*, pages 91–99, 2013.

[5] B. Ewonus, B. McCann, and N. Roth. Cs 229 final project - party predictor: Predicting political affiliation. *Stanford University*, 2013.

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

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

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

[9] 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.

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

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

[12] E. M. C. P. Michael Beißwenger, Harald Lüngen. Mining corpora of computer-mediated communication: Analysis of linguistic features in wikipedia talk pages using machine learning methods. *Workshop Proceedings of the 12th KONVENS 2014*, 2014.

[13] S. Mocherla, A. Danehy, and C. Impey. Evaluation of naive bayes and support vector machines for wikipedia. *Applied Artificial Intelligence*, 31(9-10):733–744, 2017.

[14] A. Rahman, M. E. Gabriel, and M. Thevaraja. Recent developments in data science: comparing linear, ridge and lasso regression techniques using wine data. *DISP'19 Oxford*, 2019.

[15] M. E. Robers, B. M. Stewart, and D. Tingley. stm: R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40, 2019.

[16] B. C. B. Sorin Adam Matei. *Structural Differentiation in Social Media*. Lecture Notes in Social Networks. Springer Nature, 2017.

[17] R. Wesslen. Computer-assisted text analysis for social science: Topic models and beyond. –, –(–):–, 2018.

[18] Wikipedia. Alexandria ocasio-cortez page statistics.

[19] Wikipedia. Bernie sanders page statistics.

[20] Wikipedia. Michelle bachmann page statistics.

[21] Wikipedia. Ron paul page statistics.

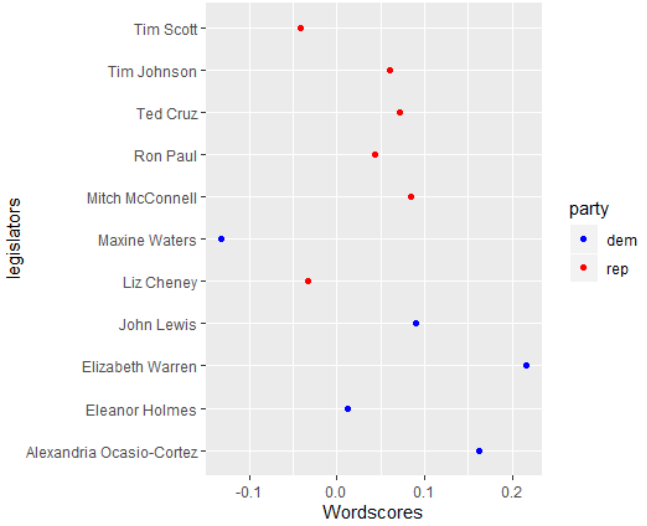
11. *Appendix

Wordscores: 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 2. Wordscore models created

The first attempt for this method was choosing Bernie Sanders (+1) and Michelle Bachmann (-1).



Wordscores model with Michelle Bachmann at the most far-right Republican, with -1, and Bernie Sanders as the most far-left Democrat, with +1 (Note: Extremes removed to facilitate analysis of the aggregation at the center, which is characteristic of the model)

LDA: In addition to the main graphs showed in the main text, we also analyzed results thrown by LDA with different specifications of K and iterations.

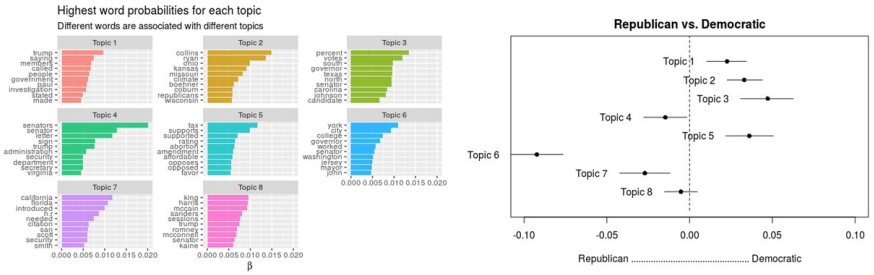


Figure 7. K=8 and 100 iterations

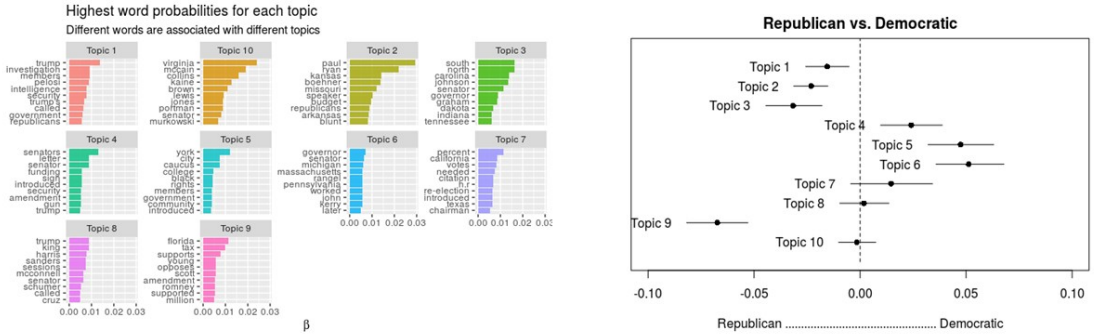


Figure 8. K=10 and 15 iterations

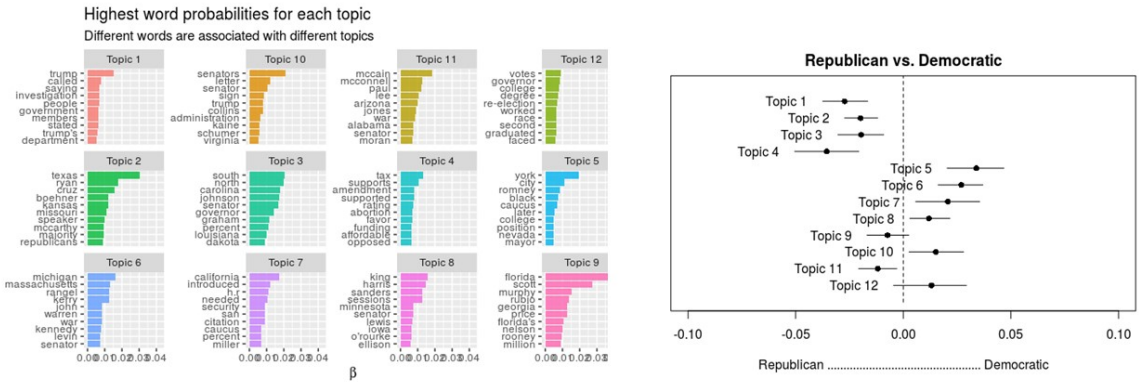


Figure 9. K=12 and 15 iterations

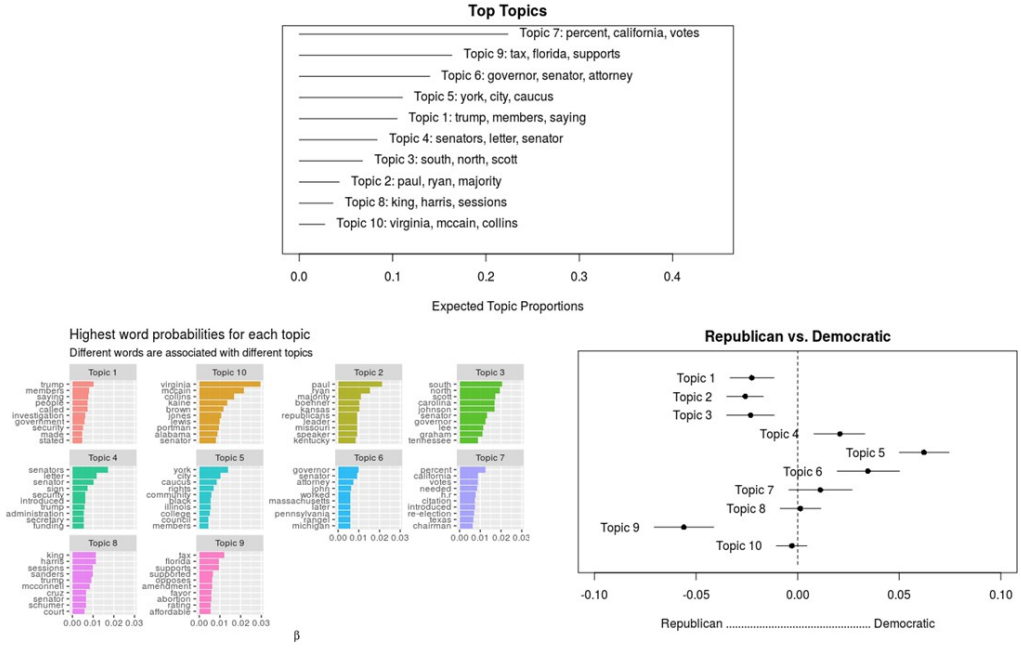


Figure 10. K=10 and 50 iterations

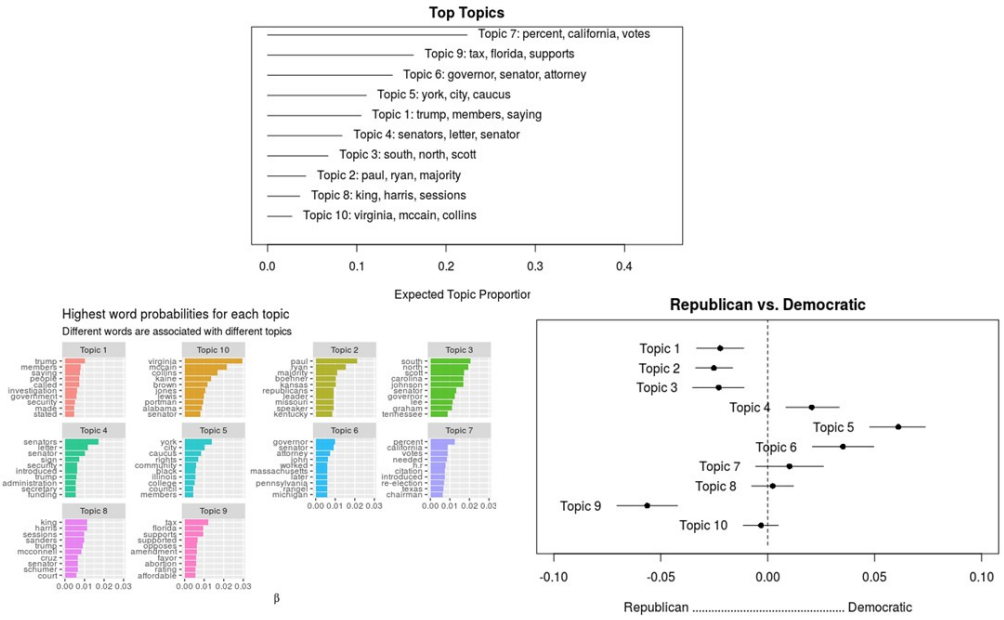


Figure 11. K=10 and 75 iterations