7-24-2023

Understanding political polarization using language models: A dataset and method

Samiran Gode
Carnegie Mellon University

Supreeth Bare
Carnegie Mellon University

Bhiksha Raj
Carnegie Mellon University & Mohamed bin Zayed University of Artificial Intelligence

Hyungon Yoo
Carnegie Mellon University

Recommended Citation

S. Gode et al., "Understanding political polarization using language models: A dataset and method," *AI Magazine*, vol. 44, no. 3, pp. 248 - 254, Jul 2023.
The definitive version is available at https://doi.org/10.1002/aaai.12104

Received: 21 July 2023 | Accepted: 24 July 2023

DOI: 10.1002/aaai.12104

SPECIAL TOPIC ARTICLE

Understanding political polarization using language models: A dataset and method

Samiran Gode[1] | **Supreeth Bare**[1] | **Bhiksha Raj**[1,2] | **Hyungon Yoo**[1]

[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

[2]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

Correspondence
Samiran Gode, Carnegie Mellon University, Pittsburgh, PA, USA.
Email: samiran.gode.applications@gmail.com

Abstract

Our paper aims to analyze political polarization in US political system using language models, and thereby help candidates make an informed decision. The availability of this information will help voters understand their candidates' views on the economy, healthcare, education, and other social issues. Our main contributions are a dataset extracted from Wikipedia that spans the past 120 years and a language model-based method that helps analyze how polarized a candidate is. Our data are divided into two parts, background information and political information about a candidate, since our hypothesis is that the political views of a candidate should be based on reason and be independent of factors such as birthplace, alma mater, and so forth. We further split this data into four phases chronologically, to help understand if and how the polarization amongst candidates changes. This data has been cleaned to remove biases. To understand the polarization, we begin by showing results from some classical language models in Word2Vec and Doc2Vec. And then use more powerful techniques like the Longformer, a transformer-based encoder, to assimilate more information and find the nearest neighbors of each candidate based on their political view and their background. The code and data for the project will be available here: "https://github.com/samirangode/Understanding_Polarization"

INTRODUCTION

Polarization among the two main parties in the United States, Republican and Democratic, has been studied for a long time (Poole and Rosenthal 1984; KhudaBukhsh et al. 2021). A lot of the discussion online has become polarized (Jiang, Robertson, and Wilson 2020), and this discussion gets the most traction online (Jiang, Robertson, and Wilson 2020). This polarization can affect the decision-making ability of a candidate if selected (Chen, Li, and Liu 2022). In such scenarios, it is important for users to be able to separate the rhetoric and understand how polar a candidate is. With this work, we set out to ask exactly these questions, "Can we measure how polarizing a candidate is?," "Can we measure how much this polarity has changed over time?," We try to answer these questions using Natural Language-based techniques and in the process, create a dataset that will be useful for the research community in trying to understand political polarization in the United States. Though we have worked on the US political system, the methods we suggest for measuring polarization would be useful for other countries with similar democratic elections in determining how polarized a candidate is. We first try classical methods such as Word2Vec (Mikolov et al.

2013) and Doc2Vec (Le and Mikolov 2014) to understand if we can find polarization using the data we have and gain more insights. We find that words that are politically sensitive (Center 2019) are related to other words which are politically sensitive (Center 2019). We thus move on to more recent and sophisticated models built using Transformers (Vaswani et al. 2017) to gain more insight into the data. We then use these, in particular, longformers (Beltagy, Peters, and Cohan 2020) to project candidate-specific data into a particular embedding space and then use this data to find the nearest neighbors of each candidate and provide one metric to find how polarized a candidate is.

RELATED WORK

KhudaBukhsh et al. (2021) talk about political polarization online and uses machine translation to interpret political polarization on the internet. Bhatt et al. (2018) discuss the impacts of hyper-partisan websites on influencing public opinion as illustrated by their ability to affect certain events in the 2016 US general elections. The authors then go on to show how certain political biases are assumed for the purpose of their study, namely overt support for either a Democrat or Republican is taken to be an indicator of the site being either Liberal or Conservative. This paper is fundamental to our research as it looks into the political division and lays the foundation for any following work in the domain of using specific features to classify an entity as being Liberal or Conservative. The features they considered were transcripts of the content being published or shared on these sites. In our case, the features will simply be the Wikipedia page content of the people. KhudaBukhsh et al. (2022) show the polarization in TV media and fringe new networks and uses anguage model-based techniques to understand them further. However, this polarization visible in the electorate stems from the candidates. DeSilver (2022) claims that the candidates become polarized and moved away from the center over the years. With this paper, we release a dataset and a few metrics that will help us understand if political polarization exists in political candidates and how we might be able to measure this political polarization. The aim of this study is to aid voters to make informed decisions before elections. And we use language-based techniques on a dataset that is classified into four eras and divided into three parts, mainly background, political, and other.

Belcastro et al. (2020) demonstrate that political polarization can be mapped with the help of neural networks. This is almost a baseline idea as we are using attention networks and Longformer models for the same. The key difference lies in the data extraction and methodology.

Khadilkar, KhudaBukhsh, and Mitchell (2022) go in depth towards finding gender and racial bias in a large sample of Bollywood (and Hollywood) movies. The author has amalgamated several known NLP models while he tries to create a reasonably robust model of his own. The portions in which this particular study differs from those before are that the sample size is fairly large. It then diverges further with its rather innovative use of diachronic-word embedding association tests (WEAT). Other techniques that are implemented include count-based statistics dependent on a highly popular lexicon cloze test using BERT as a base model (an idea we could consider after data attention) and bias recognition using WEAT. The final model is a combination of the above three. This paper is highly relevant to our project as it uses a similar idea of our own. It uses aforementioned models to predict bias, that is, sentiment prediction. In our project, we use data to predict political sentiment and attempt to classify certain features as being precursors to classification.

Rajani et al. (2019) tried to improve speech-based models on their ability to verbalize the reasoning that they learned during training. It uses the CAGE framework (Common-Sense Auto-Generated Explanations) on the common sense explanation dataset to increase the effectiveness by 10%. It introduces improvements over the use of BiDAF++ (augmented with self-attention layer) in these newer models. It further uses NLE as rationale generalization within the second phase primarily as means for sentiment analysis. In this paper, Mturk (from Amazon) is used to generate explanations for the dataset. CAGE primarily uses a question–answer format with three options, a label and the best explanation for that label. Furthermore, other evaluation parameters affecting performance are tested and may be used in our project either as verification models or otherwise. CAGE is certainly an interesting choice for verification given the higher accuracy it attains. A factor to be considered, however, is that the types of datasets and models are very different. Thus certain modifications will be made to the above framework.

Devlin et al. (2018) are the introduction paper for BERT, a model that will be used extensively. It also shows the results of fine-tuning BERT. These indirectly or directly will be used either as pretrained constraints or as tuning methods.

Petroni et al. (2019) demonstrates the ability of pretrained high-capacity models like BERT and ELMo to be used as knowledge repositories. This is mainly based on the following three observations: (1) The relational knowledge of these models is competitive to that of an NLP with access to certain oracle knowledge, (2) the effectiveness of BERT in an open domain question answer test, and (3) the fact that certain facts are easily learnable. The authors also

demonstrate the usage of other models (unidirectional and bi-directional) in the study, namely "fariseq-fconv" and "Transformer-XL." They conclude by showing that BERT-Large is able to outperform other models and compete even with supervised models for the same task.

Palakodety, KhudaBukhsh, and Carbonell (2020) demonstrate the ability of BERT and similar LMs to track community perception, aggregate opinions and compare the popularity of political parties and candidates. This is demonstrative of our work as we intend to use BERT for the purpose of sentiment analysis. The authors conclude by stating that the LM can be used as a pipeline for extracting data in the future.

In Hamilton, Leskovec, and Jurafsky (2016), the authors try to counter the problem of word meaning changing semantically with context. They propose a robust method by using embeddings. These are then evaluated with the "Law of Conformity" and "The Law of Innovation." These display the role of frequency and polysemy in the building structural blocks of language. These blocks will be crucial for the following two reasons: (1) The meaning changes may adversely affect sentiment analysis and thus affect results. Thus frequency and polysemy must be duly curtailed, (2) the embedding research is fundamental as we are using embedding-based models. Specifically Word2vec.

DATASET DESCRIPTION

Source

Our data are sourced from the individual pages of politicians (Senators and Congress members) from the 58th to the 117th congress. We divide these into four phases, chronologically, with each phase consisting of about 14 congresses. For each congress member, we scrape the section-wise data.

Data collection and processing

We scrape Wikipedia based on the list of politicians from the Wikipedia page for each congress. For each congress member in the list, we store the label, their party, and the metadata as shown in Figure 1. For each instance, this includes their personal details and all the information from their page as a dictionary, with the heading being the keys and the content being the value. This information helps with the downstream task of cleaning. We annotate this data based on the experiment, in our case, we have manually annotated the data to classify these keys into three separate categories. (1) Background data, (2) Political data, and (3) Other; in our release, we will be releasing both

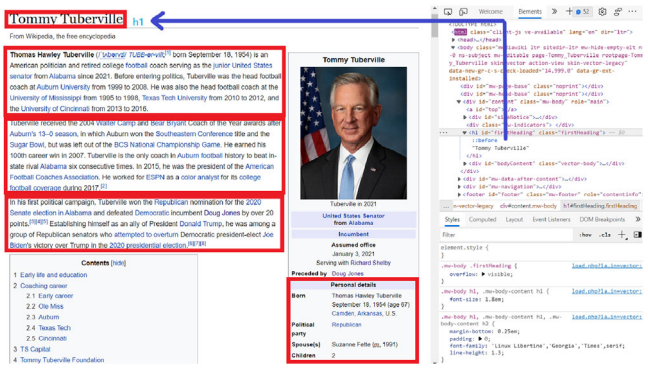

FIGURE 1 Webscraping based on each tag.

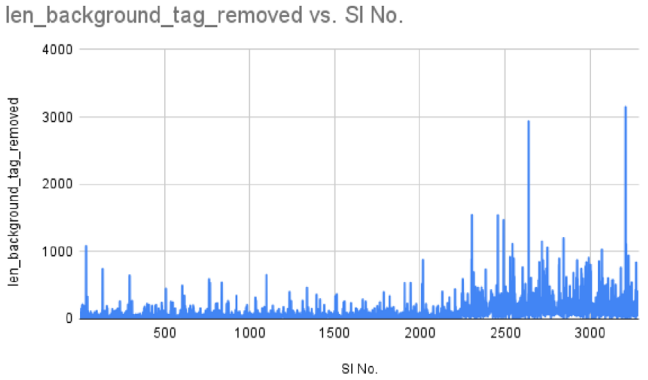

FIGURE 2 Political data length variation across candiates.

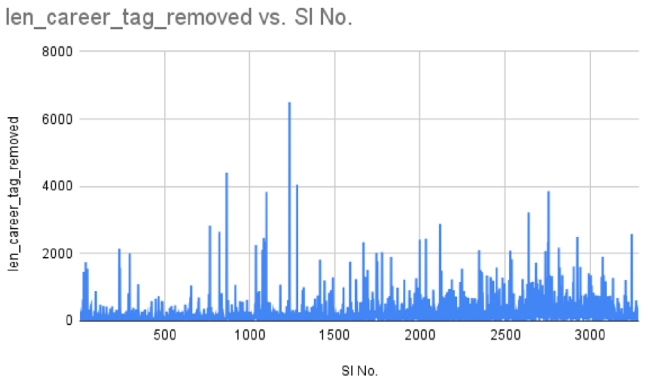

FIGURE 3 Background data length variation across candidates.

the annotated and raw versions to facilitate custom use. The distribution of the length of Political and Background text for all politicians is shown in Figures 2 and 3, respectively. Wikipedia page sections do not have a fixed format, each politician has different key sections. For instance, Early Life and Background can be split into many sections such as Education, Career, Family, Personal History, and so forth. So all these sections are grouped into a single category Background. Similarly, anything related to their political affiliation, elections, campaigns, and positions

held during their tenure are categorized into a single annotation Political Career. All other categories such as Awards, Controversies, Business-related activity, and Post political career are clubbed under the Others category. This way, only relevant data are selected under each category by manually changing the annotation based on the content inside each category. To conclude, for just Phase 4, a total of 1656 categories were merged into three categories for 1631 instances in the first pass spread over roughly 26 years (1995–2021). This data still contains information names, organizations, locations, numbers, and so forth, which need to be cleaned. We first run a NER model on the data to remove the names and organization. However, we remove location names only from the political section. The reasoning behind this is to make sure that information from the political section is not influenced by location information. However, for background, we want to understand where a person was born and raised affects their political views and for this only, this was kept but others were deleted. This information, after the NER, is passed to remove numbers and other irrelevant regular expressions. This makes sure that the data being passed for other downstream tasks are clean and gives unbiased answers.

LANGUAGE MODEL

Natural Language Processing-based applications have been dominated by transformer-based language models where models like BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019) have been state of the art since 2018 but when it comes to our dataset, these models have a drawback, that is, their ability to process longer sequences since the cost of attention grows on the order of $O(N^2)$. Longformer (Beltagy, Peters, and Cohan 2020) and other variants are useful for this task, they accept 4096 input tokens as opposed to 512 for BERT. It reduces model complexity by reformulating the self-attention computation. The performance of Longformer against the current SOTA is represented by the table present below on the raw data.

EXPERIMENTS

Preliminary experiments

Our initial experiments were aimed at gaining insights about patterns or trends that might be present in our data, and also questioning if polarization exists. We do these preliminary experiments using the Doc2Vec (Le and Mikolov 2014) and Word2Vec (Mikolov et al. 2013) models.

TABLE 1 K-means classification results.

Model	Data	Accuracy
Doc2Vec	Political	59.520
Doc2Vec	Background	61.846
Allenai/longformer-large-4096	Political	52.128
Allenai/longformer-large-4096	Background	56.383

TABLE 2 Binary SVM classification results.

Model	Data	Accuracy
Doc2Vec	Political	72.872
Doc2Vec	Background	63.564
Allenai/longformer-large-4096	Political	76.862
Allenai/longformer-large-4096	Background	69.681

The Doc2Vec model was built from scratch with the raw data, where each Wikipedia page is considered to be a document. We first use the Doc2Vec model with K-means clustering and get a classification accuracy of 59.52% with political data and 61.846% with background data as shown in Table 1. We then used the same Doc2Vec model with binary SVM classifier and achieved an accuracy of 72.872% with political data and 63.564% with background data as shown in Table 2. These results are summarized in the table presented below. The Word2Vec tests were run on pretrained models as well as models we built from scratch and trained using the data we collected. We used the Word2Vec approach to find approximate nearest neighbors and exact nearest neighbors for certain words on both the Democratic and the Republican sides. This nearest-neighbor approach led to some interesting insights. We expected to see some disparity in the nearest neighbor searches for the Republican data and Democratic data basis the assumption that there is polarization. However using the simple Word2Vec models, the 15 nearest neighbors we got were quite similar but as there were certain words for whom the order of the neighbors changed based on the party, for example, for the word "GUN," "VIOLENCE" is the 2nd nearest neighbor (approximate nearest neighbor using spotify's annoy algorithm) for democratic data; however, the same word is 9th for the republican case, similarly the word "CHECKS" is the 3rd nearest neighbor for democrats while it is the 8th for republicans. Similarly, for the word immigrant, the nearest neighbors vary with time and political party as show in Figures 4 and 5. There are more such interesting examples, which coupled with the results from the Doc2Vec classification results, prove that political polarization exists and can be learned using Natural Language Processing-based techniques.

FIGURE 4 Nearest neighboring words to the word immigrant in the democratic corpus across time from left to right, as we can see, words like Americans were closely associated in the early 20th century.

FIGURE 5 Nearest neighboring words to the word immigrant in the republican corpus, the words are similar in the 1st era, the early 20th century, but in the most recent era it is related to other politically sensitive words.

Main analysis

As part of our preliminary analysis, we use RoBERTa, we notice the removal of the words "Democratic," "Republican," and so forth. causing a drop in classification. This is expected as we lose obvious information and classifying just based on the first 512 tokens is challenging. We, hence, use Longformer since it can consider 4096 tokens at a time. As expected, this increases the score significantly, as can be seen in the table. There are two versions of Longformer—"longformer-base-4096" and "longformer-large-4096." Longformer base provides a significant improvement over other models such as RoBERTa and simpler models such as Doc2Vec as shown in Tables 1 and 2. Longformer large provided a better score and has been the best performing model when it comes to classifying a given candidate's political party. We use this information to understand how different scores are affected by different words and relate this with our broader aim of Political polarization. For that, we calculate the global attention scores of the last layer and then find the words that have the highest attention scores for self-attention with the $< s >$ token. This has shown some interesting results, for example, for Ted Stevens, a Republican, some obvious words like, "public," "federal," "legislature,", "Wisconsin" show up higher, which is expected since the main information from the political text is related to their work, but the word "abortion" showed up in the top 10 percentile words, more such analysis is being done, which we believe will give more interesting results, the above analysis for simpler models like BERT is not as impressive since the information is local and for Longformer, this not trivial since Longformer looks as context using sliding windows, however, the Longformer architecture allows for certain tokens to have global attention and choosing the CLS token allows us to look at the attention of all 4096 tokens with this word. Another hypothesis that we have been testing is that the background of a candidate can also help us identify the political leaning of a person which if the world was not polarized would not be the case and only the political information would help us classify, however as can be seen in the table, the background matters significantly as well.

RESULTS

We tested different models with the annotated raw dataset to understand the polarization in the text. The three models were tested with both the political career dataset and the background dataset to get insight into the factors that influence political polarization. The obtained results are presented in the following table.

Apart from these accuracy tests, we also leverage the attention mechanism of the Longformer model. We find the words with the highest attention scores to correlate them with our theory of political polarization. We can also design an interactive website for voters that helps you to understand if polarization exists. The website finds the nearest neighbors of the selected politicians from the Longformer output. Then depending on the ratio of Republicans to Democrats in the nearest neighbors, we estimate the politician's polarization.

In the graph, the x-axis is the rank of the 20 closest neighbors for the politician you choose given the dataset and the y-axis shows their respective closeness scores. The color blue is for Democrats and red for Republicans. The ratio is Blue versus Red points in this graph, so one of our hypotheses is that if a politician is not polarized, this ratio should be 0.5 (democrat/total) if we just look at the background data. The above graph in Figure 6 shows the neighbors for Mitch McConnell (Republican) and it is very evident that majority of the neighbors are Republican (red in color) whereas the Democrat count is only 2 out of 20. So one can infer that the polarization ratio is 0.9 for Mitch McConnell. Similarly, in Figure 7, we can see that Ayenna Presley, who is a Democrat, has 16 neighbors belonging to the same party resulting in a ratio of 0.8 for background data. Scaled-up versions of such websites with more metrics that highlight

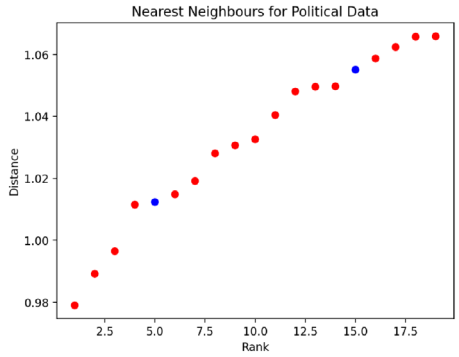

FIGURE 6 Nearest neighbors for political data belonging to Mitch McConnell (Republican).

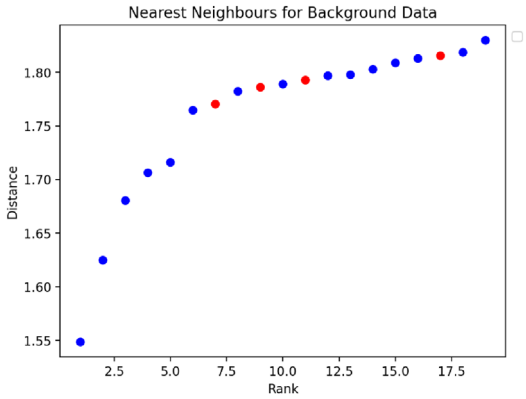

FIGURE 7 Nearest neighbors for background data belonging to Ayenna Presley (Democrat).

the political views of a member as radical or moderate will be beneficial to the voters.

Also, we show the worthiness of this data and hope that this will be useful to the research community in examining the idea of political polarization in candidates and how it is linked to other attributes of the candidate. Also, to understand the views of a candidate and measure how polarizing their views are. We also hope that the spread of this data across multiple decades will help us understand how political ideas have changed over time.

FUTURE WORK

For future work, we aim to use other metrics for finding the political polarization of individuals and communities again using the Wikipedia dataset. Specifically, we want to use the attention tokens mentioned above to look at the ratio of tokens from the background to the political given text from a candidate that is equally distributed across the background and political.

ACKNOWLEDGMENTS

We would like to thank Yash Jain and Viraj Ranade for their contributions. This was a course project part of the course 11785, Intro to Deep Learning at CMU.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

ORCID

Samiran Gode https://orcid.org/0009-0009-3324-9463

REFERENCES

Belcastro, L., R. Cantini, F. Marozzo, D. Talia, and P. Trunfio. 2020. "Learning Political Polarization on Social Media Using Neural Networks." *IEEE Access* 8: 47177–87.

Beltagy, I., M. E. Peters, and A. Cohan. 2020. "Longformer: The Long-Document Transformer." *arXiv preprint arXiv:2004.05150*. https://arxiv.org/abs/2004.05150

Bhatt, S., S. Joglekar, S. Bano, and N. Sastry. 2018. "Illuminating an Ecosystem of Partisan Websites." In *Companion Proceedings of The Web Conference 2018*, 545–54.

Chen, Z., Z. Li, and S. Liu. 2022. "The Price of Political Polarization: Evidence From Municipal Issuers During the Coronavirus Pandemic." *Finance Research Letters* 47: 102781.

DeSilver, D. 2022. The polarization in today's Congress has roots that go back decades. Pew Research Center. https://www.pewresearch.org/short-reads/2022/03/10/the-polarization-in-todays-congress-has-roots-that-go-back-decades/

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *arXiv preprint arXiv:1605.09096*. https://arxiv.org/abs/1605.09096

Jiang, S., R. E. Robertson, and C. Wilson. 2020. "Reasoning About Political Bias in Content Moderation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13669–72.

Khadilkar, K., A. R. KhudaBukhsh, and T. M. Mitchell. 2022. "Gender Bias, Social Bias, and Representation: 70 Years of BHollywood." *Patterns* 3(2): 100409.

KhudaBukhsh, A. R., R. Sarkar, M. S. Kamlet, and T. Mitchell. 2021. "We Don't Speak the Same Language: Interpreting Polarization through Machine Translation." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14893–14901.

KhudaBukhsh, A. R., R. Sarkar, M. S. Kamlet, and T. M. Mitchell. 2022. "Fringe News Networks: Dynamics of US News Viewership Following the 2020 Presidential Election." In *14th ACM Web Science Conference 2022*, 269–78.

Le, Q., and T. Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *International Conference on Machine Learning*, 1188–96. PMLR.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781

Palakodety, S., A. R. KhudaBukhsh, and J. G. Carbonell. 2020. "Mining Insights from Large-Scale Corpora Using Fine-Tuned Language Models." In *ECAI 2020*, 1890–7. IOS Press.

Petroni, F., T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel. 2019. "Language Models as Knowledge Bases?" *arXiv preprint arXiv:1909.01066*. https://arxiv.org/abs/1909.01066

Pew Research Center. 2019. In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions. https://www.pewresearch.org/politics/2019/12/17/in-a-politically-polarized-era-sharp-divides-in-both-partisan-coalitions/

Poole, K. T., and H. Rosenthal. 1984. "The Polarization of American Politics." *The Journal of Politics* 46(4): 1061–79.

Rajani, N. F., B. McCann, C. Xiong, and R. Socher. 2019. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning." *arXiv preprint arXiv:1906.02361*. https://arxiv.org/abs/1906.02361

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems*, 30.

How to cite this article: Gode, S., S. Bare, B. Raj, and H. Yoo. 2023. "Understanding political polarization using language models: A dataset and method." *AI Magazine* 44: 248–254. https://doi.org/10.1002/aaai.12104

AUTHOR BIOGRAPHIES

Samiran Gode is a MS Mechanical Engineering graduate from CMU where they focused on Peception for Robots. Their work also focuses on spatial AI and sensor fusion.

Supreeth Bare is a MS Electrical and Computer Engineering graduate from CMU now working with Oracle while at CMU they focused on Applied ML. Their work also focuses on applying Large Language Models for multimodal applications.

Bhiksha Raj is a Professor at CMU in the MLSP group where they focus on Speech recognition, Audio processing, Neural networks, and Privacy/Security for voice processing.

Clay H. Yoo is a MCDS graduate from the Language Technologies Institute, CMU, they work on applied NLP in various domains. Apart from this, they also work on applications of AI in healthcare and model interpretability/robustness.