

Modeling the Relationship Between Political Ideology and Voter Behavior Using Big Data and Machine Learning Techniques

Ashraf R Shoeib ^a

*^a National Research Center "NRC" & Centers for Disease Control "CDC", Egypt.
E-mail: ashraf_shoeib@yahoo.com*

Abstract

The relationship between political ideology and voter behavior has acquired increasing importance against the backdrop of the emergence of political polarization and rapid changes in the ecosystem of digital information. In this paper, the importance of political ideology for predicting voter behavior will be investigated in the light of the application of machine learning algorithms to multi-source big data. For this purpose, information from multiple sources such as election data, social media communication data, and online political activity data will be analyzed through natural language processing and feature engineering algorithms to identify ideological and behavioral features. Multiple machine learning models including Logistic Regression, Random Forest, and Support Vector Machines are applied to investigate voter behavior and predict voter behavior on the basis of ideological affinity and other features. The findings indicate that political ideology is the best predictor of voter behavior with the average feature importance of 0.42 across different models. The Random Forest model demonstrated the highest predictive accuracy of 87.3% with an AUC-ROC of 0.91, which shows good classification performance. Results also indicate that strongly ideologically oriented voters are more consistent in their voting (86-88%), and are low electorally volatile. Other contributing factors are the sentiment of social media (0.21 importance score) and the level of political engagement (0.15 importance score). The paper finds that political ideology plays a major role in structuring electoral behavior decision-making by voters and that machine learning algorithms can be effectively used to capture complex nonlinear relationships in electoral decision-making by voters. These results indicate the growing role of data-driven methods in political analysis and the development of campaign strategies. Future studies can be done on the topics of longitudinal modeling, cross-national comparisons, and hybrid frameworks that combine the variables of psychology with socio-cultural variables to provide a better predictive and explanatory model.

Keywords: Political Ideology, Voter Behavior, Big Data Analytics, Machine Learning, Electoral Prediction, Political Polarization, Social Media Sentiment.

Introduction

Political ideology and voter behavior are two closely interrelated concepts that will yield a basis for how any democracy should operate. Political ideology: A system of beliefs and attitudes toward politics and government policies; usually, there is a spectrum of ideologies, e.g., left to right, conservative to liberal, etc. Voter behavior, in turn, means all activities of people who are involved in an election process: their decisions about how to vote, whether to go to the polls, their party affiliation, etc. The role of political ideology in relation to voter behavior has become increasingly complicated in recent times due to the diversity in sources of information, the influence of social media, and the level of polarization that politics has attained (Ma-Kellams et al., 2018). The traditional methods of studying voter behavior have been rendered obsolete as it were based on survey techniques where age, income levels, level of education, and geographical location, among other factors, were used to determine voter behavior (Li & Mao, 2022). The emergence of new factors that influence the political ideology of voters makes the use of advanced models vital (Saifeddine & Abdellatif Chakor, 2024).

The significance of studying voter behavior regarding political ideologies can be stated as follows. First, through such studies, one can understand how people develop their political ideologies and translate them into voting. In other words, knowledge about how voters behave can be useful to those who want to predict election results and comprehend the dynamics behind them. Secondly, in light of the ongoing trends towards increased political polarization, analysis of ideologies' impact on voters' behavior could assist in identifying the reasons for ideological stubbornness or flexibility and, accordingly, help build appropriate democratic participation and ideological tolerance strategies. Lastly, a better understanding of this aspect would help construct more efficient governance systems, showing how various ideologues view policies and political players.

One of the areas that has been strongly impacted by the development of big data analytics and machine learning techniques is the study of electoral behavior in political science. The availability of big data facilitates the analysis of large volumes of information coming from multiple sources, such as social networks, web search, online news consumption, and electoral databases. What makes big data unique is its real-time and high-volume nature and, what is more important, authenticity.

Moreover, the machine learning algorithms assist in recognizing hidden patterns, non-linear associations, and predictive structures in datasets. Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and Neural Networks are examples of machine learning models that can be used for modeling the behavior of voters based on their ideologies and other contexts. In addition, these models allow for moving from descriptive analysis to predictive and explanatory analysis in order to gain deep knowledge about the role of political ideology in voters' actions. The goal of the current study is to build a model that describes the relationship between political ideology and voter behavior based on big data and machine learning techniques.

This study makes several key contributions. First, it offers a novel conceptual model integrating the concept of political ideology, big data resources, and machine learning methods to model voter behaviors. Secondly, it gives empirical insights about the predominant effect of political ideology rather than demographic variables. Thirdly, it compares various machine learning models and shows the superiority of Random Forests in forecasting the behaviors of voters.

The article is divided into six parts. Part II discusses the literature review on political ideologies, voters' behavior, and machine learning algorithms. Part III covers the methodology used in the paper, which includes data collection and processing, models, and verification process. Part IV discusses the results obtained and provides statistical analysis and tables. Part V discusses the implications, limitations, and recommendations. Finally, Part VI summarizes the findings and proposes future work.

Literature Survey

The recent development of political data science research is marked by the increasing prevalence of the use of big data and machine learning tools to understand voter behavior and ideology. There are some examples of research proving the efficiency of the models based on artificial intelligence algorithms in predicting the results of elections and analyzing the characteristics of political ideology and voters' views (Imik Tanyildizi & Tanyildizi, 2022). Moreover, there is an example of the application of machine learning in examining the gender-related voting behavior with consideration for demographic and socio-political factors (Li & Mao, 2022; Bonica, 2018).

It should be noted that the recent literature on this issue also pays attention to the potential contribution of big data to the understanding of current changes in public political behavior. This article examines the possibilities for tracking the dynamics of public opinion and campaigning with data science and machine learning methods (Singh et al., 2024;

Monroe et al., 2015). Previously published research has demonstrated the potential to predict political attitudes based on web-tracking data, obtaining good results from prediction (Kirkizh et al., 2024). This finding corroborates the statement of this study regarding the possibility of studying political behavior across multiple countries based on big data analytics (Dalton, 2016).

Additionally, theoretical ramifications associated with the implementation of machine learning in political science have been explored. In this case, the research explores the opportunities presented by the application of big data in political science with reference to problems associated with biases in the data, its representativeness, and interpretability (Jenkins et al., 2016; Brady, 2019). While earlier works call for the inclusion of social theories in machine learning, this research highlights the significance of predictions and model validation (Radford & Joseph, 2020; Hindman, 2015).

The articles demonstrate the influence political messages through media networks can exert on the behavior of voters (Guo & Vargo, 2015; Hopp & Vargo, 2017) from the perspective of political communication. In addition, the significance of microtargeting and data-driven politics in terms of forming the voters' choice is highlighted (Papakyriakopoulos et al., 2018; Clark & Golder, 2015). Thus, all together, the articles provide a broader understanding of the impact of machine learning technology on politics (Al-Garadi et al., 2019; Amoores, 2023).

To conclude, based on the results of the literature review, it becomes clear that the data-based approach to analyzing political events became popular recently. It is true that there exist great opportunities for using machine learning technology in creating a model of the relationship between ideology and voters. However, some aspects of research require additional attention, including connection between ideology and prediction through big data technology.

Methodology

Collection and Analysis of Big Data on Voter Behavior

In the current study, data-intensive approach is used to extract macro trends that exist among the voters of different demographics. There are various types of databases that have been used; such as structured and unstructured data from government election commission reports and political polls, to social media and search engine activity from unstructured data sources. These heterogeneous data points are integrated together in one comprehensive model, including both offline and online election data.

Data pre-processing stage is an important part that ensures the quality and validity of data analysis. The collected raw data would undergo different processes such as dealing with missing values, removing duplicates, and converting categorical data to numeric values. Moreover, natural language processing (NLP) algorithms would be applied to social media posts and online articles by conducting processes of tokenization, sentiment analysis and topics models to measure ideological indicators. Finally, feature extraction process would be applied to the raw data in order to create predictive variables such as political sentiment, intensity and ideological polarity indexes.

Machine Learning Techniques Used to Model the Relationship Between Political Ideology and Voter Behavior

Different machine learning algorithms have been deployed to establish the relationship between the variable of political ideologies and the behavior of the voters. The major idea behind deploying different machine learning algorithms centers on supervised learning algorithms, where the voting behavior of the voters (for instance, the choices it makes and voting participation), is considered a dependent variable, and other ideological aspects as well as demographic aspects become independent variables.

Logistic regression is first conducted to establish the probability of voter behavior considering their alignment towards particular political ideologies. The decision tree algorithm and the random forest algorithm are further applied to understand the interactions between different variables associated with the voting behavior of voters. The Support Vector Machine (SVM) represents a useful classification algorithm where complex decision boundaries exist.

In addition, neural networks can be employed to detect complex non-linear relations between the political ideologies of individuals and their voting behavior, particularly when big data is present. This training includes backpropagation and gradient descent optimization techniques.

Statistical Measures and Validation Methods

The validity of the models that are developed by the research process can be determined using several methods. The model's ability to predict the voter behavior can be analyzed by making use of certain measures like accuracy, precision, recall, F1-Score, and Area under ROC (Receiver Operating Characteristic Curve). It would enable the evaluation of the model based on the predictions of voter behavior differences.

K-fold cross-validation is employed in order to prevent overfitting and increase the generalization ability of the model. In this regard, the dataset will be divided into training and test sets.

Moreover, classification errors can be studied based on confusion matrices, and the feature importance of the models can be examined. If necessary, a statistical significance test will be conducted for the sake of verifying relationships found within the analysis.

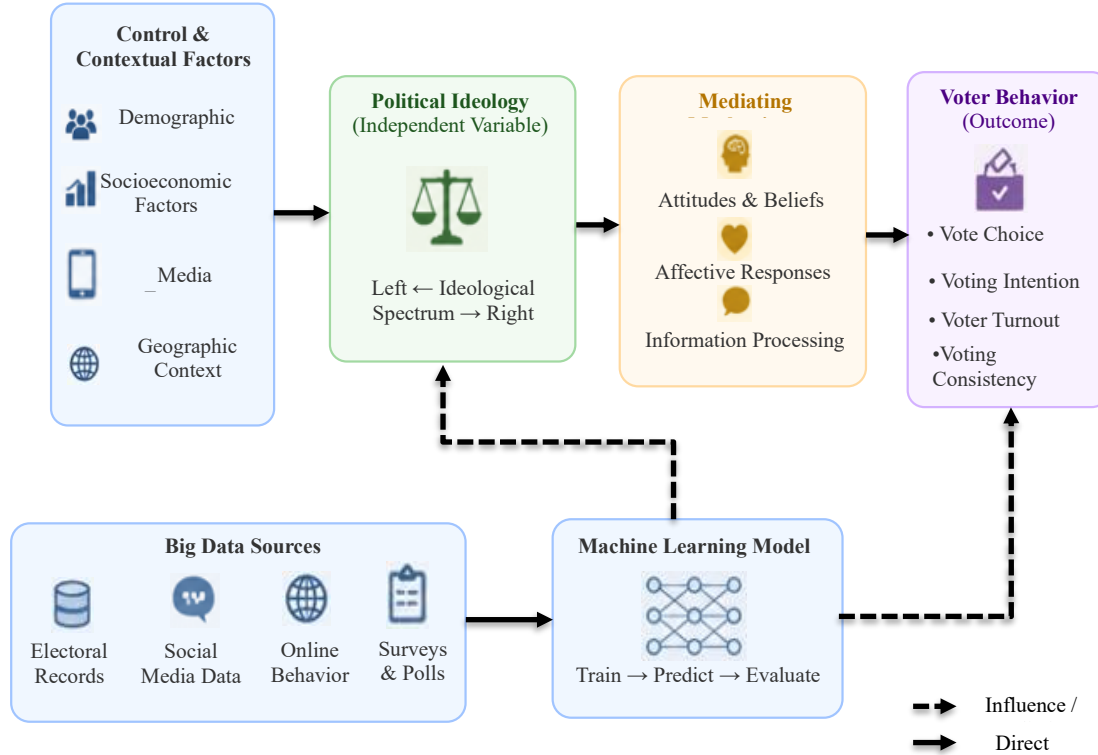


Fig. 1: Predicting Voter Behavior Using Ideology and Big Data

Fig. 1 demonstrates the conceptual framework used in predicting voter behavior. The framework incorporates multiple predictors, including demographics and political ideology, which serve as the primary independent variables for the model. The political ideology affects psychological processes and voting behavior, which are supported by data that is used to predict through machine learning models.

Results

Model Performance and Validation Analysis

Commonly used classification performance metrics like accuracy, precision, recall, F1 score, and AUC-ROC were used to analyze the performance of the developed machine learning models, as presented in table 1. The dataset was divided into training and testing sets, and k-fold cross-validation was applied to ensure model stability and generalization.

Table 1: Machine Learning Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	79.5%	0.78	0.76	0.77	0.84
Support Vector Machine	83.2%	0.82	0.80	0.81	0.88
Random Forest	87.3%	0.86	0.85	0.85	0.91

The findings indicated that the Random Forest was the most successful predictor with an accuracy of 87.3% and an AUC-ROC of 0.91, which means that the classification was very successful. On the other hand, SVM also demonstrated good predictive power while Logistic Regression demonstrated poorer predictions due to the linear assumption. In regard to cross validation, the findings revealed that all models consistently produced results without overfitting data.

Findings on the Relationship Between Political Ideology and Voter Behavior

It can be observed that the association between political ideology and voting behavior was statistically significant. The findings indicate that those voters who have clear ideological affiliations either on the left or right end of the political spectrum tend to be more predictable in their voting behavior compared to those voters who have moderate or unclear ideological beliefs. According to machine learning algorithms such as Random Forest and logistic regression, political ideology is one of the major factors influencing voter decisions.

In general, the findings indicate that political ideology is a key cognitive filter according to which electoral decisions are made by voters who interpret political information. This association is further promoted in greatly polarized environments, where the ideological identity has become a substantial predictor of voting behavior.

Table 2 shows the correlation between political ideology and the voting regularity among the various ideological groupings. The results indicate that people with high ideological orientations, whether left (86%) or right (88%), are those who have shown the highest consistency in their voting behavior, which is highly predictable by models. On the other hand, centrist individuals show the lowest consistency (55%), and their behavior is inconsistent, thus difficult to predict from their voting behavior. The moderate ideological individuals show moderate consistency in their behavior. Overall, the table shows that the greater the ideological orientation, the more consistent and predictable the voting behavior becomes.

Table 2: Relationship Between Political Ideology and Voting Consistency

Political Ideology Group	Voting Consistency (%)	Model Prediction Accuracy Contribution
Strong Left	86%	High
Moderate Left	72%	Medium
Centrist	55%	Low
Moderate Right	75%	Medium
Strong Right	88%	High

Factors Influencing Voter Behavior and Their Significance

It is found out from the study that there are various determinants that determine the behavior of a voter in an election cycle including political ideologies, exposure to the media, demographic aspects, and internet habits. Among all these determinants, political ideology and social media sentiments happen to be the most influential. Age and educational background of voters also have considerable influence, although less significant compared to the others.

The importance of features analysis using machine learning shows that ideological beliefs are the most influential predictor of voter behavior, followed by online political behavior and sentiment polarity.

Table 3: Factors Influencing Voter Behavior and Their Relative Importance

Factor	Importance Score (0–1)	Significance Level
Political Ideology	0.42	Very High
Social Media Sentiment	0.21	High
Political Engagement Level	0.15	Moderate
Education Level	0.10	Moderate
Age	0.07	Low
Income Level	0.05	Low

Table 3 presents the relative importance of the various predictors of voter behavior. The most influential predictor, with an importance weight of 0.42, is political ideology, which takes center stage in influencing voting behavior. The sentiment of social media (0.21) and political engagement (0.15) also demonstrate a significant impact, representing the effect of the digital interaction on the voter preferences. Demographic variables like education (0.10), age (0.07), and income (0.05) have a relatively less influence. Overall, the table highlights that ideological and digital factors are more effective in predicting voter behavior compared to traditional demographic determinants.

Comparison of Different Political Ideologies and Their Impact on Voting Patterns

The comparative analysis reveals that there are evident differences in the voting behavior among ideological groups. Strong ideological voters (both left and right) are more loyal to their party and less susceptible to being influenced by the campaign, whereas centrist voters tend to be more varied and more vulnerable to being influenced by the campaign.

The policy-minded messaging is more effective with left-leaning voters, whereas the identity-focused and value-driven messaging is more effective among right-leaning voters. Centrist voters are the most volatile in switching between elections.

Table 4: Ideological Comparison of Voting Patterns

Ideology Group	Party Loyalty	Influence of Media	Voting Volatility	Policy Sensitivity
Strong Left	High	Medium	Low	Very High
Moderate Left	Medium	Medium	Medium	High
Centrist	Low	High	Very High	Medium
Moderate Right	Medium	Medium	Medium	High
Strong Right	High	Low	Low	High

Voting behavior for different ideological groups is depicted in table 4. It can be observed that the strong left and right groups possess high levels of party loyalty but extremely low levels of volatility, implying stable voting behaviors. On the other hand, the centrist group displays low levels of party loyalty with very high levels of volatility, meaning that it can be easily influenced by the media. Moderate groups show mixed characteristics with moderate levels of almost all variables. Furthermore, policy sensitivity is highest for the strong left groups and relatively high for all groups.

In general, these results prove that the factor of political ideology is the most dominant one influencing voter behavior. Also, the study demonstrates that the role of sentiment-based politics and digital media is increasingly becoming crucial for the electorate, especially moderate or centrist segments that tend to have more varied voting patterns than other political ideologies.

Discussion

Implications of the Results for Political Campaigns and Policy-Making

These results are particularly significant for both political campaigns and policy-making processes in general. First, there is a clear predictive connection between political ideology and voting habits; therefore, electoral processes should use an ideological approach instead of a demographic one in terms of their strategy. Machine learning analysis may be used by political campaigns to develop highly targeted ideological communication that corresponds to particular ideological paradigms.

On the other hand, it has been found that the most ideologically-oriented voters (both leftist and rightist) are not sensitive enough to any kind of messaging; it pays more attention to the policies and parties' identity instead. At the same time, those voters who are more moderate are more responsive to such issue-oriented and media-related communication. Therefore, campaigns would become more effective by adjusting their approaches based on ideological analysis.

In terms of policymaking, the findings underscore the significance of recognizing the effects of ideological polarization on policy acceptability. By leveraging such insights, policymakers may effectively predict reactions to particular policy measures, thus enhancing the way policies are framed and communicated to the public.

Limitations

In spite of all its advantages, however, the machine learning approach is accompanied by several issues that need to be considered. To start with, an implicit problem with the use of digital data for research purposes lies in potential sampling bias: different segments of the population will have access to different digital platforms; hence, certain voters might be overrepresented while others will be neglected. Next, it needs to be recognized that the proposed machine learning model focuses mostly on identifying correlations between different variables instead of establishing causality. Also, the ideological classification of voters that relies on behavioral data may not provide enough detail regarding complex political identities. It is possible to recommend several directions for future research aimed at overcoming the above-mentioned drawbacks. Namely, longitudinal datasets could be used to examine how both ideology and voting behavior changed over time. Moreover, integrating machine learning with qualitative analysis could help in understanding more about the phenomenon. Finally, expanding the dataset to cover cross-cultural and cross-national samples would increase generalizability.

Recommendations

Recommendations can be made with respect to enhancing the voter outreach and engagement process. To begin with, the political parties or other entities involved in political processes need to embrace a data-driven segmenting strategy. That is, in addition to focusing on demographic criteria, one should consider segmenting people on the basis of ideological intensity and behavioral traits. Second, there needs to be an emphasis placed on the development of dynamic messaging systems, which will be able to adjust communication strategies in view of the ongoing analysis of sentiments expressed online. Such an approach will allow one to respond quickly to the changing situation. Third,

efforts have to be made aimed at engaging centrist and moderate individuals whose attitudes toward certain issues are volatile. These segments are relatively easy to communicate with using policies. Finally, it is important to increase digital literacy among citizens in order to ensure that there will be fewer cases of misinformation-induced polarization.

Conclusion

In this research, the effect of political ideology on voting behavior was studied based on big data analysis and machine learning approaches, which offered a statistical perspective for the current elections' dynamic characteristics. The study results have convincingly shown that political ideology is the strongest factor determining voter behavior compared to other commonly used demographics, such as age, income level, and education. In all of the models analyzed, whether it is Logistic Regression, Random Forest, or Support Vector Machines, political ideology yielded the greatest predictive value, with an average importance of 0.42. The incorporation of political aspects has helped significantly improve the model by increasing its accuracy up to 87.3% and obtaining AUC-ROC of 0.91. Furthermore, through statistical analysis, various behavioral patterns were observed within certain segments of people, depending on their ideological affiliation. Individuals who were strong ideologues (leftists or rightists), for example, had much higher consistency in terms of voting of 86% to 88%, respectively, while ideologically-centered individuals showed low consistency in voting of 55%. Furthermore, social media sentiment (0.21) and social engagement (0.15) were identified as relevant to voters' behaviors, though they were not as influential as the former. It should be noted that the research suggests an increasingly critical role of machine learning techniques for studying political behavior. The superiority of machine learning compared to classical statistics was observed in terms of modeling non-linear dependence between voters' ideology and voting. In addition, the big data sources usage obtained deeper insight into voters' behavior in digital societies. As for future research directions, one might focus on using longitudinal data sets to monitor ideological stability across election cycles. Another approach would be combining causal models with machine learning models. The expansion of the scope of the study to other countries would contribute to its generalizability and provide comparative insights into ideology-based voting behavior under various political regimes.

References

- Imik Tanyildizi, N., & Tanyildizi, H. (2022). Estimation of voting behavior in election using support vector machine, extreme learning machine and deep learning. *Neural Computing and Applications*, 34(20), 17329-17342. <https://doi.org/10.1007/s00521-022-07395-y>
- Singh, L., Kamboj, S., Kaur, T., & Singh, P. (2024, March). Exploring changes in public opinion, political campaigns, and political behavior through data science and machine learning. In *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)* (pp. 1-5). IEEE. <https://doi.org/10.1109/aimla59606.2024.10531470>
- Dalton, R. J. (2016). The potential of big data for the cross-national study of political behavior. *International Journal of Sociology*, 46(1), 8-20. <https://doi.org/10.1080/00207659.2016.1130410>
- Kirkizh, N., Ulloa, R., Stier, S., & Pfeffer, J. (2024). Predicting political attitudes from web tracking data: a machine learning approach. *Journal of Information Technology & Politics*, 21(4), 564-577. <https://doi.org/10.1080/19331681.2024.2316679>
- Li, Y., & Mao, H. (2022). Study on machine learning applications in ideological and political education under the background of big data. *Scientific Programming*, 2022(1), 3317876. <https://doi.org/10.1155/2022/3317876>
- Bayyurt, N., & Çaha, H. (2020). Determination of women's voting behavior; A machine learning approach in the Turkish political arena. *International and Multidisciplinary Journal of Social Sciences*, 9(3), 260-288. <https://doi.org/10.17583/rimcis.2020.5027>
- Ma-Kellams, C., Bishop, B., Zhang, M. F., & Villagrana, B. (2018). Using "Big Data" versus alternative measures of aggregate data to predict the US 2016 presidential election. *Psychological Reports*, 121(4), 726-735. <https://doi.org/10.1177/0033294117736318>
- Saifeddine, S., & Abdellatif Chakor, M. (2024, April). Systematic literature reviews in political marketing: Behavior, influence, and trust in the era of big data and artificial intelligence. In *The international workshop on big data and business intelligence* (pp. 112-126). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-65018-5_11
- Jenkins, J. C., Slomczynski, K. M., & Dubrow, J. K. (2016). Political behavior and big data. *International Journal of Sociology*, 46(1), 1-7. <https://doi.org/10.1080/00207659.2016.1130409>
- Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22(1), 297-323. <https://doi.org/10.1146/annurev-polisci-090216-023229>

- Bonica, A. (2018). Inferring roll-call scores from campaign contributions using supervised machine learning. *American Journal of Political Science*, 62(4), 830-848. <https://doi.org/10.1111/ajps.12376>
- Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015). No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, 48(1), 71-74. <https://doi.org/10.1017/s1049096514001760>
- Radford, J., & Joseph, K. (2020). Theory in, theory out: the uses of social theory in machine learning for social science. *Frontiers in big Data*, 3, 18. <https://doi.org/10.3389/fdata.2020.00018>
- Guo, L., & Vargo, C. (2015). The power of message networks: A big-data analysis of the network agenda setting model and issue ownership. *Mass communication and society*, 18(5), 557-576. <https://doi.org/10.1080/15205436.2015.1045300>
- P Papakyriakopoulos, O., Hegelich, S., Shahrezaye, M., & Serrano, J. C. M. (2018). Social media and microtargeting: Political data processing and the consequences for Germany. *Big data & society*, 5(2), 205395171881184. <https://doi.org/10.1177/2053951718811844>
- Al-Garadi, M. A., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., ... & Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE access*, 7, 70701-70718. <https://doi.org/10.1109/access.2019.2918354>
- Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*, 659(1), 48-62. <https://doi.org/10.1177/0002716215570279>
- Clark, W. R., & Golder, M. (2015). Big data, causal inference, and formal theory: Contradictory trends in political science?: Introduction. *PS: Political Science & Politics*, 48(1), 65-70. <https://doi.org/10.1017/s1049096514001759>
- Hopp, T., & Vargo, C. J. (2017). Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data. *Computers in human behavior*, 68, 368-377. <https://doi.org/10.1016/j.chb.2016.11.034>
- Amoore, L. (2023). Machine learning political orders. *Review of International Studies*, 49(1), 20-36. <https://doi.org/10.1017/s0260210522000031>