Thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Politics

New Empirical Methodologies for Studying Modern Political Campaigning

Musashi Harukawa

Merton College

Long Vacation 2022

Word Count: 44,267

University of Oxford

Acknowledgements

There are countless people who have been part of this thesis.

Thank you Noe, for being at my side throughout. This thesis would not have happened without your love, support and wisdom. Thank you Tom, for all of your time, insights and friendship. Thank you Andy and Ray, for your encouragement, criticism, faith, support and wisdom. Thank you Sergi, for first teaching me political science and continuing to be a mentor. Thank you Lucy and my colleagues at UCL, for bringing me into your community and giving your valuable insights into my work. Thank you to the many friends and peers who supported me through these and trying times–Gonzalo, Mitchell, Nick, Agni, Klaudia, Ash, Pablo, Andrew and countless others. This thesis also would not have been possible without the generous sponsorship of the DPIR. And finally, I want to thank my family, for their love and sacrifices.

Contents

Abstract

The evolving nature of political campaigning brings new opportunities and challenges. Political scientists are faced with questions about the implications of a new, digitalized approach to campaigning via online platforms, and have access to a greater wealth of advertising data than ever before. Each of the three papers of this thesis presents new tools and techniques drawn from distinct disciplines, adapted to these challenges and empirical social science research in general.

The first paper employs a novel experimental design incorporating machine learning to simulate a targeted campaign, and finds not only that targeting can affect unaligned voters, but evidence of within-respondent heterogeneity that informs our understanding of prior null results in this literature. The second paper combines recent advances in deep learning, natural language processing with methods from the explainable machine learning subfield to produce a tool for identifying the differentiating characteristics of political campaigns. The major contributions of this paper are 1) identifying the limitations of measurement of current approaches to text-as-data, 2) adapting methods for interpreting and validating the output of complex deep learning models to a empirical political science context, and 3) beginning

the development of a novel hybrid computationally-assisted qualitative approach to studying text in political science. The third paper draws from computational linguistics and ethical artificial intelligence to highlight the risks and challenges of applying sophisticated models for measuring meaning in text, and applies a structured approach to representing text that enables researchers to transparently classify texts for observation selection. Together, these papers not only advance methods for studying political campaigns, but contribute to a wider interdisciplinary goal of discovering, measuring and inferring patterns in complex data.

New Empirical Methodologies for Studying Modern Political Campaigning

Introduction

In the three papers of the thesis, I present a series of data-focused methodological innovations for confronting the diverse challenges faced by social science researchers in working with unstructured data, with a substantive focus political campaigns. As advertising strategies become more complex and advertisements more numerous, researchers require robust tools for analyzing and characterizing this diversity. These tools are drawn from a variety of fields, from applied machine learning and experimental design to deep learning and natural language processing, and adapted to political science applications. In each paper, the aim is not only to develop a tool for studying a different aspect of political science campaigns, but also to contribute to a nascent and growing inductive approach to social science research that seeks to recognize and make sense of the complexity of empirical social phenomena.

In the first paper, titled "Does Microtargeting Work?", I ask whether individualized allocation strategies for political advertisements can be consequential. In lieu of access to the targeting strategies of political campaigns, I implement a novel experimental design that uses machine learning to model the complex interaction between individual traits and the effectiveness of five real advertisements run by the Trump campaign in the 2020 United States (US) presidential election. This model is used to target participants in the treatment group with the advertisements predicted to be most effective given their traits. By showing that targeting can have effects on undecided voters, the experiment indicates that despite small and null aggregated effects of political advertisements, systematic searches of within-individual differences in effectiveness can yield significant effects. This result calls us to reconsider the

primacy of identifying unbiased average effects as the dominant goal in quantitative political science research, and motivates the need for tools to inductively identify patterns of observed heterogeneity observed without introducing bias that limits our ability to extrapolate beyond our sample.

The second paper explores methods for augmenting the inference of individual researchers with recent advances in computational methods for modelling language. The volume of text data produced and available to political science researchers has continued to grow in the past decade, spurred in part by changes in cost and scalability from digitalization, but also increased demands for transparency. The opportunities these data present has galvanized the development of a methodological subdiscipline in political science focusing on tools for characterizing and interpreting substantively interesting patterns in large corpora. Motivated by the limitations of existing approaches in political science text-as-data, I introduce methods for model interpretation from the explainable machine learning subfield as a means of extracting the reasoning of complex deep learning text classifiers, and adapt these methods to the frameworks for validation and inference used in the empirical social sciences. Applying this to a collection of advertisements from twelve US presidential and primary campaigns, I show how the approach can be used to enhance qualitative approaches to text by annotating advertisements with the features that characterize and differentiate the different campaigns.

In the third paper I address a critical early step in any research process; the selection of observations. Given researchers aim to study the diverse strategies and messages employed in political campaigns, tools for identifying specific phenomena are essential for extracting relevant advertisements from the large collections that

are available. In contrast to the second paper where I leverage unstructured representations of language from deep learning and natural language processing, in this paper I use a theoretically-guided approach to language from linguistics as a basis for characterizing patterns of language use, and apply this to studying the use of partisan animus and threat in a dataset of campaigning emails collected during the 2020 US election cycle.

All three papers emphasize the importance of interpretability in political science research in different ways. Although the focus on accurate prediction in computer science and machine learning has facilitated the rapid development of tools able to usefully model complex empirical phenomena, further work is needed to adapt these innovations to the questions and requirements of political science researchers. For the first paper, although I show that there is sufficient heterogeneity within individuals to make targeting effective for some groups, the clear next step in this research is to better describe this heterogeneity. This is essential to understanding when and why targeting strategies can be consequential for elections. The second paper speaks directly to the question of interpretability, and I seek to harness the logic of complex language classifiers to enhance human understanding of what differentiates campaigns. The third paper takes a more critical approach to what we can learn from these models, based on findings that models encode the contextual biases of the data they are trained on. In applications where we seek to make unbiased characterizations of political phenomena, uncritical applications of these models leads not only to erroneous inference, but also risks propagating societal biases in potentially damaging ways. There is therefore a trade-off between the power and flexibility of complex but opaque models with the predictability of

more straightforward approaches. I pursue a middle way that leverages complex models for identifying patterns that exist in empirical data, but engages with and challenges these inferences of these models with a range of tools for validation and interpretation.

Literature Review

Political campaigns have long been an object of study for political scientists. From explorations of the content of advertisements and increasingly negative messaging (Ansolabehere and Iyengar 1995; Jerit 2004), scholars have studied the implications of the moments in which parties attempt to communicate the masses not only in persuading voters at the polls (Gerber et al. 2011; Coppock, Hill, and Vavreck 2020; Broockman and Kalla 2020), but also in influencing attitudes and social divisions outside the electoral cycle (Iyengar and Krupenkin 2018).

The advertising strategies employed by political campaigns have become increasingly diverse, with the past decade seeing a particular rise in data-driven and digitalized campaigning (Wood and Ravel 2017; Fowler et al. 2021). I focus on two aspects of this change that have created new questions and challenges for researchers.

The first relates to changes in the way that advertisements are delivered, tailored and targeted to voters. Revelations surrounding the role of *Cambridge Analytica* and related firms in several watershed political events of 2016 propelled a new form of high-tech campaigning into the spotlight (Simon 2019). A normative scholarly literature highlights the dangers of a new form of campaigning that uses extensive

data on voters, collected via data brokers and social media platforms, to produce and target advertisements that are individually persuasive. The harms discussed in this literature range from threats to individual privacy (Wachter 2017) and civic discourse through the creation of informational "filter bubbles" (Burkell and Regan 2020), to undermining the very foundations of democracy by weakening individual autonomy, with elites effectively determining political outcomes (Sunstein 2015; Krotoszynski 2020).

As I discuss, some of these arguments presuppose that the claims by the firms of being able to influence electoral outcomes are true. This stands in contrast to a sizable body of political science research that finds that the persuasive effect of political advertisements is typically small (Broockman and Kalla 2020) and short-lived (Gerber et al. 2011). In particular, a large study by Coppock, Hill, and Vavreck (2020) finds that the effect of political advertisements does not appear to vary greatly from advertisement to advertisement. If this is the case, then it is unlikely that sophisticated tailoring and targeting strategies will make any difference.

However, given the opaque nature of the advertising platforms, access to information on the extent, content and manner of targeting is limited (Ali et al. 2019; Edelson et al. 2019). This complicates efforts to determine its effect observationally. Experimental approaches have provided a means for researchers to nevertheless make inferences about the effectiveness of targeting. Zarouali et al. (2020) simulate the tailoring of advertisements to latent psychological types, a technique called *psychometric profiling*, and find that advertisements matched on this basis can be more effective. Meanwhile, in an experiment run during the 2021 German elections, Gahn (2022) tests the effect of targeting on variable numbers of traits, and finds that over-

personalization of messages can induce a backlash that reduces their persuasiveness. The first paper likewise contributes to this line of research with a novel experimental design. Taking an agnostic approach to the exact strategy of targeting, I employ machine learning to target survey participants with individually optimal advertisements. The purpose of this study is to determine whether there is enough variation between the effects of different advertisements on the same person to make matching the "right" advertisement with a viewer worthwhile for campaigns.

The second challenge faced by researchers of political campaigns is the scale of data. Digitalization has provided campaigns not only with a greater number of narrow audiences, but also means of automating the testing of advertisements and messages. Moreover, increased demands for transparency from the public, regulators and academics has led the Google and Meta to increase access to the content of advertising that takes place on their platforms. Additionally, journalistic and academic projects have produced several large corpora of political advertisements. The following table, taken from the second paper of this thesis, illustrates the scale of data that is now available to researchers studying political campaigns.

Large Political Advertising Corpora.

Dataset	N Docs
Google Transparency (Politics, US)	619K
Facebook Ad Library (Politics, US)	13.3M
Princeton Corpus of Emails	435K
ProPublica	225K

The challenges of analyzing large and complex text data is not unique to the study of campaigns. Many areas of political science, from legislatures to courts to media,

require tools for studying data that is largely *textual*. In part driven by this demand, a methodological subfield for computationally-assisted text analysis referred to as text-as-data (Grimmer 2013) has become an increasingly popular and important area of political and other social sciences. The contributions of this field are not solely tools for analyzing text data; key authors have outlined a new direction for social science research that embraces an inductive, empirically-guided paradigm that incorporates tools from machine learning and related disciplines for the tasks of discovery, measurement and inference (Grimmer, Roberts, and Stewart 2021; Grimmer, Marble, and Tanigawa-Lau 2022).

Advances in deep learning, a subfield of machine learning and artificial intelligence, have yielded powerful models for language that excel at a broad variety of applications with minimal task-specific adaptation and training (Vaswani et al. 2017; Devlin et al. 2019; Rogers, Kovaleva, and Rumshisky 2020). These developments present many exciting applications for the study of political campaigns; despite the increasing availability of politically relevant corpora, data labelled with substantively useful features is both rare and expensive to produce. Transfer learning–training a model on one data and then applying it to another–offers the possibility to leverage powerful inferential models against comparatively small and unlabelled datasets.

However, two major challenges limit the applicability of these developments to political science research. The first is a lack of interpretability. Whereas these models excel at optimizing a quantifiable metric, such as predictive accuracy, their complex nature and lack of interpretable parameters limits their application for explanation, a key task in political science research (Hofman, Sharma, and Watts 2017). This requirement for explainable predictions is not unique to the social sciences, and in

the second paper of this thesis I introduce techniques from a subfield of applied machine learning for extracting the "reasoning" of complex models. By adapting and contextualizing these explanation tools for political science research, I open the door to researchers harnessing the advances in deep learning to the study of large datasets of political advertisements.

There is another challenge for the application of these models in political science research. Complex language models have been shown to encode the social biases (Caliskan, Bryson, and Narayanan 2017). This means that the societal biases present in the texts used to train these models, such as the associations between gender and profession, is exhibited in their behavior and predictions. I discuss in my third paper the pitfalls of using a computational-statistical approach to measuring meaning (Garg et al. 2018; Rodman 2019; Rodriguez, Spirling, and Stewart 2020) or identifying groups of similar entities (Ash, Gauthier, and Widmer 2021), in social research applications.

Nevertheless, tools for constructing representations of text data that are amenable to quantitative analyses continue to be essential to the task of studying large political campaigning corpora. Computational linguistics has much to offer in this regards; in my third paper I also extend the work of Ash, Gauthier, and Widmer (2021) on applications of semantic roles, a linguistic framework for abstracting the meaning of a sentences from its specific phrasing (Jurafsky and Martin forthcoming). I argue that these frameworks and representations offer opportunities for identifying and studying the range of diverse phenomena and questions in political campaigns.

More broadly, the way to confront the changes and further develop our understand-

ing of modern political campaigns requires methodological diversity and interdisciplinarity. The innovations of this thesis are not method, but application and adaptation to a social context. This transfer need not be one way either; the frameworks and standards of proof for empirical social science research are robust and mature, shaped by the need to make valid inferences on consequential questions in frequently observational settings. The next steps for the discipline and the field relate to those stated in Grimmer, Roberts, and Stewart (2021); the development of a novel framework and discipline combining computational methods and rigorous theory to generate new insights into social phenomena.

References Cited in Introduction

Ali, Muhammad, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. "Ad delivery algorithms: The hidden arbiters of political messaging." *arXiv Preprint arXiv:1912.04255*.

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going negative : how attack ads shrink and polarize the electorate*. New York ; London: Free Press.

Ash, Elliott, Germain Gauthier, and Philine Widmer. 2021. "Text semantics capture political and economic narratives." *arXiv Preprint arXiv:2108.01720*.

Broockman, David, and Joshua Kalla. 2020. "When and Why Are Campaigns' Persuasive Effects Small? Evidence from the 2020 US Presidential Election." OSF Preprints. https://doi.org/10.31219/osf.io/m7326.

Burkell, Jacquelyn, and Priscilla M. Regan. 2020. "Voting public: Leveraging personal information to construct voter preference." In *Big Data, Political Campaigning and the Law: Democracy and Privacty in the Age of Micro-Targeting*, edited by Normann Witzleb, Moira Paterson, and Janice Richardson, 47–68. Routledge.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356 (6334): 183–86.

Coppock, Alexander, Seth J Hill, and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6 (36).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Under-

standing." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Edelson, Laura, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. "An Analysis of United States Online Political Advertising Transparency." *arXiv Preprint arXiv:1902.04385*.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. "Political advertising online and offline." *American Political Science Review* 115 (1): 130–49.

Gahn, Christina. 2022. "How much targeting is 'too much'? Voters' punishment of highly targeted campaign messages." Presented at MPSA.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115 (16): E3635–44.

Gerber, Alan S, James G Gimpel, Donald P Green, and Daron R Shaw. 2011. "How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment." *American Political Science Review* 105 (1): 135–50.

Grimmer, Justin. 2013. "Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57 (3): 624–42. https://doi.org/10.1111/ajps.12000.

Grimmer, Justin, William Marble, and Cole Tanigawa-Lau. 2022. "Measuring the Contribution of Voting Blocs to Election Outcomes." SocArXiv. https://doi.org/10.31235/osf.io/c9fkg.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24 (1). https://doi.org/10.1146/annurev-polisci-053119-015921.

Hofman, Jake M, Amit Sharma, and Duncan J Watts. 2017. "Prediction and explanation in social systems." *Science* 355 (6324): 486–88.

Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39 (S1): 201–18. https://doi.org/10.1111/pops.12487.

Jerit, Jennifer. 2004. "Survival of the fittest: Rhetoric during the course of an election campaign." *Political Psychology* 25 (4): 563–75.

Jurafsky, Dan, and James H. Martin. forthcoming. "Speech and Language Processing." https://web.stanford.edu/~jurafsky/slp3.

King, Gary, Patrick Lam, and Margaret E Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–88.

Krotoszynski, Ronald J., Jr. 2020. "Big Data and the electoral process in the United States: Constitutional constraint and limited data privacy regulations." In *Big Data, Political Campaigning and the Law: Democracy and Privacty in the Age of Micro-Targeting*, 186–213. Routledge.

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis*, 1–25.

Rodriguez, Pedro L, Arthur Spirling, and Brandon M Stewart. 2020. "Embedding Regression: Models for Context-Specific Description and Inference in Political Science."

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A primer in bertology:

What we know about how bert works." *Transactions of the Association for Computational Linguistics* 8: 842–66.

Simon, Felix M. 2019. "'We power democracy': Exploring the promises of the political data analytics industry." *The Information Society* 35 (3): 158–69.

Sunstein, Cass Robert. 2015. "Fifty shades of manipulation." *Journal of Behavioral Marketing* 213 (1).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in Neural Information Processing Systems* 30.

Wachter, Sandra. 2017. "Privacy: Primus Inter Pares Privacy as a Precondition for Self-Development, Personal Fulfilment and the Free Enjoyment of Fundamental Human Rights." *SSRN*, January.

Wood, Abby K, and Ann M Ravel. 2017. "Fool Me Once: Regulating Fake News and Other Online Advertising." *S. Cal. L. Rev.* 91: 1223.

Zarouali, Brahim, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. "Using a personality-profiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media." *Communication Research*.

Does Microtargeting Work? Evidence from an Experiment during the 2020 United States Presidential Election

Political science research consistently shows that political advertisements have small and uniform effects. This contrasts with claims that microtargeted campaigning can sway electoral outcomes and poses a threat to society. This paper introduces a novel experimental design simulating a targeted campaign to empirically test whether targeting political advertisements can be effective. Participants in an online survey experiment view one of five anti-Biden advertisements. Respondents assigned to control are allocated advertisements at random. This data is used to train a model predicting Biden favorability from advertisement and pre-treatment traits. Respondents in the treatment group view the advertisement that this model predicts to be the most effective. The difference between targeted and random allocation is an 8.7 percentage point increase in Biden dislike and a 7.1 percentage point decrease in intent to vote Biden among unaligned voters who had not yet cast their vote ($N = 586$). The effect is negligible among partisan voters.

Introduction

Following the Brexit referendum and the election of Donald Trump in 2016, observers raised questions surrounding the role of sophisticated digital campaign strategies employed by *Cambridge Analytica* and others (Simon 2019). Watchdogs, journalists, and scholars identified numerous issues stemming from these technologies: threats to privacy from the systematic harvesting of individual data by corporations (Zuboff 2015), threats to public accountability from personalization of political information (Fowler et al. 2021), and threats to autonomy from mass manipulation (Burkell and Regan 2019).

Although the normative discussion frequently proceeds with the assumption that digital campaigning strategies are effective (Krotoszynski 2020), empirical political science research indicates this conclusion is at best premature (Nickerson and Rogers 2020). A large panel study by Coppock, Hill, and Vavreck (2020) on US presidential election campaign advertisements finds not only small average persuasive effects, but a lack of heterogeneity that seems to leave little room for targeting to operate. In contrast, research in psychology finds that campaigns leveraging knowledge of receiver characteristics can improve message effectiveness (Zarouali et al. 2020). Efforts to reconcile these results are hampered by the challenges of studying campaigns and their activities online.[1]

I introduce a novel experimental design to directly estimate the effect of targeting political advertisements. The experiment runs in two immediately sequential

[1]See, for example, difficulties faced by Edelson et al. (2019, 3).

stages. Respondents in the first stage are randomly assigned one of five real anti-Biden advertisements. I use this data to train an online targeting algorithm, which optimally allocates advertisements on the basis of respondent characteristics in the second stage. Provided the participants in the two stages are comparable, this design yields a consistent estimate of the effect of microtargeting on candidate preference and voting intention.

The results provide evidence that matching respondents with rptimal advertisements increases the effect of the advertisement in some subgroups. The treatment had no significant average effect on partisan respondents. Among unaligned respondents who had not already voted at the time of the survey, optimized allocation caused an 8.7 percentage point (pp) increase in Biden dislike and 7.1pp decrease in intent to vote for Biden relative to receiving a random advertisement.

The findings of this paper contribute to several strands of literature. In political science, these results suggest that prior studies finding no effect of political advertisements may have insufficiently considered heterogeneity on respondent characteristics. For the legal and ethical debate regarding this technology, the magnitude of the estimates found in this experiment implies that microtargeting may be sufficient to change the result of close elections with a sufficient number of unaligned voters.

Motivation

I begin by clarifying the distinction between *tailoring* and *targeting*. *Tailoring* a message is constructing it to appeal to a specific audience. *Targeting* is delivering a

message only to the intended audience (Fowler et al. 2021). I use "microtargeting" to mean targeting on the basis of individual-level data such as gender or ethnicity, as opposed to aggregate data such as postal code. This article focuses on the effect of the advertisement allocation strategies of data-driven political campaigns, or *political microtargeting*.

A dramatic rise in online political campaigning brought with it a myriad of problems (Wood and Ravel 2017). Of particular concern is the ability of political campaigns, operating on platforms such as Facebook, to show narrowly tailored advertisements to the individuals they are designed to affect (Fowler et al. 2021). Legal and ethical scholars note systemic threats to democracy arising from this technology and the difficulty of limiting these harms (Burkell and Regan 2019; Hindman 2018). Although these arguments largely assume that microtargeting works, many political scientists remain skeptical of the claims of campaigners and their ability to alter electoral outcomes (Nickerson and Rogers 2020).

A decade of field experiments studies the effect of political advertising in various mediums (Gerber et al. 2011; Edelson et al. 2019), discovering largely small or null effects (Kalla and Broockman 2018). In a recent field experiment Coppock, Hill, and Vavreck (2020) conclude: "expensive efforts to target or tailor advertisements to specific audiences require careful consideration. The evidence… shows that the effectiveness of advertisements does not vary greatly from person to person or from advertisement to advertisement" (6). Correspondingly, targeting the delivery of advertisements should be unlikely to yield any gains in effectiveness. It is difficult to square these null findings with claims that microtargeting is a threat to democracy.

Research on *psychometric profiling*, the building of psychological voter "profiles" for tailoring advertisements, provides evidence that gains should be possible. Zarouali et al. (2020) use a personality profiling algorithm to label participants as introverts or extroverts, and then show a personality-congruent advertisement based on this label, finding that correctly matched advertisements have a stronger effect. The psychological models of persuasion underlying this technique (e.g. Petty and Cacioppo 1986; Cialdini 2007) emphasize that whether a message is processed *emotively* or *cognitively*[2] depends on an interaction between message content, receiver characteristics and receiver social context. Knowledge of receiver characteristics allows a campaign to tailor an advertisement to increase the likelihood that the message will be processed in a manner that results in opinion or attitude change.

The psychometric approach maximizes the persuasive effect of advertisements on the basis of a deductive theoretical model associating psychological types and advertisement response. In contrast, in my approach I train a model to directly optimize the outcome (candidate preference), and thus avoid relying on the validity of a particular theoretical model to produce optimal respondent-advertisement matchings. By demonstrating whether a campaign can improve their performance by optimizing advertisement allocation on receiver traits, my theoretically agnostic approach shows whether there are possible gains from microtargeting regardless of the specific technique employed.

[2] *Elaboration Likelihood Model* (Petty and Cacioppo 1986): modes of processing information range from central (critical, rational) to peripheral (heuristic, emotive).

Design

Experiment

In order to estimate the effect of microtargeting, I compare the average outcomes between an untargeted control group and targeted treatment group. I operationalize targeting to mean that targeted respondents are shown the optimal advertisement conditional on their traits, while untargeted respondents are shown an advertisement independently of their traits (i.e. at random).

Respondents in both groups are shown one of five real anti-Biden advertisements (see table 2.1), with the allocation rule differing between treatment and control. These advertisements are chosen from the set of all Trump advertisements run in the final five months of the election, on the criteria that they are clearly tailored to different audiences, and no advertisement is likely to outperform all others.[3]

Table 1.1: Advertisements and descriptions.

Title	Description
They Mock Us	Clinton and Biden are mocking you (In-Group)
Why did Biden let him do it?	Hunter Biden's ostensible corruption
Biden will come for your guns	Second Amendment; Biden will steal guns
Insult	Biden: Black Trump supporters not Black
Real Leadership	Obama/Biden caused wars, neglected veterans

[3]See Appendix 1A: Case and Advertisement Selection for details.

Respondents assigned to control are randomly allocated an advertisement using permuted block randomization with uniform probability. This data is used to train a model that predicts the outcome, Biden favorability, as a function of the respondent's characteristics[4] and the advertisement. For respondents in the treatment group, this model generates predictions for each of the five advertisements and then allocates the one that minimizes Biden favorability.[5]

Algorithm

Five models are fitted to predict Biden favorability as a function of advertisement allocation and pre-treatment traits. The first three are decision tree-based algorithms chosen for their ability to learn highly conditional response surfaces: Random Forest (RF), AdaBoost and GradBoost. A Multi-Layer Perceptron and Support Vector Machine are included for comparison, with the latter hedging on the possibility of a smooth response surface. All models are standard, "off-the-shelf" algorithms implemented in the popular `scikit-learn` library (Pedregosa et al. 2011). This highlights that campaigns with low data science sophistication can easily implement similar approaches.

The models are compared on root mean squared error, maximum error and prediction time,[6] in order to find the model best able to give quick and accurate predictions.

[4]Characteristics measured are age, gender, race, income group, state, news interest, whether they think the country is on the right track, partisanship, and ideology.

[5]In this case, I minimize the outcome because the advertisements are negative.

[6]Prediction time mattered because long queues during traffic bursts could create uncontrolled variation in the respondent's experience of the survey.

RF and AdaBoost performed the best on all of these metrics. Between RF and AdaBoost, RF performed weakly better and was less likely to predict ties, and was thus chosen for predicting optimal advertisements.[7]

Sample Balance

The first sample is necessarily assigned to the control group because randomized allocation of advertisements is necessary to train a model that produces unconfounded predictions of the outcome for targeting respondents in the treatment group. Ideally, respondents in the second sample would be randomly split between treatment and control (and the first sample used only for training the algorithm). Due to resource constraints, I instead used the first sample as the control group and assigned all respondents in the second sample to treatment. Thus treatment is assigned based on when the subject took the survey, raising the possibility that the two groups differ systematically in politically relevant ways.

The online format of the experiment allowed for the rapid deployment and simultaneous testing of many respondents, and the switch-over between stages where the targeting model was trained and uploaded was completed in under an hour. Given this and the results of a power analysis and simulation of the experiment based on the Coppock, Hill, and Vavreck (2020) data, I opted to assign the groups sequentially in order to have enough observations to train the targeting algorithm (300 per advertisement) and detect potentially small effects of targeting (900 in treatment).[8]

[7] In the case of ties, an optimum was randomly chosen.
[8] See Figure 1.7 in Appendix 1B.

I conduct multiple checks for covariate imbalance; a chi-squared test, covariate balance check, and logistic regression predicting whether a respondent was collected in the first or second sample. None of these tests provide evidence of systematic difference between the samples. To search for evidence of an unobserved time-variant confounder, I regress the time at which a respondent took the survey interacted on the group they were in and their time zone. This check shows no evidence of a linear time trend in the outcome within any group or time zone, which indicates the data are inconsistent with there being a time-variant unobserved confounder on the outcome.[9] I therefore conclude that it is extremely unlikely that the sequentiality of samples significantly affected my estimate.

Hypotheses

I measure the effect of targeting on three outcomes that a campaign is likely to want to influence. For a given respondent, I hypothesize that targeting anti-Biden advertisements will *increase* Biden dislike, *decrease* intent to vote for Biden, and *decrease* turnout intention, relative to showing an advertisement at random. Note the latter two outcomes are only measurable for individuals who have not already voted at the time of the survey. Therefore, all hypotheses are tested for the subset of respondents who had not already voted at the time of the survey, but only Biden dislike is tested with the full sample.

In line with theories of motivated reasoning, whereby partisan voters resist information contrary to their identity orientations, and findings in related studies that the

[9]See SI-A: Randomization Check and Time-of-Day Effects

effects of advertisements depend on the partisan affiliation of the recipient (Coppock, Hill, and Vavreck 2020; Broockman and Kalla 2020), I test all hypotheses conditioning on partisanship.

Results

Effect of Targeting

The experiment took place on 28 October 2020. $1,500$ and 900 responses were collected for the control and treatment groups. After filtering for irregularities,[10] the experiment yielded $2,261$ valid responses, with $1,416$ in control and 845 in treatment.

Figure 1.1 shows the estimated effect of targeting on Biden dislike, intent to vote for Biden, and turnout intention among voters who had not already voted at the time of the survey ($N = 1,160$), conditional on partisan self-identification. Each pair of bars shows the predicted level of the outcome untargeted and targeted respondents, with 95 percent confidence intervals, for Democrat, Unaligned and Republican respondents in the corresponding facet.

The numbers along the top show the estimated conditional average treatment effect (CATE) of targeting for each outcome and partisanship group. The central facet shows that the estimated CATE of targeting among unaligned respondents who had not voted at the time of the survey on Biden dislike and intent to vote for Biden are

[10] Irregularities include incomplete surveys and failed attention checks. See SI-A.

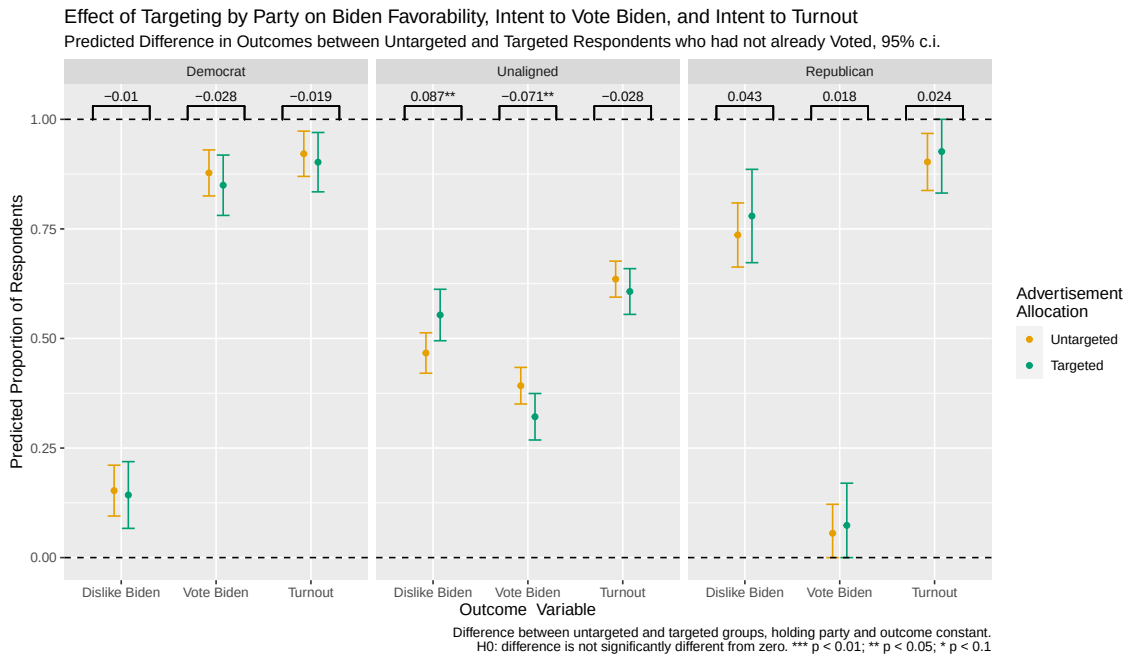

Figure 1.1: Effect of targeting on biden dislike, Biden vote, and turnout; predicted levels and differences.

0.087 and −0.071 respectively. This translates to targeting resulting in an 8.7pp increase in the proportion of unaligned respondents who state that they dislike Biden, and a 7.1pp decrease in the proportion of unaligned respondents stating intent to vote for Biden, comparing to the untargeted scenario. The remaining differences are smaller and insignificant.

When I include respondents who had voted at the time of the survey, the estimated CATE of targeting on Biden dislike among unaligned voters is smaller ($\beta = -0.054$, $s.e. = 0.027$), but this result does not remain significant at $\alpha = 0.05$ if the outcome is operationalized in any other way. The estimated CATE of targeting is null for Democrat and Republicans, and the estimated ATE of targeting is null for all outcomes.

All results are checked for robustness to alternative operationalizations of the outcome as a five-point ordinal or continuous scale. Among unaligned voters who had not pre-voted, the effect of targeting on Biden favorability is robust to all checks, but the coefficient on treatment for intent to vote Biden does not remain significant at $\alpha = 0.05$ when operationalized as a logistic regression ($p = 0.08$). Robustness checks, regression tables and coefficient plots are in Appendix 1C.

Decomposing the Effect

Microtargeting works by leveraging the fact that for a given individual, certain messages are more persuasive than others. By learning an individual's probable reaction to different advertisements, it becomes possible to match them with the most effective advertisement. This is the strategy of the design of this experiment.

Note that because each individual is matched with their optimal advertisement, this design allows the proportion of each advertisement shown to vary between stages. If one advertisement is the most effective for a greater proportion of people, then it will be shown more. This means we can decompose our estimator of the average effect of targeting into two parts.

The first part is the change due to improving individual allocations. At the individual level, the treatment "moves" the outcome from the average point on the advertisement response surface to the optimum. I refer to this as the improvement due to "better allocations". The second part is the change due to altering the proportions of the advertisements. If one advertisement is more effective and it is shown to

a greater proportion of respondents, then the average outcome will correspondingly increase. I refer to this as "better advertisements".[11]

This decomposition matters because our notion of microtargeting–showing the right advertisement to the right person–is closer to the effect of "better allocations". A campaign could achieve the effect of "better advertisements" without large amounts of individual data by using focus groups to find their best advertisement and showing it to everyone.

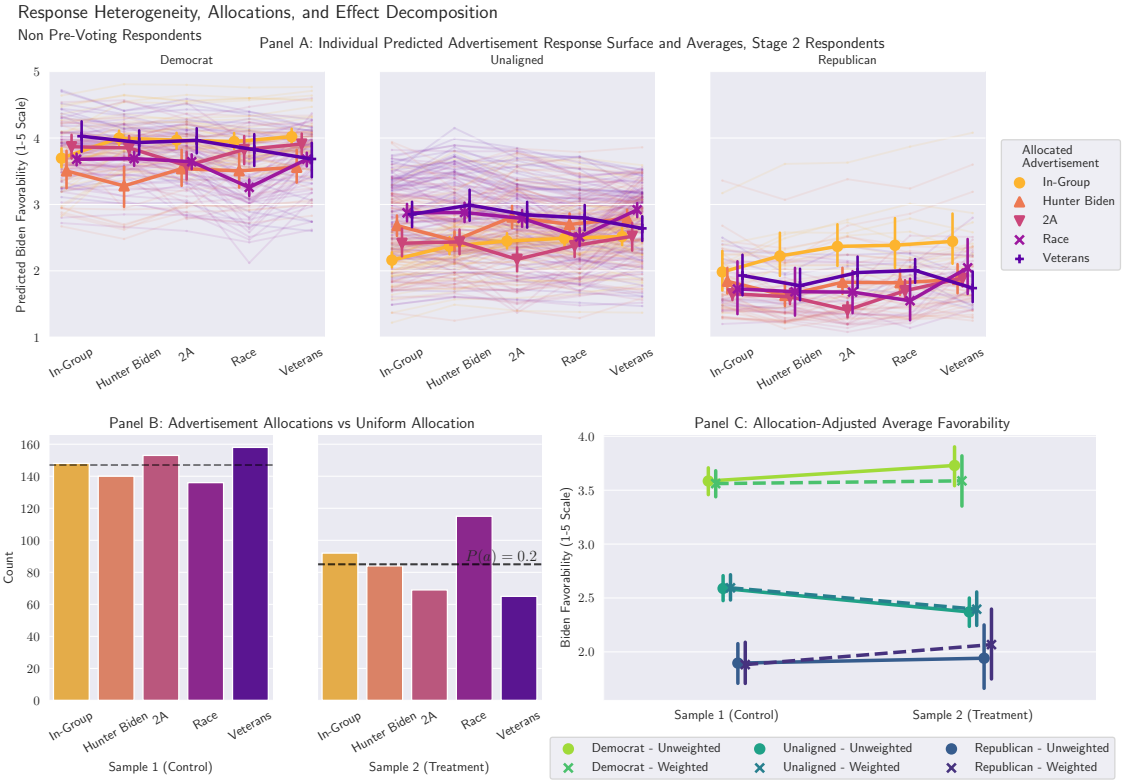

Figure 1.2: Response heterogeneity, advertisement allocations and effect decomposition.

Figure 1.2 provides insight into this decomposition and its relation to within-

[11] As shorthand for "more of the better advertisements".

respondent heterogeneity. For each respondent in the treatment group, the targeting model calculates their predicted Biden favorability given each of the five advertisements. Each of these individual predicted response surfaces is depicted in Panel A, Figure 1.2 by a thin line joining five points. The shape of this line determines which advertisement was shown to the respondent; the lowest point corresponds with the advertisement assigned. The lines are colored and grouped according to this assignment. The dot-and-whiskers connected by the thicker lines denote the group averages and 95% confidence intervals.[12]

The lines are not flat; this shows that each individual is predicted to respond differently to each of the five advertisements. We also see that no single advertisement is the most effective, nor is the relative effectiveness of each advertisement the same for all people. The dot-and-whiskers emphasize both of these facts, showing that the difference between the optimal advertisement and other advertisements is not marginal, nor are the individuals receiving the same advertisement homogeneous.

If the effect of targeting were entirely due to "better advertisements", then the model would have assigned everyone the same (best) advertisement. Panel B, Figure 1.2 shows this is clearly not the case; although the proportion of each advertisement shown varies between samples, it remains remarkably uniform in the targeted group. Therefore at least some of the effect is due to "better allocations".

To visualize the change in outcomes between samples holding the change due to "better advertisements" constant, we weight observations to make the advertisement assignment probability uniform in each stage. The circular markers connected by

[12]A descriptive characterization of these groups is explored in Appendix 1B: Characterization.

solid lines in Panel C, Figure 1.2 show the unweighted average Biden favorability conditional on partisan self identification and sample. The x-shaped markers connected by dotted lines show these averages after weighting each observation by the inverse probability of seeing the advertisement they were shown given the sample in which they were drawn.

There is little difference between the unweighted and weighted conditional means on the left because, as seen in panel B, the advertisement allocation in the first sample is nearly uniform. Notably, the conditional mean for unaligned respondents in the second sample hardly moves, indicating that there was little change in the proportions of each advertisement shown for unaligned respondents between stages. I find that 98 percent of the effect of targeting on unaligned voters who had not already voted is attributable to "better allocations" with the decomposition in SI-B. Thus the effect of microtargeting in this experiment is mostly attributable to optimized matching of dissimilar advertisements to heterogeneous respondents.

Discussion

In sum, this experiment shows that for these five advertisements, optimizing allocation on the basis of respondent traits causes a 7.1pp decrease among unaligned voters reporting that they intend to vote for Biden, and an 8.7pp increase in those reporting Biden dislike.

These magnitudes may have serious implications. Given that this experiment used

a convenience sample[13] and does not account for decay (Gerber et al. 2011) a precise estimate of how these effects translate to electoral outcomes is difficult to give. Under moderately conservative assumptions, a 7.1pp decrease in unaligned voters voting for Biden is sufficient to flip the result of the 2020 presidential election in Arizona, where 35.1 percent of voters are registered as unaligned and Biden won by a 0.3 percent margin.[14]

Do these findings contradict prior research? I argue that this may not be the case. Coppock, Hill, and Vavreck (2020) demonstrate in their experiments that there is little variation in the effectiveness of advertisements, but mostly check for heterogeneous treatment effects for different advertisement types. Among respondent characteristics, the authors only test for heterogeneity on partisanship. In contrast, the design employed in this paper leverages response heterogeneity across nine respondent pre-treatment covariates, and searches the space of all possible covariate combinations for the optimal outcome. These results are likewise consistent with Broockman and Kalla (2020), who find that while advertisements about Trump demonstrate "identical, small effects", "treatments for Biden show large, clear dispersion".

Future research will look at the extent to which the results in this paper are specific to this setting and these advertisements. Policy researchers can utilize this experimental design to test the effectiveness of interventions and regulation designed to limit the harms of microtargeting, such as clearer disclaimers or civic awareness campaigns. So long as this technology continues to be a part of our elections, un-

[13]Sample representativeness is addressed in SI-A.

[14]The assumptions behind this back-of-the-envelope calculation are explained in SI-B.

derstanding its effects is crucial to safeguarding electoral integrity.

References

Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going Negative : How Attack Ads Shrink and Polarize the Electorate*. New York ; London: Free Press.

Ansolabehere, Stephen, Brian Schaffner, and Samantha Luks. 2020. "CCES Common Content, 2019." Harvard Dataverse. https://doi.org/10.7910/DVN/WOT7O8.

Baldwin-Philippi, Jessica. 2017. "The Myths of Data-Driven Campaigning." *Political Communication* 34 (4): 627–33.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Broockman, David, and Joshua Kalla. 2020. "When and Why Are Campaigns' Persuasive Effects Small? Evidence from the 2020 US Presidential Election." OSF Preprints. https://doi.org/10.31219/osf.io/m7326.

Burkell, Jacquelyn, and Priscilla M. Regan. 2019. "Voter Preferences, Voter Manipulation, Voter Analytics: Policy Options for Less Surveillance and More Autonomy." *Internet Policy Review* 8 (4).

Cialdini, Robert B. 2007. *Influence: The Psychology of Persuasion*. Vol. 55. Collins New York.

Coppock, Alexander, Seth J Hill, and Lynn Vavreck. 2020. "The Small Effects of Political Advertising Are Small Regardless of Context, Message, Sender, or Receiver: Evidence from 59 Real-Time Randomized Experiments." *Science Advances* 6 (36).

Dobber, Tom, Ronan Ó Fathaigh, and Frederik Zuiderveen Borgesius. 2019. "The

Regulation of Online Political Micro-Targeting in Europe." *Internet Policy Review* 8 (4).

Edelson, Laura, Shikhar Sakhuja, Ratan Dey, and Damon McCoy. 2019. "An Analysis of United States Online Political Advertising Transparency." *arXiv Preprint arXiv:1902.04385*.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. "Political Advertising Online and Offline." *American Political Science Review* 115 (1): 130–49.

Gerber, Alan S, James G Gimpel, Donald P Green, and Daron R Shaw. 2011. "How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *American Political Science Review* 105 (1): 135–50.

Goyal, Vineet, and Rajan Udwani. 2020. "Online Matching with Stochastic Rewards: Optimal Competitive Ratio via Path Based Formulation." In *Proceedings of the 21st ACM Conference on Economics and Computation*, 791. EC '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3391403.3399531.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis*, 25–46.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Hindman, Matthew. 2018. *The Internet Trap*. Princeton University Press. https:

//doi.org/10.1515/9780691184074.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, 65–70.

Hunter, J. D. 2007. "Matplotlib: A 2d Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Kalla, Joshua L, and David E Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112 (1): 148–66.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. "Jupyter Notebooks - a Publishing Format for Reproducible Computational Workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, edited by Fernando Loizides and Birgit Scmidt, 87–90. Netherlands: IOS Press. https://eprints.soton.ac.uk/403913/.

Krotoszynski, Ronald J., Jr. 2020. "Big Data and the Electoral Process in the United States: Constitutional Constraint and Limited Data Privacy Regulations." In *Big Data, Political Campaigning and the Law: Democracy and Privacty in the Age of Micro-Targeting*, 186–213. Routledge.

Kruikemeier, Sanne, Minem Sezgin, and Sophie C Boerman. 2016. "Political Microtargeting: Relationship Between Personalized Advertising on Facebook and Voters' Responses." *Cyberpsychology, Behavior, and Social Networking* 19 (6): 367–72.

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–44. https://doi.org/10.1111/ajps.12361.

Nickerson, David W., and Todd Rogers. 2020. "Campaigns Influence Election Outcomes Less Than You Think." *Science* 369 (6508): 1181–82. https://doi.org/10.1126/science.abb2437.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Petty, Richard E, and John T Cacioppo. 1986. "The Elaboration Likelihood Model of Persuasion." In *Communication and Persuasion*, 1–24. Springer.

Simon, Felix M. 2019. "'We Power Democracy': Exploring the Promises of the Political Data Analytics Industry." *The Information Society* 35 (3): 158–69.

Sunstein, Cass Robert. 2015. "Fifty Shades of Manipulation." *Journal of Behavioral Marketing* 213 (1).

team, The pandas development. 2020. *Pandas-Dev/Pandas: Pandas* (version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134.

Wachter, Sandra. 2017. "Privacy: Primus Inter Pares Privacy as a Precondition for Self-Development, Personal Fulfilment and the Free Enjoyment of Fundamental Human Rights." *SSRN*, January.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wood, Abby K, and Ann M Ravel. 2017. "Fool Me Once: Regulating Fake News and Other Online Advertising." *S. Cal. L. Rev.* 91: 1223.

Zarouali, Brahim, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. "Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media." *Communication Research*.

Zuboff, Shoshana. 2015. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30 (1): 75–89.

Appendix 1A: Design and Implementation

Design Summary

Each of the respondents $i \in N$ was allocated one of five advertisements $D_{i,a}$, $a \in [1, 5]$. The five advertisements, detailed in Table 2.1 of the main article, were selected from the set of all anti-Biden advertisements run by the Donald Trump campaign on Facebook and YouTube. The criteria for selection are: all advertisements are clearly tailored to different audiences, but no single advertisement is likely to outperform all others. Advertisements focusing on Joseph Biden were chosen in favor of Donald Trump because two weeks prior to the election it was uncertain would Trump would drop out due to COVID-19. Although Coppock, Hill, and Vavreck (2020) find no evidence of asymmetric effects for attack versus promotional advertisements, attack advertisements were chosen because there is an extensive literature focusing on the normative issues surrounding negative advertising (Ansolabehere and Iyengar 1995).

In the first stage, respondents first answered demographic and political questions regarding age, gender, race, income group, state, interest in news, whether they thought the country was on the right track for the past four years, a seven-point partisan identification scale, and a five-point liberal-conservative ideological self-identification scale. The format and wording of these items was adopted from the American National Election Study.

Respondents were then given a prompt stating that they would be shown a short advertisement, and then served one of the five advertisements detailed above at ran-

dom. Randomization was done using permuted block randomization over a discrete uniform distribution; $P(a) \sim \mathcal{U}\{1, 5\}$.

After viewing the advertisement, respondents were asked three post-treatment items (plus a manipulation check). These included their favorability rating of Biden and Trump on one to five scale, and their voting intention. Given that the survey was run very close to the actual election and there were high rates of early voting, respondents were given the option to state whom they had already voted for.

The data gathered in the first stage was then used to train a predictive model to learn Biden favorability as a function of the pre-treatment covariates and advertisement shown. Thus given any profile of pre-treatment covariates, the predicted outcome under each of the five advertisements would be calculated $\{\hat{Y}(D_1), \hat{Y}(D_2), ...\hat{Y}(D_5)\}$, and the advertisement that minimized Biden favorability, $D^*(X_i)$, would be allocated:

$$D^*(X_i) := argmin_a \ f(X_i, D_{i,a})$$

In the second stage, respondents first answered the same pre-treatment questions. These answers were sent to a server-side `Python` kernel, which given the pre-treatment covariates X_i, used the pre-fitted RF model to allocate the optimal advertisement $D^*(X_i)$. The respondent then watched the advertisement and answered the same three post-treatment items. The first and second stage were therefore indistinguishable from the perspective of the respondent, as they were

unaware which advertisements they were not shown and how the advertisement was allocated to them until an end-of-survey debrief.

Provided that there are no systematic differences between the first and second stage, we can compare the outcome between randomly assigned stage 1, $\mathbb{E}_a[\mathbb{E}_i[Y_i(D_{i,a})]]$, and optimally assigned stage 2, $\mathbb{E}_i[Y_i(D^*(X_i))]$, as an estimator of the average effect of targeting:

Stated in potential outcomes notation:[15]

$$ATE = \mathbb{E}_i[Y_i(D^*(X_i))] - \mathbb{E}_a[\mathbb{E}_i[Y_i(D_{i,a})]]$$

Following a power analysis based on a simulated run of the experiment using the replication data for Coppock, Hill, and Vavreck (2020), the total N was set at $2,400$,[16] with $1,500$ respondents allocated to the control group and 900 respondents allocated to the treatment group. All participants are United States citizens, resident in the United States, of voting age. Because the design requires a sizable sample to undertake the experiment in a short window of time, the experiment was conducted online rather than in-person or over the phone. Respondents were recruited via the survey provider Prolific and redirected to a custom-built website at `https://survey.polinfo.org`.

[15] *Note on Notation:* In the control group (right-hand term of equation), I average the value of the outcome over all individuals ($\mathbb{E}_i$) and an equal probability of seeing any particular advertisement ($\mathbb{E}_a$). In the treatment group (left-hand term of equation), the advertisement allocation is not random, but a function of pre-determined respondent traits ($D^*(X_i)$).

[16] See Figure 1.7.

Case and Advertisement Selection

The 2020 US presidential election was chosen as the setting for this experiment for several reasons. Enormous campaign spending (Baldwin-Philippi 2017) and a relative lack of regulation (Dobber, Ó Fathaigh, and Zuiderveen Borgesius 2019) made the US a relevant setting with a wide selection of advertisements made for a targeted campaign.

The five advertisements were selected from nearly one hundred videos with a length between 15 and 35 seconds posted by the Trump campaign to their YouTube channel in the final five months of the 2020 United States presidential contest. These were downloaded using the `youtube-dl` tool and hosted on the survey website listed above.

After narrowing down the videos to 10 likely candidates, I asked a panel of ten doctoral students at the our university to help select five advertisements based on the criteria that:

- The advertisements were clearly tailored to different audiences.
- No one advertisement was likely to outperform all others for all respondents.

Irregularities and Attention Check

Responses that failed one of two checks have been omitted from the data used for this article. Prolific provides basic demographic data on respondents that can be downloaded after respondents have completed the survey. Responses where there were

considerable discrepancies between answers and supplied demographic information were rejected.

There was also a attention check immediately after the advertisement, which asked respondents which campaign ran the advertisement ("My name is _____ and I approve of this message"). Given that the answer was provided in the last few seconds of the advertisement, and the question was asked less than a few seconds later, I assumed that respondents who failed this were not paying attention to the video and therefore rejected their responses from the final data.

Sample Representativeness

As mentioned, due to resource constraints it was not possible to procure a representative sample. Nevertheless, the sample provided by Prolific covered all 50 states and Washington DC, with a roughly proportional number of respondents to the population of each state (see Table 1.4).

Table 1.2: Race and ethnic representation.

Race	Census (2019)	Sample
White	60.1%	66.29%
Black	13.4%	9.16%
Hispanic	18.5%	7.03%
Asian	5.9%	12.8%

Table 1.3: Income bound representation.

Income Bound	Census (2019)	Sample
$0 to $53.5k	40%	54.67%
$53.6k to $109.7k	30%	29.4%

Table 1.4: Number of respondents per state (or district), with percentage over/under representation.

State	N	Distortion	State	N	Distortion
Alabama	17	0.74%	Montana	5	0.1%
Alaska	3	0.09%	Nebraska	12	0.06%
Arizona	53	-0.13%	Nevada	31	-0.43%
Arkansas	10	0.48%	New Hampshire	6	0.15%
California	307	-1.54%	New Jersey	79	-0.79%
Colorado	33	0.29%	New Mexico	8	0.28%
Connecticut	15	0.42%	New York	189	-2.43%
Delaware	10	-0.15%	North Carolina	78	-0.25%
District of Columbia	8	-0.14%	North Dakota	1	0.19%
Florida	185	-1.64%	Ohio	73	0.33%
Georgia	69	0.18%	Oklahoma	15	0.54%
Hawaii	10	-0.01%	Oregon	38	-0.4%
Idaho	7	0.23%	Pennsylvania	91	-0.12%
Illinois	105	-0.78%	Rhode Island	6	0.06%
Indiana	42	0.19%	South Carolina	28	0.33%
Iowa	15	0.3%	South Dakota	2	0.18%

State	N	Distortion	State	N	Distortion
Kansas	14	0.27%	Tennessee	36	0.49%
Kentucky	27	0.17%	Texas	175	1.09%
Louisiana	18	0.62%	Utah	7	0.67%
Maine	12	-0.12%	Vermont	4	0.01%
Maryland	54	-0.55%	Virginia	73	-0.63%
Massachusetts	58	-0.47%	Washington	54	-0.07%
Michigan	54	0.65%	West Virginia	6	0.28%
Minnesota	25	0.61%	Wisconsin	43	-0.13%
Mississippi	12	0.38%	Wyoming	3	0.04%
Missouri	35	0.32%			

Table 1.5: Age and gender representation.

Age		Census			Sample		
>=	<	Both	Male	Female	Both	Male	Female
15	19	8.09%	8.39%	7.81%	3.23%	3.41%	2.77%
20	24	8.25%	8.53%	7.98%	21.3%	21.3%	20.5%
25	29	9.03%	9.38%	8.7%	20.1%	19.4%	20.6%
30	34	8.51%	8.7%	8.33%	19.6%	21.6%	17.8%
35	39	8.32%	8.46%	8.19%	13.1%	14.7%	12%
40	44	7.6%	7.66%	7.54%	8.25%	9.05%	7.76%
45	49	7.9%	7.95%	7.84%	5.41%	4.82%	6.1%
50	54	7.9%	7.9%	7.9%	3.79%	2.23%	5.43%
55	59	8.21%	7.99%	8.42%	2.06%	1.29%	2.88%
60	64	7.99%	7.81%	8.16%	1.62%	0.823%	2.44%
65	69	6.74%	6.52%	6.94%	1.06%	1.18%	0.998%
70	74	5.48%	5.32%	5.64%	0.335%	0.118%	0.554%
75	79	3.63%	3.37%	3.88%	0.112%	0.118%	0.111%
80	84	2.35%	2%	2.68%	0%	0%	0%

Using the CCES data (Ansolabehere, Schaffner, and Luks 2020) and entropy balancing (Hainmueller 2012), I also attempted to simulate the results after adjusting covariate balance to a nationally representative sample. Due to the lack of elderly respondents in our sample, some of the resulting weights were over 700, making the results of this model to be meaningless.

Randomization Check

The causal identification of the treatment relies on there not being any systematic differences between the treatment (targeted) and control (untargeted) groups. That the control data had to be gathered prior to the treatment step leaves open the

possibility of bias due to the difference in time of day. In order to mitigate this bias, the entire experiment was run in as small a window as possible.

The results of Chi-Squared tests of independence on assignment to treatment or control against all of the pre-treatment covariates are presented in Table: 1.6. All hypotheses fail to achieve significance except for ideology, but this is not robust to the Holm (1979) or Benjamini-Hochberg (1995) multiple comparisons corrections. Given the strong correlation between partisanship and ideology ($\rho = 0.747$), I find it unlikely that time-of-day effects can confound ideology while not interacting with partisan identification. I therefore conclude that there was a successful randomization, but for each model I additionally test a variant controlling for all pre-treatment covariates.

Table 1.6: Chi-squared test of treatment on pre-treatment independence, with Holm and Benjamini-Hochberg corrections.

	p	Holm	BH
Age	0.345066	1	0.7111003
Gender	0.9753692	1	0.9753692
Race	0.5646555	1	0.8873158
Income	0.9483788	1	0.9753692
Region	0.234541	1	0.7111003
NewsInt	0.2219223	1	0.7111003
On-Track?	0.9516303	1	0.9753692
Party	0.3878729	1	0.7111003
Ideology	0.02123161	0.2335477	0.2335477

	p	Holm	BH
General Vote	0.7472937	1	0.9753692

A second randomization check took the form of covariate balance checks (see figure 1.3). For no variable did the (standardized) mean difference exceed 0.05.

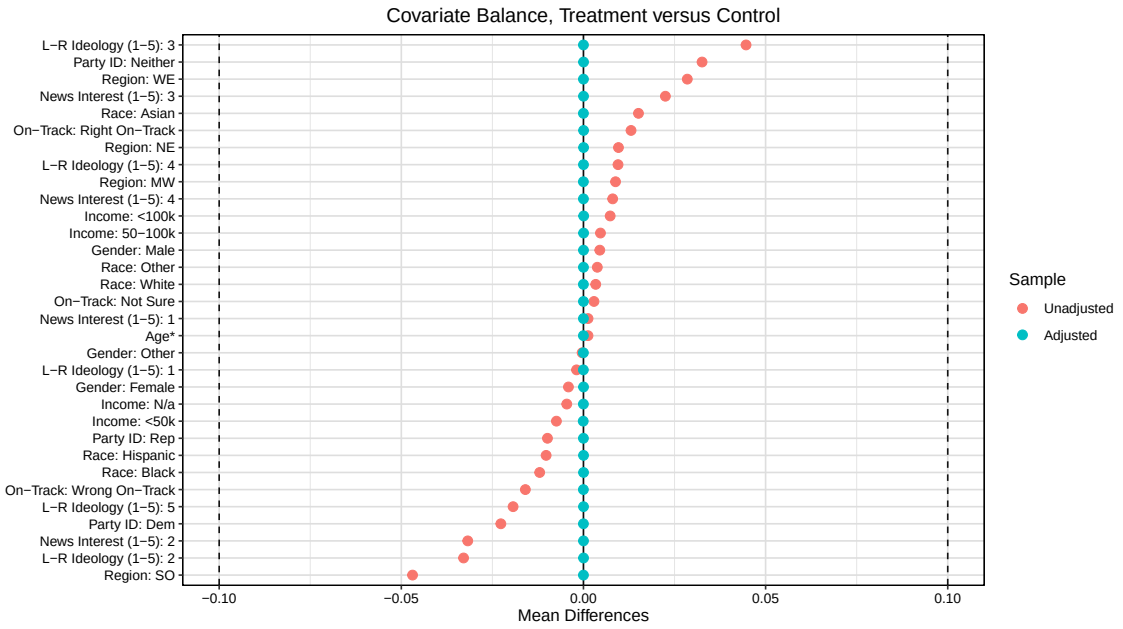

Figure 1.3: Covariate balance on treatment indicator.

A third randomization check took the form of a logistic regressions predicting the advertisement assignment as a function of pre-treatment covariates. The coefficients are displayed in 1.7. It can be seen that no feature is a significant predictor of assignment to treatment or control.

Table 1.7: Logistic regression: predicting stage as function of pre-treatment covariates.

Intercept	−0.26 (0.28)
Age	−0.00 (0.00)
Gender: Female	0.00 (0.09)
Gender: Neither	−0.03 (0.32)
Race: Asian	0.08 (0.14)
Race: Black	−0.12 (0.16)
Race: Hispanic	−0.16 (0.18)
Race: Other	0.06 (0.21)
Income: <50k	−0.01 (0.10)
Income: >100k	0.04 (0.15)
Income: N/A	−0.19 (0.29)
Region: Northeast	−0.02 (0.14)
Region: South	−0.17 (0.13)
Region: West	0.07 (0.14)
News: Most Days	−0.11 (0.15)
News: Sometimes	−0.21 (0.14)
News: Hardly	−0.00 (0.24)
Nation: On-Track	0.19 (0.20)
Nation: Wrong Track	−0.02 (0.16)
Party: Dem.	−0.04 (0.11)
Party: Rep.	−0.10 (0.19)
Ideo: Lib.	−0.11 (0.12)
Ideo: Neither	0.18 (0.16)
Ideo: Con.	0.03 (0.18)
Ideo: Very Con.	−0.45 (0.27)
AIC	3013.90
BIC	3156.99
Log Likelihood	−1481.95
Deviance	2963.90
Num. obs.	2261

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Time-of-Day Effects

The above checks indicate that the observed data are inconsistent with imbalance on observed covariates, but do not address the possibility of the two groups differing systematically in an unobserved way. However, the only kinds of unobserved systematic differences that matter for unbiased estimates of the ATE are those that either directly or indirectly affect the outcome. We can therefore leverage the fact that not all members of each sample took the survey simultaneously to look for evidence of within-sample time trends in the outcome.

Put differently: the threat to inference of the sequential samples is that the time of day at which I collected my samples may be causing systematic differences in the outcome, therefore I cannot be sure that the differences observed between the first and second sample is due to the treatment. However, if there is some systematic relationship between time and the outcome, then we should be able to observe it by viewing the relationship between the outcome and the time at which people took the survey.

But what step changes due to people returning from work, or watching the news? If these coincide with the switch over between stages this might still cause the difference without there being within-sample time trends in the outcome. For this, we can use the fact that the sample was spread over multiple time zones, meaning that the switch over happened at different times of the day for different groups.

I do not find robust evidence for a time trend in the outcome. Figure 1.4 shows the time trend plotted for each sample in the four main time zones the survey

was run (there were also a small number of samples in Alaska and Hawaii, but too few to make inferences). The only group with visual evidence of a time trend is for the Treatment group in Mountain Time. Table **??** shows that this trend is significantly different from zero at $\alpha = 0.05$, but this result is not robust to a multiple comparisons adjustment (Benjamini-Hochberg adjustment: $p = 0.233$; Holm adjustment: $p = 0.446$). Moreover, it seems substantively implausible that this time trend would occur only in one of the four groups and only in one period, especially given that it is between other time zones and the same "shift" should be observable in a neighboring time zone at the corresponding local time.

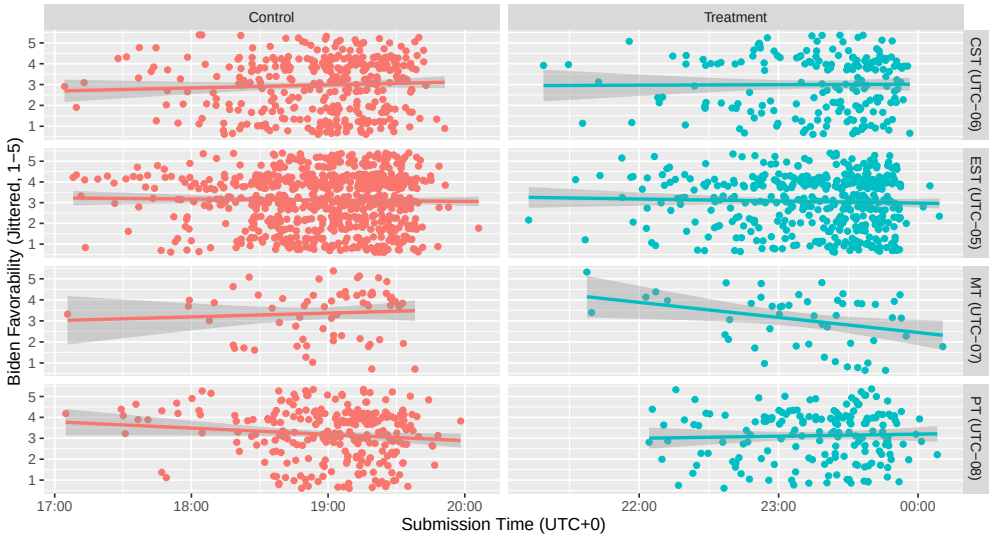

Figure 1.4: Qualitative robustness check of time trends.

Appendix 1B: Theoretical Notes

Predictive Model

As mentioned in the main article, our claims of successful emulation of microtargeting depend in part on the targeting model being sufficiently accurate. There are two scenarios where the treatment effect would not be attributable to optimized allocation. First is zero heterogeneity; if the relative effectiveness of the advertisements is constant across respondents, then the change in average favorability may be due to allocating a greater proportion of respondents a more effective advertisement. This is addressed in the main article, as well as a subsequent section of this appendix. The second is poor predictions, in which case the model cannot be credited with the improvement. Neither of these situations could convincingly be described as emulating microtargeting.

Figure 1.5 addresses these objections. Panel 1 reports per-feature normalized Mean Gini Decrease[17] (nMGD). Note all advertisements have high nMGD, indicating that the algorithm "learned" more about a respondent's Biden favorability from the advertisement they were shown than their gender, income group, or race (with the exception of *Race: White* versus *Ad: Hunter Biden*). Although feature importance does not correspond to any causal estimand, the randomized advertisement assignment in this first stage leaves no explanation for high feature importance other than the heterogeneous effects between advertisements.[18]

[17] See technical explainer (Appendix 1D).
[18] A causal interpretation of stage 1 results is discussed below.

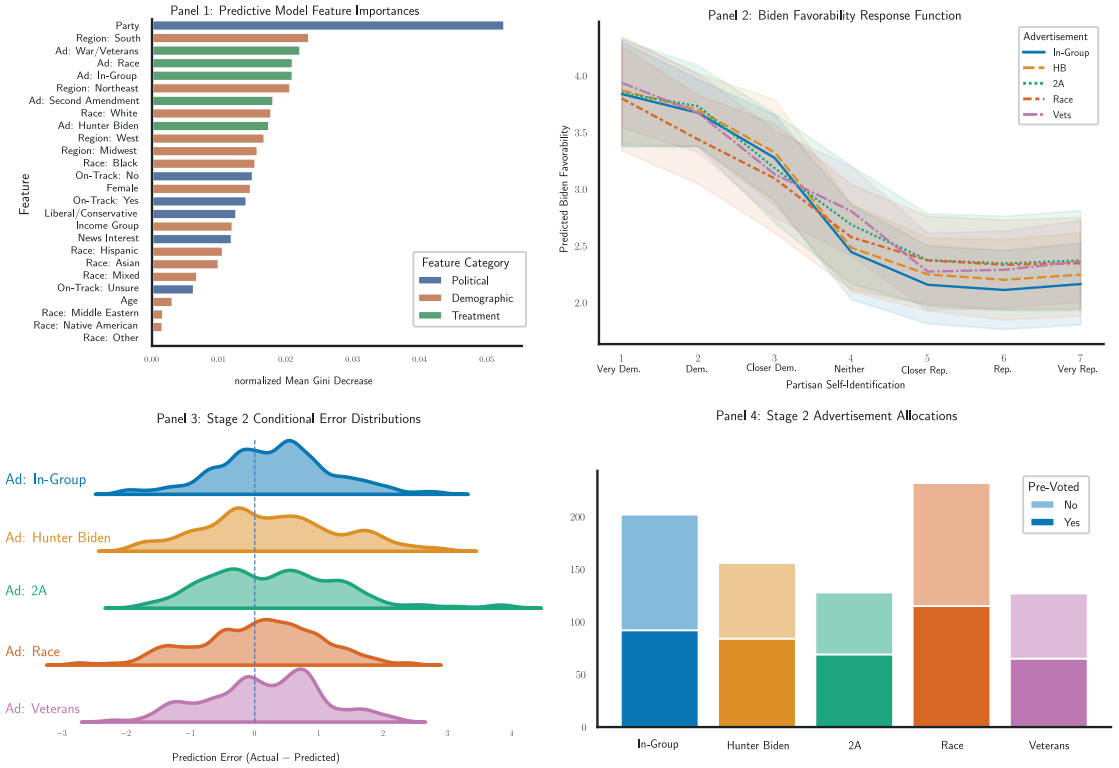

Figure 1.5: Panel 1: feature importances for predictive model used to allocate optimal advertisements; Panel 2: predicted Biden favorability response curve as function of partisanship for stage 2 respondents; Panel 3: conditional prediction error distributions; Panel 4: stage 2 advertisement allocation counts.

The favorability-partisanship response curves in Panel 2 provide further insight into the logic of the targeting model. This panel shows predicted Biden favorability for each advertisement as partisanship is varied, averaged across all stage 2 respondents.[19] The overlapping curves demonstrate between-advertisement and between-respondent heterogeneity over partisanship; one advertisement (*Race*) is more effective on average when the voter is Democrat, while a different advertisement (*In-Group*) is more effective on average when Republican. The large overlapping

[19]See SI-D for interpretation.

standard errors indicate considerable level differences on between-respondent heterogeneity for variables other than partisanship and advertisement.

The conditional error distributions on stage 2 predictions (Panel 3) exhibit a weak positive bias ($\mathbb{E}(e) = 0.21$, $p < 0.001$), indicating that on average the model overestimated the effectiveness of the advertisements. Interpreting this value is difficult because the realization of predicted outcomes is confounded by the model itself. Note the *2A* advertisement has a long right tail, reflecting its greater assignment to individuals with a lower baseline level of Biden favorability, increasing the magnitude of possible prediction error.

It is unclear whether bias is a cause for concern; what is important is that the algorithm did not selectively favor one advertisement. The first three advertisements demonstrate similar conditional average prediction errors (0.31, 0.28 and 0.34 respectively), reducing the likelihood of distorting the ranking between each other. Crucially, Panel 4 demonstrates that no single advertisement dominated the allocations, reducing the likelihood that the difference between treatment and control is due to a more effective advertisement being shown to more respondents.

Characterization

While Panel 1 of figure 1.5 tells us which features mattered, it does not tell us how they mattered. In order to characterize the kinds of respondents that were shown each advertisement, I fit a linear regression for each advertisement to predict

assignment to that advertisement as a function of pre-treatment traits.[20]

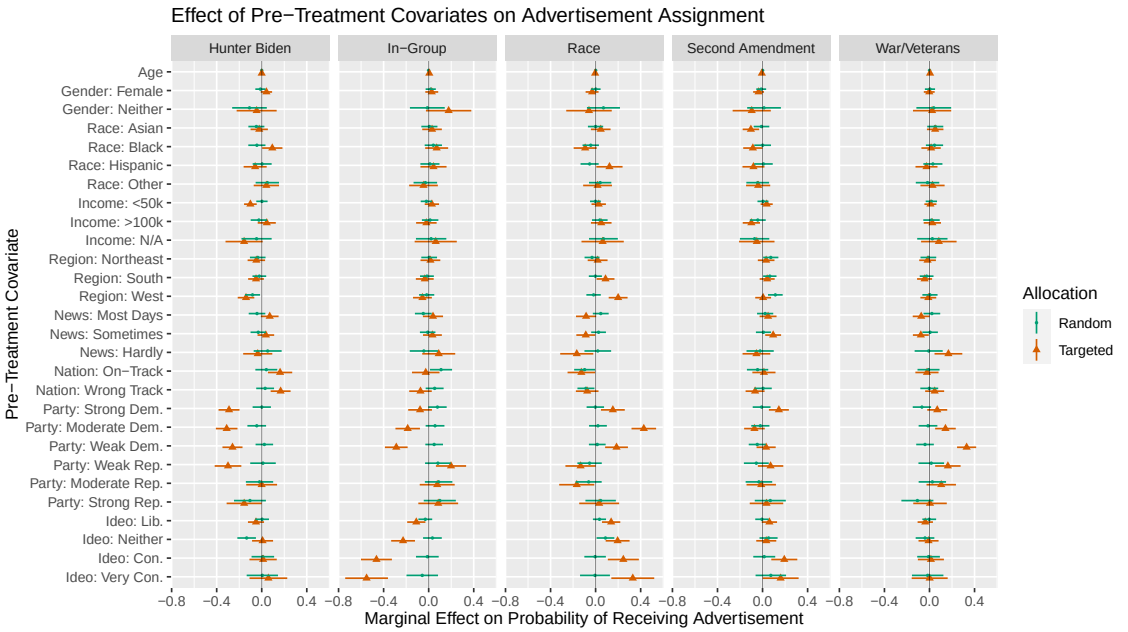

Figure 1.6: coefficient plot of ten linear regressions predicting advertisement assignment

Each panel of the figure represents an advertisement. The red horizontal bars with a triangular marker at the center show the coefficient with 95 percent confidence intervals for the targeted sample, and the blue horizontal bars with the circular marker show the coefficient and 95 percent confidence intervals for the sample receiving randomized advertisements. These coefficients can be interpreted as the average effect of a one-point change in the corresponding pre-treatment variable denoted on the y-axis label on the probability of receiving the advertisement in the label at the top of the panel.

For instance, looking at the red bars, we can see that relative to being completely

[20]In simpler terms: to fit the model predicting the effect of pre-treatment traits on receiving the Hunter Biden advertisement, a new dummy variable was created that took the value 1 if the respondent received the Hunter Biden advertisement, and 0 otherwise.

unaligned, identifying as a Democrat or a weak Republican reduced the probability of receiving the Hunter Biden advertisement by roughly 0.3. A few patterns are unintuitive. In particular, it appears that respondents identifying as ideologically conservative were more likely to receive the Race advertisement.

The blue bars provide evidence of sample balance in the first stage; given that advertisement allocation was randomized, it should not be predictable from any pre-treatment covariates.

Decomposition

As noted in the design and results section, the average treatment effect of targeting in this experiment can be decomposed into two parts; the change due to between-advertisement heterogeneity, and the change due to within-individual heterogeneity. In this section I explain this decomposition and show how the results translate.

In theory, it is possible to control for this effect by fixing the proportion of advertisements between samples. I opt not to do this because a) this is not a realistic constraint that campaigns face, b) the online bipartite matching with stochastic vertex arrivals problem still has no globally optimal solution (see Goyal and Udwani 2020 for current state of the literature) and c) these issues can be addressed post-hoc with the following simple decomposition.

Denoting the average treatment effect as τ, I note that it can first be decomposed into the sum of per-advertisement outcomes times the probability of receiving that advertisement:

$$\tau = \sum_{j \in A} (p_{j2}\bar{y}_{j2}) - (p_{j1}\bar{y}_{j1})$$

Here $A = [a, b, c, d, e]$, the set of the five advertisements, and the second subscript $[1, 2]$ indicates the stage of the experiment. Note that 2 is the treatment stage.

Rearranging the terms, we get

$$\tau = \sum_{j \in A} (p_{j2} - p_{j1})\bar{y}_{j1} + (\bar{y}_{j2} - \bar{y}_{j1})p_{j1} + (p_{j2} - p_{j1})(\bar{y}_{j2} - \bar{y}_{j1})$$

This can be broken down into three parts:

- $(p_{j2} - p_{j1})\bar{y}_{j1}$: the change in advertisement proportions weighted by the stage 1 outcomes.
- $(\bar{y}_{j2} - \bar{y}_{j1})p_{j1}$: the change in outcome weighted by the (uniform) stage 1 frequencies.
- $(p_{j2} - p_{j1})(\bar{y}_{j2} - \bar{y}_{j1})$: an interaction term.

The first term represents the change due to a greater proportion of individuals being shown a given advertisement; in other words, the change due to between-advertisement heterogeneity. The second term represents the change in individual outcomes for each advertisement between stages; in other words, the effect of allocating individually better advertisements.

This table shows that 98 percent of the increase in Biden dislike was attributable to changes in individual outcomes; in other words, the effect of showing respondents

individually best advertisements. Likewise, roughly 2 percent of this change is attributable to changing the mixture of advertisements. This pattern holds across models, where in all cases the change is almost entirely attributable to within-individual heterogeneity.

Relation to Coppock, Hill, and Vavreck (2020)

This paper makes extensive reference to Coppock, Hill, and Vavreck (2020) (within this section of the appendix, CHV20). CHV20 and its replication materials, published just over a month prior to this experiment, were essential to its design and testing. The replication data was used as mock results for the first stage, which were then used to train predictive algorithms and simulate the targeting in stage two.

The simulated targeting experiment based on the CHV20 data indicated that targeting would have an effect, with a magnitude of roughly 0.4 on a 1-5 scale. This result, however, is difficult to interpret because the expected outcome under targeting was given by the same model that was used to optimize advertisement allocations. To increase certainty that this result was not an artifact of the algorithm selectively sampling off the right-hand-side of the standard error, I conducted a permutation test ($n = 30,000$) in which the treatment vector was randomized. The null hypothesis being tested by this permutation test was row-level treatment independence, which would make the targeting irrelevant. This null hypothesis was rejected with $p = 0.99$.

We subsequently used the estimated effect size and variance as the basis for a pre-experimental power analysis, shown in Figure 1.7. On this basis I removed two additional treatment categories and a control advertisement, which I intend to test in a follow-up experiment.

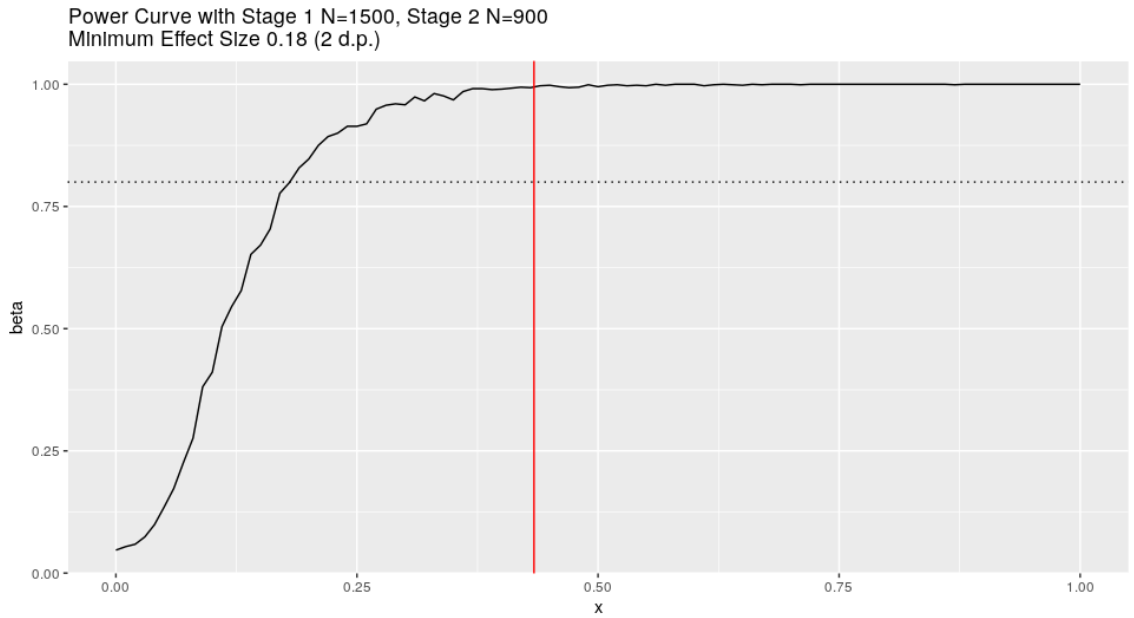

Figure 1.7: *power analysis from CHV20 data*

As noted, the conclusions of this paper seem at odds with CHV20, who find little evidence of treatment effect heterogeneity. They infer from this that there is little room for micro-targeting to work; if the effect of advertisements does not vary greatly from one individual to another, why should micro-targeting make any difference?

We suggest that there are two main reasons for this difference. The first is that whereas CHV20 focuses primarily on heterogeneity over advertisements, this experiment maximizes heterogeneity within voters. Thus, although CHV20 search for conditional effects over the characteristics effect of 49 different advertisements, the only respondent characteristic they test for heterogeneity over is partisanship.

In comparison, the targeting algorithm trained for this experiment searches a space of up to $362,880$ ($9!$) combinations of respondent characteristics to find connections between response to the advertisement and combinations of traits. The respondent is then shown the advertisement predicted to maximize (or minimize) this difference. As referenced in the main body, CHV20 does indeed note that exhaustively searching traits may expose sources of heterogeneity.

Advertisement Effectiveness

A natural estimand of interest is the "effect" of each advertisement. In contrast to other papers in this literature, this paper does not offer an estimate of the persuasive effect of each advertisement.

A theoretical quantity of interest is the effect of each individual advertisement used in this experiment. Randomized advertisement assignment in the first stage allows for a causally identified interpretation of the coefficient on each advertisement, but because it is unclear which advertisement should serve as a reference category, these coefficients are not meaningful. The original design for the experiment included a control advertisement in stage 1 in order facilitate these comparisons, but this was omitted to prioritize the key components of this experiment: maximizing the number of observations for training the predictive algorithm and identifying the effect of targeting. The implementation of the control group can be seen in the source code of this project on GitHub.

Envelope Calculation

The reader may be wondering what are the substantive implications of the estimated effect size. The experiment shows that in a hypothetical campaign where the Trump campaign ran just five advertisements, optimizing allocation with the method presented in this paper would have resulted in 8.7 percentage points more of unaligned voters stating that they do not like Biden, and 7.1 percentage points fewer of unaligned voters stating that they intend to vote for Biden, when compared to a scenario when advertisements are assigned independently of individual traits.

What do 8.7 and 7.1 percentage points mean? While it is not possible with the available data to give a rigorous estimate, the following calculation gives an illustration of how meaningful a 7.1 percentage point swing could be. In the 2020 United States presidential election, 31 states included information on partisan affiliation for voter registration. Assuming (unrealistically) that all registered voters turn out to vote, that stated intention not to vote for Biden translates into voting for neither candidate (i.e. no switching), then we can multiply the proportion of unaligned voters in each state by the effect size in order to estimate how the result would have changed.

For example, in Florida, where 26.7% of registered voters affiliated with neither of the major parties, a swing of 7.1 percentage points among unaligned voters could lead to a diminution of $0.267 \times 0.0708 = 0.0189$, or 1.9 percentage points in the Biden vote share. In North Carolina and Arizona, this estimated change is larger than the percentage point difference between the shares of the two leading candidates

(in Arizona, 35.1% unaligned with a 0.3% margin of victory, and in North Carolina 30.6% of voters are unaligned with a 1.3% margin of victory).

There are many obvious caveats and limitations to this calculation beyond the unrealistic assumptions already stated. For one, changing answers on a survey is radically different to changing voting intention. For another, this survey design and calculation do not account for decay and counter-acting effects, meaning the actual effect is likely smaller. A further point is that the 7.1 percentage point effect is based on a distribution of Biden preferences that were present in the sample; this is likely not the same for all possible distributions of starting preferences. Finally, the sample used in this survey is not representative, meaning that the distribution of effects will likely be different. Nevertheless, this calculation is meant to illustrate that the magnitude of effect is not trivial in a political context like the United States, where consequential races have incredibly small margins.

Normative Implications

These results highlight the potential for a multitude of harms. Arguably the clearest threat is to voter privacy, which this technology creates incentives to disregard (Wachter 2017). Moreover, because those using this technology are more likely to win power, this creates a self-reinforcing cycle (Hindman 2018, 5) in which private actors (Zuboff 2015; Baldwin-Philippi 2017) and public officials (Krotoszynski 2020) broker access to such data.

There is a second, less clear, set of normative issues that speaks to manipulation

and the legitimacy of democratic processes. It begins with asking "what does it mean to show a voter the *right* advertisement?" Proponents of targeted advertising often present it in a benign and efficiency-improving frame; targeted advertising reduces the informational burden by showing the consumer only the goods that are most relevant to their interests (Hindman 2018, 39). Analogously, targeted political advertising has the potential for good; it can be used to show voters elements of candidates' policies that are most relevant to their interests, lowering the cost of civic and political engagement.

Targeted advertising can also be malevolent. Combining emotive advertisements with knowledge of the fears and anxieties of the voters,[21] targeted negative advertising could fairly be called a form of manipulation (Sunstein 2015). By seeking to identify ways to elicit strong negative reactions from voters, campaigns are attempting to engage peripheral, or emotive, and not central, or cognitive/logical, processing of the information in advertisement. This approach to advertising attempts to bypass the deliberative and cognitive capacity of the viewer, ultimately seeking to make the decision for the viewer.

This negative and manipulative strategy is a more suitable description of the calculus employed by CA and demonstrated in the five Trump campaign advertisements used in this experiment. The psychometric profiling employed by CA targeted "neurotic" voters that would be especially susceptible to fear-based advertising (Hindman 2018). The five advertisements similarly appear to be attempting to elicit negative emotions

[21]Note that with the black box algorithmic approach to targeting employed here, the advertisers may not specifically know what aspects of an advertisement make it resonate with an individual. This does not mean, however, that the algorithm is not implicitly leveraging these patterns or identifying these psychological traits.

to be associated with Biden, from the fear that he will take away the viewer's guns, to emphasizing that Biden and Clinton laugh at and mock people like the viewer. To determine the extent to which this strategy characterizes the various United States campaigns requires further research.

However, these are issues with manipulation and not targeting. At an individual level, manipulative political advertising could be seen as disenfranchising, by depriving voters of the opportunity to make their own decision. The results from this experiment show that such manipulation is possible on a massive scale, and can therefore affect the outcome of an election. That election outcomes could be determined by micro-targeted manipulative advertisement campaigns undermines the legitimacy derived from the democratic process, as such outcomes arguably no longer reflect the public will. In a nutshell, although campaigns can engage in manipulative practices without targeting, to the extent that a manipulative campaign is effective at a societal level the threat transcends from individual to systemic.

Appendix 1C: Regression Tables and Robustness

The following appendix contains the full results of the various models and operationalizations, as well as various robustness checks.

All Operationalizations and Specifications

In this section of the appendix, the various regression models reflect different methods of operationalizing the outcome. As these decisions should not, and were not made arbitrarily, we explain the details and rationale here. For the candidate favorability question, after watching the advertisement, survey respondents were posed the following question:

> *"On a scale of 1-5, how would you rate Democratic Presidential Candidate Joseph Biden? (A rating of 4 or 5 means that you feel favorable and warm towards the candidate. A rating of 1 or 2 means that you don't feel favorable and warm towards the candidate. You would rate them at 3 if you don't feel particularly warm or cold toward them.)"*

This item is operationalized in four ways. The first two keep the five categories, but treat it either as a continuous linear scale, or as an ordinal categorical scale. The latter two treat it as a binary indicator of Biden dislike, with 1-2 being coded as 0, and 3-5 being coded as 1. The direction of the measure and inclusion of 3 in the positive category are not arbitrary; Biden *dislike* should be of particular interest to

campaigns. If they can get voters to actively dislike the candidate (1-2), then this is more significant than simple indifference (3).

The vote choice question simply allowed them to indicate whether they intended to, or had already voted for Trump, Biden, another candidate, that they did not intend to vote, or that they did not wish to answer. The intent to vote Biden variable is simply a binary indicator on whether they chose "I intend to vote for Biden" for this question.

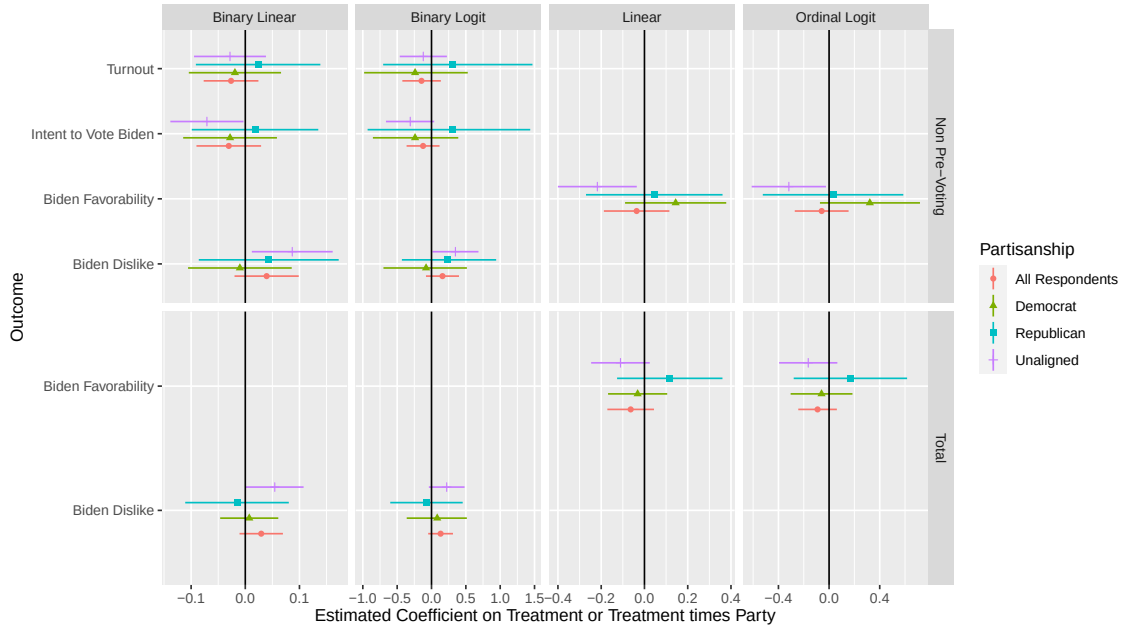

Figure 1.8: Coefficient and 95% confidence interval on treatment indicator for all specifications.

Each dot-and-whisker in Figure 1.8 shows the coefficient and 95% confidence interval on the treatment indicator for one of 48 different models. The top four panels contain models fitted to the subset of respondents who had not voted prior to the survey ($N = 1,160$), whereas the bottom four panels contain coefficients for the full sample ($N = 2,261$). The left-to-right, the panels show linear models fit to a binary outcome

(Binary Linear), logit models fit to a binary outcome (Binary Logit), linear models fit to a continuous outcome, and an ordinal logit model fit to a Likert outcome. The ticks on the y-axis designate the outcome of the model; *Turnout* is a binary indicator on whether the respondent stated that they intend to cast a vote in the election, *Intent to Vote Biden* is a binary indicator on whether the respondent stated that they intended to vote for Biden, *Biden Dislike* is a binary indicator rating Biden lower than "Neither" higher on a favorability scale, and *Biden Favorability* Likert scale ranging from "Strongly Dislike" to "Strongly Like".

The coefficients are presented in groups of four, with each shape/color corresponding to which covariates were included in the model. The red circle indicates the coefficient on treatment for a model regressing only treatment on outcome. The green triangle indicates the coefficient on treatment for a model regressing treatment interacted on partisanship with Democrats set to the base partisanship category. The blue square and purple vertical line indicate the same with Republicans and Unaligned voters set as the base partisanship category respectively.

Effect on Candidate Favorability

Table 1.10 reports the effect of targeting conditional on partisanship for the subset of respondents who had not already voted at the time of the survey ($N = 1,160$). The coefficients in this table reveal a substantial amount of treatment effect heterogeneity. The second row of coefficients, *Targeting (Unaligned)*, shows the effect of targeting on respondents who had not already cast their votes and self-identify as neither party; the group that a campaign would aim to target. The estimated CATE is -0.218

points on a five-point scale ($SE = 0.093$), *which translates to an* 8.7 *percentage point increase in respondents saying they dislike Biden* ($SE = 3.8$pp). Both of these coefficients are significantly different from zero at standard confidence levels of $\alpha = 0.05$ ($p = 0.0192,\ 0.0338,\ 0.0228,\ 0.0416$).

The remaining coefficients reveal unsurprising patterns. The effect of self-identifying as Democrat and Republican has large and significant effects on Biden favorability in the expected directions. Looking at the CATEs of self-identification as Democrat or Republican, we discover that their coefficients are in the same direction and diminish the effect of targeting. In other words, targeting anti-Biden advertisements has the strongest effect on unaligned voters, but has a relatively weak effect on voters who identify with a party.

Effect on Voting Preference

The dependent variable for the models in this section is the proportion of respondents stating their intention to vote for Biden in the general election, out of the respondents who had not voted at the time of the survey. Table 1.11 reports the effect of targeting on intention to vote for Biden among respondents who had not voted at the time of the survey. The two columns on the left report the models regressing targeting on voting preference, and the two columns on the right report the models regressing targeting interacted on partisan self-identification on voting preference. The columns alternatingly report the results for OLS and logistic regression models.

The first pair of models (1) and (2) show a weak, insignificant and negative ATE of

targeting on voting preference. When we search for heterogeneity over (non-) partisanship, we observe a similar pattern to the CATE of targeting on candidate favorability. The effect of targeting on intention to vote for Biden is -0.071 ($SE = 0.034$, $p = 0.03967$), meaning that among respondents who identified with neither party and had not voted at the time of the survey, being targeted increased the proportion of respondents not intending to vote for Biden by 7.1 percentage points. As with the previous section, the effect of aligning with either the Democratic or Republican party largely nullifies the effect of targeting. That the pattern is persisting in a separate but related outcome increases confidence that targeting is in fact increasing the likelihood that the targeted advertisements are persuasive.

Effect on Turnout

The final set of results, shown in table 1.12, looks at the effect of targeting on turnout on respondents who had not voted yet. This is operationalized as a dummy variable indicating the proportion of respondents stating that they will vote for Biden, Trump, or a third candidate. The models are presented in the same as the previous section, with the first two columns testing an uninteracted model and the latter two testing a model interacting turnout intention on partisan self-identification.

These models indicate that if there is an effect of targeting on turnout, then it is not significantly different from zero in the total sample nor among non-partisan voters. It is worth noting that partisan voters were more likely to indicate that they intended to vote by 28.6 and 26.7 percentage points for Democrats and Republicans respectively, from a baseline of 63.5% for non-partisan voters.

Appendix 1D: Technical Explainers

This appendix contains brief explanations of technical concepts not assumed to be known by a reader with a political science background.

(normalized) Mean Gini Decrease

Unlike ordinary least squares (OLS), the random forest model employed in the second stage does not have regression coefficients to facilitate interpretation of the predicted outcome as the sum of linear components. Nevertheless, decision-tree based algorithms do provide metrics that allow substantive interpretation (Montgomery and Olivella 2018).

Panel 1 of figure A.2 article shows the per-feature normalized Mean Gini Decrease (nMGD) statistic from the RF model used to assign allocations in the second stage. nMGD tells us the degree to which partitions on a feature produce homogeneous sub-spaces at each node of the individual decision trees. In other words, this statistic tells us the extent to which a feature produces systematic and separable regions of the label space. Because Mean Gini Decrease is biased upwards for features with high cardinality, we normalize the score by dividing by cardinality.

A question of substantive interest to both researchers and campaigners alike is why the predictive algorithm assigned a particular advertisement and not another. Features with high nMGD provide much of the predictive power for RF models. Parti-

san self-identification has the highest feature importance, followed by the respondent being from the South.

Average Response Curve

Panel 2 of Figure 1.5 presents, for each advertisement, the average response curve of the outcome (Biden favorability) across values of partisanship on a seven-point scale. This was calculated by using the fitted RF model to predict Biden favorability for each of the stage 2 respondents at each level of partisanship (1 through 7), and then averaging the favorability across respondents for each partisanship-advertisement combination. The shaded areas represent the standard deviation of the predicted response curves for each advertisement.

Focusing just on the solid lines (and not the shaded areas), the interpretation of the figure is as follows. At each level of partisanship, the vertical ordering of the lines represents the relative average predicted effectiveness of the various advertisements. For example, at partisanship *2: Dem*, the line corresponding to the *Race* advertisement is lower than the other four. This indicates that on average, for the stage 2 respondents, if we set partisanship to *2* on the seven-point scale and leave all other traits as they were, the targeting algorithm predicts that showing the *Race* advertisement will produce a lower Biden favorability other advertisements. Similarly, when we move up the scale to partisanship of *5*, *6* or *7*, the *In-Group* advertisement is predicted to produce a lowest Biden favorability on average, followed by the *Hunter Biden* advertisement.

The shaded areas, as mentioned, indicate the standard deviation of each of these response curves. You may note that they are largely overlapping, perhaps with the exception of *Race* at partisanship *2*. The takeaway here is that there is still a great deal of heterogeneity in individual response curves (the predicted levels of Biden favorability for each individual, artificially varying their partisanship). This should not be surprising; we would expect other demographic traits, such as state of residence, to correspond with level shifts in baseline levels of Biden favorability. Note also that artificially varying partisanship results in some very implausible trait combinations (e.g. a young ideologically left-wing hardcore Republican).

Ultimately, the takeaways from this figure are as follows:

- *Crossing Curves*: the algorithm predicts that different advertisements are the most effective at different levels of partisanship. This provides evidence supporting the hypothesis that the effect of micro-targeting is a "substitution" effect, and not an "income" one.
- *Large Overlapping Errors*: the algorithm predicts that there is a high level of heterogeneity in response corresponding with traits other than partisanship. In other words, while partisanship and the advertisement shown matter, they do not explain the full range outcomes.
- *Small Difference at Very Democrat*: that the *Race* advertisement is comparatively much more effective at moderate Democrat partisanship (*2*) than extreme Democrat partisanship (*1*) indicates that there may be no persuasion effect for any advertisement when the respondent identifies as strongly Democrat. This is consistent with the findings of Broockman and Kalla (2020).

Appendix 1E: Ethics

As an experiment involving mild deception, and dealing with difficult ethical issues, every step was taken to ensure that this research was conducted in a way that respected and protected participants and avoided impacting wider processes such as the ongoing election. Where possible, the researcher went beyond the requirements of the ethical review board and erred heavily on the side of caution.

IRB Approval

Prior to the experiment, ethical approval was applied for and obtained from the Oxford University Research Ethics Committee, Departmental Research Ethics Committee (DREC).

Recruitment and Payment

Participants were recruited via academically-focused online survey provider Prolific, who require that researchers provide fair compensation for task completion. Accordingly, respondents were compensated at an average rate of USD 11.51 per hour (with a median completion time of three minutes). A call for responses was posted via the Prolific portal with the following text:

```
Political Ads Survey
```

Does Microtargeting Work?

```
Summary:
    - 2-4 minute academic survey.
    - Involves 13 multiple-choice questions and a 30-second ad.
Compatibility:
IMPORTANT: Please disable your ad blocker,
           as some participants have had
           issues watching the video while it is enabled.
    - Tested on Linux/Mac/Windows for Firefox/Chrome.
    - There are no cookies/external ads on this survey website.
Questions:
    - 5 demographic
    - 4 political
    - 4 about the ad
```

Respondents were therefore given a great deal of up-front information regarding the content and objective of the task, so that they could decide whether they wanted to participate.

Respondents who selected this study were provided with a link to an external website, `survey.polinfo.org` along with a `GET` url parameter to forward their prolific user id. This website was built and hosted by the researcher, and all connections to it were forced to use encrypted `https` in order to minimize the risk of a data leak. The aforementioned url parameter is only linkable to real identities by the survey provider, Prolific.

Consent

The first page of the survey, `https://survey.polinfo.org/consent.php`, explained in broad terms the purpose of the research (researching the effects of political advertising), obtained participant consent, and made clear that respondents had the option of freely revoking their consent at any point with no penalty. It also made clear to participants that a longer debrief would be available at the end of the survey, and that this would contain additional information regarding the purpose of the experiment.

Deception

This design involves deception by omission. The real objective of this study was to study the effect of targeted advertising. Participants did not consent to having their data used to train a targeting algorithm, or being targeted by said algorithm, prior to these things being done.

To address this, at the end of the survey, participants were shown an extensive debrief explaining the intention of the experiment (measuring the effect of targeted advertising) and that either a) their responses had been used to train a targeting algorithm or b) they had been targeted with the advertisement they were shown based on the answers they gave. That participants were able to revoke consent at any point was emphasized again, and the contact details of the researcher were provided again to answer questions and concerns.

Impact

The debrief emphasized that these advertisements were run by the Trump campaign, and that the purpose of this experiment was to show the effect of targeting the advertisement. Research into political microtargeting finds that informing participants of the targeting alters engagement with the message (Kruikemeier, Sezgin, and Boerman 2016), such that revealing the intention of the research is likely to counteract the effect of the advertisement itself.

Moreover, from follow-up discussions with participants, more than 100 indicated that they had been "inundated" with political advertisements in the period leading up to the survey, and as such were unable to remember the content of the advertisement they were shown even half an hour after the experiment. Specifically, many were unable to even recall whether we had shown them an advertisement run by the Trump or Biden campaign.

Confidentiality and Data Privacy

All connections were made to the survey website via secured connection, and all answers were stored locally. Pseudo-anonymised information taken during the course of the survey was a unique user id provided by Prolific, which the researcher does not have the means to convert to real identities.

Responses were securely transferred to the researcher's local device via SQL over SSH. A copy and backup are held by the researcher in password-protected databases.

The reproduction data will be scrubbed of any metadata provided by Prolific, to only contain the answers to the questions.

A copy of the consent responses, linked to Prolific IDs, will be held for the required period of time in order to permit the researcher to take action in the case of revoked consent.

Funding

This study was made possible from the researcher's own expenses and a generous grant from the Merton College Graduate Research fund.

Conflicts of Interest

The author hereby declares that they have no conflicts of interest, personal or professional, relating to this study.

Appendix 1F: Reproducibility

Reproduction code for both the experimental design (website front end and back end) as well as scripts for analysis will be hosted on Github at `https://github.com/<REDACTED>`. Data will be made available via Dataverse once appropriately sanitized and the period stated in the consent form for revoking consent has expired.

Technical Implementation

The design of this survey required a high degree of interactivity. In the second stage, the respondents' answers were sent to an on-line pre-trained machine learning model that would respond with the optimal advertisement assignment in real time. Given that there was no commercial survey software that provided this integration functionality, the survey website was built by the researcher from the ground up.

The front-end of the website[22] was largely written in PHP and hosted on an AWS Lightsail instance running a Linux-Nginx-MariaDB-PHP (LEMP) stack. PHP was likewise used to communicate with the back-end in real-time, which consisted of a Jupyter kernel and MariaDB database.

The machine learning algorithm was implemented in Python using the scikit-learn library (Pedregosa et al. 2011). Due to severe resource constraints on the Lightsail

[22]CSS and other styling elements were copied then modified from `https://surveyjs.io/`.

server, the algorithm was trained during the switch-over on a local machine, and the trained model was stored as a `joblib` binary then uploaded via `ssh` connection.

Interactivity between PHP and the algorithm was implemented using a modified Jupyter interface, which also handled queue management. Prediction response time during peak loads never exceeded 100ms.

The source code for the website is hosted on Github at `https://github.com/<REDACTED>`.

Open Source Attributions

This research would not have been possible without the enormous contributions of the open-source software community. In particular, the following libraries were integral to this research:

- `Scikit-Learn` (Pedregosa et al. 2011)
- `Jupyter` (Kluyver et al. 2016)
- `Numpy` (Harris et al. 2020)
- `Pandas` (team 2020)
- `Seaborn` (Waskom 2021)
- `Matplotlib` (Hunter 2007)
- `ggplot2` (Wickham 2016)
- `Tidyverse` (Wickham et al. 2019)

Table 1.8: Robustness check: outcome as function of time survey completed additional models interacted by group and time zone.

	Model 1	Model 2	Model 3	Model 4
Intercept	3.15***	4.00	3.17***	4.00
	(0.05)	(2.77)	(0.13)	(2.77)
Time (secs.)	−0.00	−0.00	−0.00	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)
CST (UTC-06)		−0.84		−0.77
		(2.77)		(2.77)
EST (UTC-05)		−1.05		−1.30
		(2.77)		(2.78)
PT (UTC-08)		−0.45		−0.98
		(2.78)		(2.84)
HST (UTC-10)		−0.77		−0.23
		(2.77)		(2.79)
MT (UTC-07)		0.31		−0.50
		(2.86)		(3.51)
Time*CST		−0.00		−0.00
		(0.00)		(0.00)
Time*EST		0.00		0.00
		(0.00)		(0.00)
Time*PT		−0.00		0.00
		(0.00)		(0.00)
Time*HST		−0.00		−0.00
		(0.00)		(0.00)
Time*MT		−0.00		0.00
		(0.00)		(0.00)
Stage 2			0.32	−1.26
			(0.55)	(1.34)
Time*Stage 2			−0.00	0.00
			(0.00)	(0.00)
Stage 2*CST				1.70
				(1.55)
Stage 2*EST				1.42
				(1.75)
Stage 2*HST				5.65*
				(2.45)
Stage 2*MT				−5.51
				(8.00)
Time*Stage 2*CST				−0.00
				(0.00)
Time*Stage 2*EST				−0.00
				(0.00)
Time*Stage 2*HST				−0.00*
				(0.00)
Time*Stage 2*MT				0.00
				(0.00)
R^2	0.00	0.01	0.00	0.01
Adj. R^2	0.00	0.00	−0.00	0.00
Num. obs.	2252	2252	2252	2252

Table 1.9: Effect decomposition for four models. Subset refers to unaligned and non pre-voting respondents.

	Dislike Biden		Vote Biden	
	Total Sample	Subset	Total Sample	Subset
$(p_2 - p_1)y_1$	0.006649	0.006460	0.006649	-0.001744
$(y_2 - y_1)p_1$	0.034626	0.085389	0.034626	-0.069409
$(y_2 - y_1)(p_2 - p_1)$	-0.011849	-0.005128	-0.011849	0.000316

Table 1.10: Effect of micro-targeting on candidate favorability, interacted on partisan self-identification among respondents who had not voted.

	OLS: Five-Point	Ordered Logistic	OLS: Binary	Logistic
Intercept	2.588^{***}		0.467^{***}	-0.133
	(0.057)		(0.024)	(0.105)
Targeted (Unaligned)	-0.218^{**}	-0.318^{**}	0.087^{**}	0.348^{**}
	(0.093)	(0.150)	(0.038)	(0.171)
Democrat	0.997^{***}	1.625^{***}	-0.314^{***}	-1.580^{***}
	(0.092)	(0.159)	(0.038)	(0.212)
Republican	-0.693^{***}	-1.276^{***}	0.269^{***}	1.159^{***}
	(0.108)	(0.189)	(0.044)	(0.216)
Targeted × Democrat	0.362^{**}	0.640^{**}	-0.097	-0.427
	(0.151)	(0.251)	(0.062)	(0.353)
Targeted × Republican	0.263	0.352	-0.043	-0.112
	(0.186)	(0.320)	(0.076)	(0.388)
R^2	0.258		0.190	
Adj. R^2	0.254		0.187	
Num. obs.	1160	1160	1160	1160
AIC		3259.343		1363.093
BIC		3304.849		1393.430
Log Likelihood		-1620.671		-675.547
Deviance		3241.343		1351.093

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1.11: Effect of micro-targeting on intention to vote for Biden among respondents who had not voted.

	OLS	Logit	OLS Interacted	Logit Interacted
Intercept	0.478^{***}	-0.090	0.392^{***}	-0.438^{***}
	(0.018)	(0.074)	(0.021)	(0.108)
Targeted (Unaligned)	-0.030	-0.123	-0.071^{**}	-0.309^{*}
	(0.030)	(0.122)	(0.034)	(0.179)
Democrat			0.485^{***}	2.409^{***}
			(0.034)	(0.229)
Republican			-0.337^{***}	-2.395^{***}
			(0.040)	(0.379)
Targeted × Democrat			0.043	0.070
			(0.056)	(0.363)
Targeted × Republican			0.089	0.609
			(0.069)	(0.617)
R^2	0.001		0.346	
Adj. R^2	0.000		0.343	
Num. obs.	1160	1160	1160	1160
AIC		1605.846		1158.461
BIC		1615.958		1188.798
Log Likelihood		−800.923		−573.230
Deviance		1601.846		1146.461

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1.12: Effect of micro-targeting on turnout among respondents who had not voted.

	OLS	Logit	OLS Interacted	Logit Interacted
Intercept	0.777***	1.250***	0.628***	0.523***
	(0.015)	(0.086)	(0.020)	(0.105)
Targeted (Unaligned)	−0.021	−0.118	−0.019	−0.082
	(0.025)	(0.139)	(0.033)	(0.170)
Democrat			0.298***	2.002***
			(0.032)	(0.267)
Republican			0.285***	1.821***
			(0.037)	(0.299)
Targeted × Democrat			0.004	−0.118
			(0.053)	(0.416)
Targeted × Republican			0.035	0.303
			(0.065)	(0.568)
R^2	0.001		0.126	
Adj. R^2	−0.000		0.123	
Num. obs.	1241	1241	1241	1241
AIC		1343.142		1184.504
BIC		1353.389		1215.246
Log Likelihood		−669.571		−586.252
Deviance		1339.142		1172.504

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Learning from Machines: Differentiating US Presidential Campaigns with Attribution and Annotation

Identifying and characterizing ways in which political actors and groups express themselves is a key descriptive task in the study of legislatures, campaigning and communication. A variety of computational tools exist to help find and describe these patterns, typically summarizing differences with weighted word lists representing either lexical frequencies or semantic fields. I identify two limits to the inferences that can be made based on these methods: the ambiguity of the semantic value of words without wider context and an inability to detect differences outside of lexical semantics. As an alternative means of identifying differentiating language usage in political texts, I present a combination of text annotation and deep-learning feature attribution–a set of techniques for evaluating the relative importance of data

inputs to the prediction of a neural network classifier. This utility of this method is demonstrated with application to a dataset of twelve US presidential campaign advertisements from Facebook between 2017 and 2020. The contributions of this paper is 1) tools to conduct statistical, comparative, and substantive validations of the output of a complex black-box model for a social science context and 2) the development of a novel approach to qualitative content analysis with annotation by the logic of deep learning models.

Introduction

How can we summarize differences in rhetoric between Donald Trump and Joe Biden? Observers familiar with both politicians might contrast the policy issues they choose to talk about, such as Trump favoring illegal immigration and Biden favoring infrastructure development. One can also point to differences in the sentiment they try to evoke, with Biden's rhetoric emphasizing unity, especially among Democrats, versus Trump emphasizing the existence of a movement and its enemies; or, specific phrases or slogans associated with the respective candidates "Build Back Better" versus "Make America Great Again". These descriptive differences in rhetoric are key to characterizing and understanding political actors, and are core to a common notion of what constitutes *politics*.

The field of political science has seen the development of a rich methodological field combining tools from applied linguistics, statistics and machine learning for the operationalization of large text corpora and frameworks for producing systematic and robust inferences from these representations (Grimmer, Roberts, and Stewart 2022). These tools have enabled novel inferences about a wide range of political phenomena occurring or recorded in text form, such as the study of legislator attention (Grimmer 2013) or policy framing in news media (Barnes and Hicks 2018).

I argue that while the tools developed in political science for computationally-assisted text analysis are effective for identifying some kinds of linguistic differences, they are limited in their ability to identify and describe others. I attribute this to two design decisions. First, current methods typically use words as a unit of analysis, discarding linking information relating to order, context and syntax. Second,

current methods primarily output reductive numerical summaries, which are then used to weight word lists or documents. This approach links theory and empirics by inferring the semantic values and contexts associated with individual words, and then making claims about the prevalence of these phenomena based on numerical weights. As a result, researchers face two widespread issues: difficulty validating the inferred semantic value of a word because of a disconnect between the source texts and the summaries, and difficulty identifying or describing patterns that occur outside of lexical semantics.

To address these gaps, I adapt several tools for model explanation in deep learning (DL) to produce and validate a computationally-assisted approach for identifying and describing differences in language usage between political actors. In this approach, texts are annotated with feature attribution scores representing whether a region contains language that differentiates the speaker. These feature attribution scores are generated using Integrated Gradients (Sundararajan, Taly, and Yan 2017), a technique for measuring the importance of individual data inputs to the prediction of neural network model, extracted from a DL sequence classifier based on the transformer architecture (Vaswani et al. 2017; Devlin et al. 2019). Instead of using these scores as a scoring method to rank the extent to which words are differentiating, I annotate the source texts with these scores in order to visualize the logic of the DL classifier in context and detect a broader range of linguistic phenomena.

I demonstrate the utility of these methods with an application to characterizing the advertising styles of twelve US presidential campaigns on Facebook in the 2020 election cycle. I first compare the highly contrastive Trump and Biden campaigns as sense-check, to show that the method produces plausible results. Next, I compare the

more similar Sanders and Warren campaigns to show that the method is capable of detecting subtler and unexpected differences. Finally, I compare eleven Democratic primary candidate campaigns, including Biden, Sanders and Warren, to show both the substantive importance of reference category when identifying characterizing features, and that the method can be generalized to the multiclass case.

I find a number of differences between the ways the campaigns express themselves. At the lexical level, I find patterns such as Sanders using the words "campaign" and "donate" versus Warren using the phrases "grassroots movement" and "chip in". At a thematic level, comparing the Biden campaign to other Democratic primary campaigns suggests that the use of Christian-moral and nationalistic language is more predictive of Biden than other candidates.

The main methodological contribution of this paper goes beyond presenting a new method for differentiating language use that is capable of detecting patterns above the individual word level. Despite the importance of DL in machine learning (ML) and natural language processing (NLP) as a flexible, efficient and powerful approach to modelling empirical phenomena, the application of complex neural network architectures in political science has been limited because of the interpretability and validation requirements of quantitative social science research. In this paper, I adapt strategies for interpretation and validation from the explainable ML subfield (Jain et al. 2020) to the social science domain and explain how, why and when they function analogously to standards of proof from a quantitative political science paradigm.

Looking forward, this paper calls for the development of a novel hybrid approach to computational text analysis. By annotating documents with patterns of attribution

permits valid inference and bridges qualitative and quantitative approaches to text, combining the reliability of human validation with the scalability of algorithmic analysis.

The rest of this paper is organized in four parts. In the first section I address the motivation for developing another tool for differentiating language patterns by discussing the limits of inference of existing ones, and why social science researchers want to go beyond these limits. In the second section, I present my novel approach based on identifying phrase-level differences with sequence classifiers, and crucially how to interpret and validate the outputs of this approach. In the third section I present the results of the validation and the substantive interpretation of the outputs of my method. In the final section I discuss the future directions and limitations of this approach.

Motivation

Prior to presenting the approach, I elaborate the need for a novel method by considering the aims of automated text analysis in political science and the limits of existing tools. I argue that the utility of automated tools for text extends beyond problems of scale, and that these tools should be thought of as more than a means of *amplifying* human intuition. I then consider advantages of the reductive numerical approach to text used in political science text-as-data, but also establish the kinds of claims that we are unable to make within this paradigm.

Why Text, and Why Automate?

In their 2013 survey of the burgeoning political science text-as-data field, Grimmer and Stewart (2013) note the increased availability of large political corpora in electronic form as a driving factor behind the increased popularity of automated content analysis methods. This "big data" trend is the case for political campaigns as well. An indirect consequence of the major growth in internet-based political campaigning (Fowler et al. 2021) has been the creation of datasets several orders of magnitude larger than those previously available; four such examples are detailed in Table 2.1 below.[1]

Table 2.1: Large Political Advertising Corpora.

Dataset	N Docs
Google Transparency (Politics, US)	619K
Facebook Ad Library (Politics, US)	13.3M
Princeton Corpus of Emails[2]	435K
ProPublica	225K

The logistical challenges presented by these large corpora makes clear that one motivation for the development and application of automated content analysis methods is scalability. Automated tools make it possible for researchers with limited funding support to analyze large quantities of unstructured data. However, justifying automated content analysis methods in terms of feasibility and efficiency presents

[1]Note that such large numbers can be misleading in both directions; on the one hand they are underestimates of the total number of all political advertisements shown on these platforms (Edelson, Lauinger, and McCoy 2020) but on the other hand, most research questions are typically interested in a considerably smaller subset of advertisements, such as those run by candidate campaigns or PACs.

[2]See Mathur et al. (2020).

them as a second-rate option for researchers unable to hire and train large numbers of human coders or research assistants, which misses several important parts of the picture.

For one, political scientists study texts because many political phenomena are textual. In some cases, this is in the form of incidental trace data, such as the transcripts of political activities (e.g. Diermeier et al. 2012; Gentzkow, Shapiro, and Taddy 2019). Other political phenomena are inherently of text form, such as laws (Lapinski 2008) and court decisions (Clark and Lauderdale 2012), or the communications that construct a political environment such as legislator press releases (Grimmer 2013) and news articles (Barnes and Hicks 2018). Studying these phenomena requires a range of robust descriptive tools[3] specifically designed to deal with the complex way information is encoded into natural language.

While structured qualitative approaches to text can use human natural language understanding to trivialize the complexity issue at the phrase and document level, their challenge is aggregation and the identification or description of corpus-level patterns. One common approach, referred to as (Qualitative) Content Analysis (QCA), systematizes the human estimation of counts and proportions of concepts of interest within a corpus. This strategy can be applied at various levels. Examples of document level coding include Barabas and Jerit (2009) and Hayes and Lawless (2015), who respectively code news articles by event or tone. Classification can also be done at the quasi-sentence level (e.g. John and Jennings 2010), or flexibly

[3]The complexity of measurement and operationalization of text data has contributed to a focus on largely descriptive tools in the text-as-data field, although recent works study causal inference with texts (Egami et al. 2018; Fong and Grimmer 2019; Margaret E. Roberts, Stewart, and Nielsen 2020).

between levels as an information retrieval task for finding all texts and extracts relevant to a particular issue (Stanig 2015).

What do we Estimate?

Grimmer and Stewart (2013) name quantitative analogs for tasks done with QCA, and where the estimand is the same between qualitative and quantitative approaches to a question, text-as-data methods can be thought of as tools for "*amplifying* and *augmenting* careful reading" (Grimmer and Stewart 2013, 268). But as with non-text quantitative methods used in political science, quantitative text analysis developed for political science applications by political science methodologists primarily produce numerical or statistical summaries. Three such categories of approach include: scaling models, topic models and embeddings. Scaling models, such as Word Scores (Laver, Benoit, and Garry 2003; Lowe 2008), Wordfish (Slapin and Proksch 2008) or Wordshoal (Lauderdale and Herzog 2016) score documents onto a single dimension, which is then used to compare their positions. Topic models (Blei, Ng, and Jordan 2003) decompose a corpus into k distributions over words and documents, which are generally interpreted to correspond with separate semantic fields. These topics are then used to describe documents as a mixture of these fields, or to compare different documents within the same topic. Popular variants include seeded topic models (Eshima, Imai, and Sasaki 2020), which can be used for semi-supervised lexicon discovery and parameterised topic models (Margaret E. Roberts et al. 2014), used to model topic mixture as a function of speaker (and lexical) covariates. Finally, embedding models represent tokens in a corpus as the dense representation of the

conditional likelihood of a token given its context and other metadata in the training corpus. These vector representations are used to model words, documents and document covariates; Rodman (2019) describes the semantic shifts of key terms in newspaper corpora, whereas Rheault and Cochrane (2020) include speaker metadata to produce "party embeddings".

These approaches are all reductive in the sense that they produce a lower-dimensional summary of the source texts. They are also appropriate to their respective applications. If our estimand is ideology and we treat parliamentary speech (e.g. Peterson and Spirling 2018) or manifesto text (e.g. Slapin and Proksch 2008) as a noisy signal of ideology, then the appropriate estimator will be one that decomposes or otherwise filters the input. Likewise, if our goal is to describe the semantic shift of key terms (Garg et al. 2018; Rodman 2019) then distances in a semantic space will be an appropriate estimator, and if our goal is to discover and label the semantic fields expressed in a corpus (Blei, Ng, and Jordan 2003; Grimmer 2013; Margaret E. Roberts et al. 2014), then topic models will be well-studied estimators for this task. In the case of this article, where the objective is to select the linguistic features that differentiate and characterize political actors, past approaches have been based on lexical scoring methods, such as Monroe, Colaresi, and Quinn (2008) who compare four different approaches for identifying words that differentiate Democrats from Republicans in a congressional speech corpus.

The theoretical estimand in all of the above papers is semantic in nature, such as the ideology of a text (Laver, Benoit, and Garry 2003; Lowe 2008; Slapin and Proksch 2008; Lauderdale and Herzog 2016) or the words most closely related in *meaning* to a latent concept expressed by a target word or list of words (Rodman 2019; Es-

hima, Imai, and Sasaki 2020). The link between theoretical and empirical estimands (Lundberg, Johnson, and Stewart 2021) in these studies is made by leveraging the fact that individual words tend to be associated with meanings. Thus researchers combine the semantic values of words with numerical values provided by models to produce weighted or ranked word lists, and then use these weights and meanings infer the relative prevalence of semantic quantities in the data.

Many of the challenges and limitations of these methods–e.g. polysemy, ambiguity–result directly from an approach that begins with discarding word order and ends with describing the meaning of language in terms of the relative rates of usage of different words or phrases. These challenges are exacerbated by the fact that summary and reduction necessarily entail informational loss. Particularly when the representation of the results is in a different medium to the original data (i.e text to numeric), tasks such as sense disambiguation become difficult because of the disconnect between the summarized quantity and the original text.

It is not my intention to claim that inference leveraging numerical summaries for aggregation and lexical semantics for linking theory and empirics is useless or invalid; on the contrary, the careful application of these methods has produced many valid works in the study of important textual political data and phenomena. Rather, my point is that it is not always necessary to disconnect the source texts and numerical summaries, and opting not to do so has various benefits. Some questions require the contrasting of corpus-level patterns with document-level phenomena. In this article, the goal is to identify the characteristics of an advertisement that differentiate it as belonging to its campaign. This compares corpus- and document-level features. Identifying the linguistic features that typify a political advertising campaign re-

quires knowledge of the entire corpus, and identifying which features of a document make it (a)typical of that campaign requires application of those corpus-level patterns to every aspect of the document. In such cases, I argue that there is a need for annotative methods that link high-level summaries to individual observations.

Differentiating Language

In this section I present a new approach to differentiating the language use of political actors, based on the feature attribution scores of a DL sequence classification architecture. The presentation is divided into two portions; in the first, I discuss existing approaches to the task based on lexical scoring methods, and identify the gaps in our methodological toolkit detailed in the previous section. In the second, I discuss the challenges of applying DL in a social science context, and provide several justifications for its application in this case.

Motivating Sequence Models

Differentiating language usage between political campaigns is a broad task. One approach considers it a feature selection task, where the labels are the campaign, the data are the content of the advertisements, and the goal is to select the features of the data that are indicative of a label. I add the restriction that these features should be substantively interpretable, given that we eventually want to use them to characterize the respective campaigns.

Monroe, Colaresi, and Quinn (2008) approach this task by modelling lexical, i.e. word-usage, differences conditional on group. Their method treats words as the unit of analysis in language, and the semantic values associated with words facilitate the interpretation of resulting outputs. Monroe, Colaresi, and Quinn (2008) use additional contextual information about their corpus, such as the fact that the primary topic of debate in the particular session of Congress was regarding abortion, to validate their measure and make additional claims about speaker pragmatics, such as strategy, inferable from observed word frequencies. In this example, they find that Republicans are associated with the pseudowords `kill`, `babi`, `mother` and Democrats are associated with `woman`, `decis`, and `viabil`. We are able to infer the framing strategies (Chong and Druckman 2007) employed by politicians on respective sides of the aisle based on these single words both because we infer a wider context in they occur.

However as noted, there are limits to what we can infer about the style or strategy of a campaign based on the meanings of individual words sans context. Although in the above examples, the broader intent of a statement containing the respective words can be inferred because it is less likely that the word "kill" will occur in a pro-choice statement in the context of a debate on abortion, using the word "kill" in a sentence does not inherently make it a pro-abortion stance. The semantic context of a word in an isolation can be highly ambiguous even where the word is not polysemous. For example, it is hard to guess whether the tokens `doctor` or `procedur` should be associated with Republican or Democrat speech in an abortion context (as it turns out, `doctor` is Democratic and `procedur` is Republican).

As noted in the previous section, the challenge faced by text-as-data strategies link-

ing their theoretical and empirical estimand based on individual words and their inferred semantic context is that in most cases, individual words do not *produce* the phenomenon of interest, but only *correlate* with it. When what is being measured (the empirical estimand) is only linked to what is being studied (the theoretical estimand) by empirical regularity and not causality, it becomes important for the researchers to consider the conditions under which this regularity may not hold.[4]

One approach to directly circumvent this ambiguity is to recover the full phenomenon of interest (or if the phenomenon is latent, the full context that signals this phenomenon). This approach is challenging using current political science text-as-data methods for two reasons. The first reason relates to pre-processing: most text-as-data methods discard word order, representing documents as counts of tokens. Trying to recover span boundaries using such models is difficult, because the model does not treat language as an ordered sequence. The second reason relates to how we use the outputs of text-as-data models. Even if we are able to perfectly extract relevant spans from each text, these outputs cannot be aggregated in a straightforward way like a numerical summary. In order to interpret the output of a span-extracting model, we can either use a secondary model to aggregate features of interest from these spans. I do not opt for a strategy based on aggregating spans because they are likely to be even higher sparsity than n-gram data, meaning that the aggregation procedure will need to be highly reductive (such as sampling the most frequent words of these spans, which returns to the original challenge). Instead, I advocate an *annotative* approach, where the researcher highlights the relevant spans

[4]It is not my intention to say that researchers must therefore always prefer deductive strategies! On the contrary, the inferential leverage of an inductive approach provides many opportunities for progress in empirical social science research, and regardless a deductive approach to a process as complex as language will come with its own challenges.

within each document. The time cost of this approach is mitigated by the reduced time spent on validation, and can be further mitigated by randomly sampling a subset of texts to determine broader trends. I expand on this approach in a subsequent section.

Deep Learning and Interpretation

Because of the limitations in modelling spans introduced by the standard preprocessing steps detailed above, I turn to a computational language model that models words within sequences. Among these, I use several variants of the DL `BERT` architecture (Devlin et al. 2019)[5], which has become dominant in NLP for achieving state-of-the-art performance in a broad variety of tasks with minimal task-specific engineering.

As noted earlier, despite the increasingly widespread application of ML tools in quantitative social science methodology, there has been limited adoption of complex neural network-based DL tools, themselves an important part of ML, in our field.[6] In particular, whereas NLP has taken a strong turn towards DL in recent years, text-as-data tools developed in political science and adjacent disciplines applications remain primarily focused on non-DL methods. There are reasons for the lack of DL

[5] *Variants of BERT*: for the sake of brevity, in this paper I refer to all transformer-based encoders for language modelling tasks as `BERT`, after the most famous architecture in Devlin et al. (2019). However, the models employed in this paper are all variants of the original model, not limited to DistilBERT, `RoBERTa`, `MiniLM`, and the Reimers and Gurevych (2019) modifications of all these. For an extensive explanation of how `BERT` works, see Rush (2018).

[6] Recent exceptions include Lall and Robinson (2021) and Torres and Cantú (2022), who use DL for multiple imputation and images-as-data respectively. Additionally Rodriguez, Spirling, and Stewart (2020) include `BERT` as a benchmark to compare their method against.

applications in political science which need to be addressed in order to justify its use in this paper.

In brief, DL refers to a broad class of neural network models that pass data through *multiple* layers of nodes.[7] Each node (sometimes called an artificial neuron) takes in a set of inputs, calculates their weighted sum, and then passes this value onto the next layer of the network. Often, before passing on the data, an "activation function" is applied to constrain the value of the output. The model is trained by passing the data through the entire network, calculating the loss compared to the true outcome, and then adjusting the weights on the inputs at each node to minimize this error, iteratively. The loss function itself can be defined by the user. While each individual node is therefore relatively simple (essentially a generalized linear model), the ability to combine many of these units into complex configurations (called architectures) creates models that are both highly flexible and predictive.

However, in many of the typical applications for quantitative models in social science research, the ability to describe the relationship between the inputs and outputs is more important than the ability to accurately predict the output (Hofman, Sharma, and Watts 2017). Despite individual neurons containing weights analogous to regression coefficients, it is difficult to make inferences about the relationship between the inputs and outputs on the basis of a DL network. In part this is because it is unclear what the coefficients in intermediary layers in the network describe, and in part because DL architectures typically have an enormous number of parameters[8] that make systematic evaluation impractical.

[7]For recent and accessible introductions to neural networks and deep learning, I point the reader to Aggarwal et al. (2018) or Skansi (2018).

[8]Depending on the configuration, `BERT` has between 110 and 345 million parameters.

These challenges are not limited to the political science domain, and the need for explainability in various applications has motivated a "interpretable ML/DL" subfield (Linardatos, Papastefanopoulos, and Kotsiantis 2021 provide a recent review) focusing on the development of tools to interpret complex DL models. Of particular interest in this paper is a method for extracting the logic of DL text classifiers referred to as *rationale extraction* (Lei, Barzilay, and Jaakkola 2016).

This method uses the fact that when classifying a document, humans and algorithms typically base the label on a subset of the words in the text. This subset of the text that contains sufficient information to assign a label is referred to as a *rationale*. Various authors have developed axioms that rationales should fulfil (Jain et al. 2020). The first, called *faithfulness* (Lipton 2018), concerns the quality of the rationale extraction strategy. For a given classifier and rationale, a rationale is *faithful* if the extracted span "reflects the information actually used by said model" (Jain et al. 2020, 1). Two further criteria concern the quality of the rationale itself: it should be concise (*brevity*, Lei, Barzilay, and Jaakkola 2016) and contain all relevant information (*comprehensiveness*, Yu et al. 2019). A fourth condition concerns the relation between machine-generated rationales and human intuition; Wiegreffe and Pinter (2019) argue that rationales should make sense to human readers (*plausibility*). A further study establishes a benchmark for comparing machine and human-generated rationales (*ERASER*, DeYoung et al. 2020).

Although rationale extraction is designed as a tool for the interpretation and evaluation of complex DL language models, I argue that it can be used to characterize the differences in language usage between political campaigns. Given a classifier for political advertisements that predicts the campaign as the label, we want to

know the linguistic features in the advertisement that the model used to (correctly) label the advertisement. Provided that the extraction strategy is faithful, concise and comprehensive, rationales are the minimal span of text within a document that contains all of the information used by the model to make that prediction. When that prediction is correct, that span contains valid information for labelling the text. To the extent that the rationale is *plausible*, we can describe the relevant linguistic features that it captures.

Why ask `BERT`?

Not all differentiating regularities are relevant or useful. A critical reader may point out that we cannot specify the *kinds* of differences the classifier should use to differentiate the texts. While some differences will be reflective of the latent properties of the respective speakers that we want to capture, others may be the result of "artifacts" of the data generating process that is substantively uninformative. An example of such an artifact is clear for an image classification text: if all Warren advertisements have the same grey border along the bottom of their images, then a model will learn to differentiate on this basis (and possibly miss much subtler traits that provide a smaller marginal improvement to classification accuracy). A text analog for this kind of artifact could be a formatting by-line that ends each text. While these kinds of artifacts can be caught and removed with careful pre-processing of the data, the fact remains that there may be less obvious latent artifacts that are nevertheless difficult to interpret because they do not inform the researcher of a useful difference.

There are several encouraging reasons to believe *a priori* that this will not be the kind of difference that BERT models identify, relating to the BERT architecture and how it is used. Typically, researchers will download a BERT model that has been trained on large corpora and stored this information in the weights of the model. The pre-trained model is then "fine-tuned" on a smaller task-specific corpus. In most cases, a BERT model that has been pre-trained then fine-tuned will outperform a BERT model that has only seen the task-specific corpus. This technique of using non-task-specific information to improve performance on other tasks is referred to as *transfer learning*, and has been crucial in enabling the application of neural network architectures requiring enormous training data to data-scarce applications (Rogers, Kovaleva, and Rumshisky 2020). This strategy appears to work by encoding linguistic regularities in the weights of the model. Although researchers have found that BERT does not interpret language identically to humans, they have found that it has a hierarchical notion of syntax (Lin, Tan, and Frank 2019) and is able to infer semantic relations (Ettinger 2020). As such we have evidence that BERT is able to identify the kinds of relations that we are interested in, although we are unsure whether this will hold across all applications.

Nevertheless, even if BERT does not differentiate campaigns using the same linguistic patterns as a human coder would, there are reasons to at least study the patterns that it does identify. For one, classifiers using BERT-like models have achieved superhuman accuracy on a variety of natural language benchmarks (Linardatos, Papastefanopoulos, and Kotsiantis 2021; Storks and Chai 2021). This means that researchers have found that for tasks where there is an objective truth (such as the task in this paper), BERT is able to classify correctly at a higher rate than human

coders. Moreover, provided that the rationale extraction strategy is faithful, then the rationale text is in fact something that differentiates the classes. This ability to identify non-obvious differences is likely to be especially useful in cases where the differences are likely to be subtle, such as the comparison between Sanders and Warren. Despite all this, the possibility remains that differences are substantively unhelpful "artifacts" of the data-generating process. This weakness to artifacts is not unique to this approach, and in all cases reinforces the need for careful human validation of model outputs. I set out these validations in the following section.

Method and Validation Strategy

In this section I explain the implementation of the rationale-based strategy for identifying differentiating patterns of language use, and discuss the ways in which the results of this model may be misleading. To address these challenges, I outline the three forms of validation to mitigate these challenges: statistical, comparative and substantive.

Method: Rationale Extraction and Integrated Gradients

Rationale extraction consists of two components: a feature attribution strategy that scores the input features in terms of their "contribution" to a prediction, and an aggregation strategy for generating rationales from the scored input features. In this paper, I use a feature attribution method known as Integrated Gradients

(IG; Sundararajan, Taly, and Yan 2017), shown to provide faithful and consistent explanations for text classifications (Atanasova et al. 2020).

IG is a feature attribution method designed for neural network models that given an input and model prediction, gives the per-feature importance relative to a user-defined baseline input. In the case of image models (one of the three example applications in Sundararajan, Taly, and Yan 2017), this baseline is typically a black image (of all zero inputs), whereas for `BERT` models we use an input with all tokens replaced with the special `<pad>` token used to denote an empty input.

The "explanation" provided by IG scores is analogous to understanding the relationship between the inputs to and prediction of a linear model in terms of the partial derivative of the prediction with respect to each input.[9] In less technical terms, for a given IG score on an input feature, strongly positive values indicate that the input "pushes" the prediction more towards the target class, near-zero values indicate that the token provides little information, and strongly negative values indicate that the token reduce the likelihood of the target class. In the binary classification case, these values are symmetric between target classes, so negative values can be interpreted as being indicative of the negative class. In the multiclass case, positive scores have the same interpretation, but negative scores should be interpreted as being indicative of "not-" the target class.

For the aggregation step, I compare three strategies for extracting rationales from documents based on IG scores. The first two, used in Jain et al. (2020), are selecting the **top-k** tokens by attribution score and the **contiguous** length-k region

[9] I explain the derivation of and intuition behind Integrated Gradients and how they relate to OLS models in detail in the appendix.

with the highest summed attribution, where k is one-tenth the length of the document. The third strategy is to use the contiguous **positive** region with highest average attribution score. This has the advantage of being non-parametric, order-and-context-preserving and excluding any negative regions, but has the disadvantage of potentially returning no tokens (if all attribution scores are zero or negative).

Validation 1: Comparative

The first form of validation I provide is comparative. This validation is motivated by the intuition that to the extent that different models treat the data differently but arrive at similar results, the researcher can be certain that the patterns they are observing are inherent to the data and not an artifact of the method they have chosen.

For this, I evaluate IG as a lexical scoring method by comparing its attribution scores to the lexical scoring methods presented in Monroe, Colaresi, and Quinn (2008). I briefly explain the two methods used for lexical scoring in this paper. The first, log-odds (LO), provides a symmetric measure of the likelihood of a token given the class of the speaker, centered at zero when the likelihood of use is equal between classes. Although I prevent tokens from having an infinite or undefined log-odds ratio by adding $1e-2$ to the odds, in presenting my results I only include the union of the vocabularies; i.e. tokens occurring in both classes (one or rest for the multiclass case). This is because including tokens only occurring in one class gives unstable results.

The second method, which I refer to as Fightin' Words (FW), is a Bayesian model for identifying distinguishing word usage presented in Monroe, Colaresi, and Quinn (2008). It builds on a simple comparison of usage rates (via LO, above) to include tokens that only occur in one set of documents by introducing a smoothing Dirichlet prior over all tokens. The z-scores indicate the number of standard deviations a word usage is away from the mean of no difference in usage between groups.

To use IG scores for lexical scoring, I normalize the IG scores within documents so that they add up to one, and calculate the average normalized score for each token lemma. Although this loses the interpretation associated with the zero threshold, we can infer that tokens with high average normalized IG for a given candidate or campaign are strongly indicative of that campaign in all instances where the token occurs.

These IG scores are compared with the Monroe, Colaresi, and Quinn (2008) scores both quantitatively for inter-correlation and qualitatively for *prima facie* informativeness. Although strong correlation and overlap would be indicative that IG is as valid as the established approaches in Monroe, Colaresi, and Quinn (2008), diverging results are more difficult to interpret. In one sense, we want a method that provides *additional* insights that existing methods cannot find; in another sense, diverging results require additional validation to demonstrate that the new results are also informative.

Validation 2: Statistical

Evidence from Adebayo et al. (2018) and Atanasova et al. (2020) indicates that evidence provided by attribution on DL models can be unstable, meaning a measure of uncertainty is essential. In particular, given that a IG score of zero has a clear substantive interpretation–the input does not contribute to the prediction–the ability to conduct hypothesis tests on these values is critical to the application of this method in a quantitative social science context.

However, although the training of neural networks can be probabilistic, the predictions of these models and the attribution scores are deterministic. A clever solution in the DL literature relies on the fact that many DL models include dropout layers. A dropout layer is an operation on a tensor that returns the input with some random subset of elements set to zero, with the probability of zeroing typically[10] uniform over the inputs with the probability set parametrically. Dropout layers are a proven tool for reducing overfitting (Srivastava et al. 2014), but Gal and Ghahramani (2016) shows that they can be used as a computationally efficient bootstrap for neural networks. This method is referred to as Monte Carlo Dropout (MC Dropout). I use MC Dropout to draw 30 estimates of the predicted class and attribution score. I use these 30 draws to generate bootstrapped confidence intervals for hypothesis testing, where I test whether the attribution score of a given feature is significantly different from zero.

[10]Typically, but not necessarily, e.g. Li, Gong, and Yang (2016) propose multinomial dropout.

Validation 3: Substantive

Thus we need a validation does not require comparison to existing methods. For this, I adapt the FRESH strategy described in Jain et al. (2020). This strategy consists of three models.

The first model is trained on the full training dataset (with a validation split left out to assess training progress). The weights and predictions of this model are used to generate IG scores, which are then used to construct rationales from the texts in the three ways described above (contiguous, positive and top-k). Then for each rationale strategy, two new datasets are constructed. One is a dataset of only rationales, and the other is the texts minus the portion assigned to the rationale (*text-sans-rationale*). For each of these datasets, we train a completely new classifier: one that only sees rationales, and the other that only sees *text-sans-rationale*. We then compare the performance of the new models.

Table 2.2: Substantive validation summarized.

Model Trained On	Model Accuracy	Strategy Precision	Strategy Recall
rationale	high	high	–
text-sans-rationale	low	–	high

Table 2.2 summarizes the logic of this strategy. If the accuracy of the classifier does not drop significantly compared to the model trained on the original dataset (we may expect some drop because the total entropy of the data will significantly decrease when we only retain one-tenth of each text), then we can infer that the rationales

do contain an informative signal of the class (i.e. the attribution method has high precision). If the model trained on *text-sans-rationale* has lower accuracy than the *rationale* model, then we can infer that the rationale extraction strategy is not missing information for predictions in the remaining data (i.e. the attribution and rationale methods have high recall). Together, this test can show that the attribution method must be detecting information that is actually relevant to differentiating the classes in the data.

Data and Task

I use ProPublica's publicly available collection of political Facebook advertisements from 2016 to 2020. These advertisements were collected using a browser plugin installed by volunteers. Whenever a participating individual was served an advertisement on Facebook, the advertisement was scraped and sent to servers hosted by ProPublica, who then used a mixture of human and automated classification to determine the advertisement to be political or not. The raw content plus some pre-parsed information of the advertisements classified as political is published on their website as a continuously updating dataset.

Note that the distribution of unique advertisements in this dataset is heavily skewed towards liberal/Democrat audiences. I presume that this is because recruitment of volunteers was more successful among liberal/Democratic individuals than conservatives/Republicans.[11] Given that I have a large number of advertisements for

[11] At the time of writing, the accompanying documentation to the dataset does not address this imbalance.

Democratic candidates at all levels, from gubernatorial and state legislature to presidential level, and I have close to no advertisements for republican candidates other than Donald Trump, I do not try to make inferences about campaigns other than the largest one, the Trump campaign. Although there are a number of advertisements likely tailored for conservatives in the dataset, the imbalance in coverage indicates to me that there is a large amount of missing data for smaller races.

I filter the dataset down to twelve campaigns, and remove near-duplicate advertisements using a Levenshtein similarity threshold of 0.95 to avoid biasing towards advertisements that had a high degree of A/B testing. From these twelve campaigns I make three comparisons. The first comparison is Donald Trump and Joe Biden (527 advertisements). This comparison is made both because they were the eventual presidential nominees of their respective parties, but is presented first because we have the strongest preconceptions about what differences we should find between Trump and Biden, especially given the idiosyncratic style of Trump. Thus differences we find between the campaigns are less likely to be surprising, and more likely to be confirmatory.

The second comparison is Bernie Sanders and Elizabeth Warren (829 advertisements). This comparison is made because they represent the two most left-wing primary candidates in the Democratic primary, and therefore we have fewer expectations about the differences we expect to find in their campaign advertisements. The differences identified in this comparison therefore shows the utility of the applied methods to identify subtler differences when the expected similarity is high.

The final comparison is between Amy Klobuchar, Bernie Sanders, Beto O'Rourke,

Cory Booker, Elizabeth Warren, Joe Biden, Kamala Harris, Kirsten Gillibrand, Micheal Bloomberg, Pete Buttigieg and Tom Steyer (3165 ads). This comparison shows the utility of these methods beyond the binary case, which is particularly relevant for the applicability of these methods in a multiparty setting. It also reveals information about the segments of the Democrat primary electorate that the respective primary candidates view as their target constituency. In order to adapt LO and FW to the multiclass case I employ a one-vs-rest strategy.

Pre- and Post-Processing

Pre-processing steps in automated text analysis broadly encompass the initial operationalisation of the text in a numerical (and usually tabular) format prior to its use in a model. Denny and Spirling (2018) show that steps taken at this stage typically have non-trivial impacts on the final output of the model. In order to simplify comparisons, I unify pre-processing steps between the three models where possible. I discuss two key parts of the pre-processing: text cleaning and tokenization.

Text cleaning helps to remove artifacts from the original data collection process and reduces noise, but also risks introducing systematic bias by omitting particular features. For all three models I simply removed all HTML tags, and otherwise kept the text intact (leaving artifacts such as consecutive blank spaces and so on). A custom stopword list was created to remove non-words and other parsing errors.

Tokenization refers to the conversion of sequence of characters in a text to a sequence of "tokens", the base input for all models considered. Relevant decisions in-

clude conversion to lowercase, stemming/lemmatization, and even decision of what constitutes a token (i.e. determining token boundaries). For all models, I did *not* convert to lowercase because capitalization it conveys relevant information about tone, emphasis and style to the reader. Instead of using a blunt stemming method, I used the lemmatization algorithm provided by the SpaCy library. This reduces homonymy and leaves tokens more interpretable.

For the LO and FW approaches, I also used SpaCy's language model to determine token boundaries. This has the advantage over splitting on whitespace of handling contractions as two words, and then reduce them to their respective lemmas. For IG, because the classification model is based on `RoBERTa` (Liu et al. 2019), tokenization is done with a pre-trained byte-pair encoding (BPE) algorithm. I also experiment with `MiniLM` (Wang et al. 2020) and `MPNet` (Song et al. 2020), and train my own model and tokenizer on the full ProPublica dataset based on the `DistilRoBERTa` (Sanh et al. 2019) architecture, but find that the pre-trained `DistilRoBERTa` model performs the best.

In order to make direct comparisons between the three methods, in post-processing I match the token scores produced by the pre-trained BPE to match the boundaries of SpaCy tokens. For the most part, because BPE produces subword tokens, this is a simple case of concatenating tokens where a word is split into several word-parts, but in the rare reverse of a single BPE token corresponding to multiple SpaCy tokens (such as an apostrophe in the middle of a word, which is represented by two tokens in `RoBERTa`), I transfer the score to the first token and set the score of the latter as NaN. There were no cases of split boundaries between the two tokenization schemes.

With all comparisons, there is the decision of how to deal with tokens that occur in documents belonging to one class and not the others. This is a particular issue for LO, as the log-odds ratio is undefined/infinite. This is typically remedied by adding noise to the odds ratios (Monroe, Colaresi, and Quinn 2008 add 0.5 to counts). This still results in very large log-odds scores for tokens that only occur in one class, meaning that word lists of the top words from each class will simply capture this.

Monroe, Colaresi, and Quinn (2008) argue that this is a limitation, and account for this in the FW approach by giving a Dirichlet prior for the baseline probability of tokens occurring in a document. For transformer (and other DL) models, the baseline probability of a token is learned from the pre-training corpus, with no word containing only characters in the tokenizer vocabulary having a uninitialized corresponding gradient (although obviously, combinations of tokens may not have been encountered during the pre-training).

During fine-tuning, we should expect transformer models to converge towards predicting documents on the basis of individual tokens that perfectly predict class in the training data. In order to mitigate this, I use both a low learning rate ($1e-5$ to $5e-5$) and dropout layers both in the embedding block at the beginning of the model and in the classification block at the end of the model, to penalise localized learning.

Results

I divide the results into two parts. The first concerns the outcome of the three validation exercises. For the comparative validation, I assess whether IG performs similarly to LO and FW as a lexical scoring method. I find that whereas LO and FW provided highly correlated rankings (both by Pearson and Spearman correlations), IG is relatively only weakly correlated with either. Nevertheless, by inspecting the words that each ranks highly, I see that IG does not provide obviously different substantive results.

For the statistical validation, I present the application of bootstrapped confidence intervals for the integrated gradient scores for a specific advertisement and discuss the substantive implications. Although there is some variation, salient tokens within rationales are all salient for this strategy. For the substantive validation, I compare the accuracies of the *rationales-only* and *text-sans-rationales* model and find evidence that is consistent with the IG strategy identifying portions of the text that are actually differentiating.

In the second part, I use the IG scores to annotate the features of an advertisement that differentiate it from other campaigns, and characteristic of its own campaign. I highlight particular patterns that come of changing the reference category; Biden versus Trump against Biden versus all Democratic primary candidates, and Sanders versus Warren against Sanders versus all Democratic primary candidates.

Comparative Validation

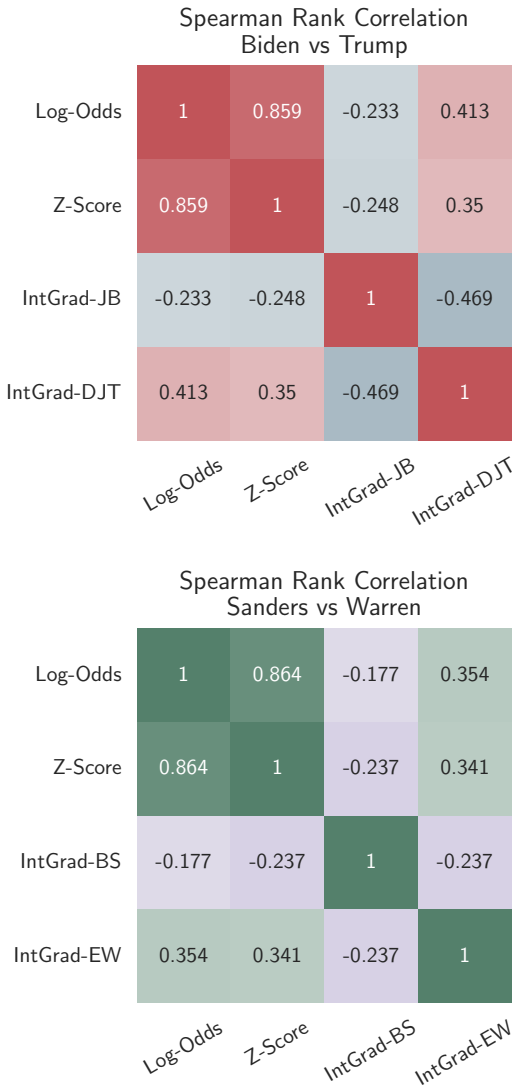

Figure 2.1: Spearman rank correlation matrices for LO, FW and IG scores.

I first compare the Spearman rank correlation of the three measures. Figure 2.1 shows the correlation matrix for the Biden vs Trump and Sanders vs Warren comparisons; the correlation matrix of all primary candidates is in the appendix. In

each matrix, there are four columns/rows, corresponding to the ranking of tokens by Log-Odds Ratio, FW Z-Score, and the non-normalized feature attribution scores for each candidate (note that the rankings that these will not be symmetrical, as each is calculated only within the speaker's own documents).

Several patterns stand out. For both pairs, the rankings provided by LO and FW are highly correlated (approximately 0.86). For Biden and Sanders, there is a negative correlation due to the alignment of the LO and FW rankings treating Trump/Warren as the "positive" pole. In general IG is weakly correlated with LO and FW, but there is no fixed pattern to this weak correlation.

How do these rankings compare at the top end of the distribution? I present the top ten tokens by candidate for each comparison in Figure 2.2 In contrast to the visualizations of Monroe, Colaresi, and Quinn (2008), I plot the information as annotated heat maps to convey the information more concisely. The figure consists of subplots, each containing the top 10 ranked tokens by candidate for each comparison, colored by score. The top row (orange) shows the absolute log-odds ratio. The middle row shows the absolute z-score from Monroe, Colaresi, and Quinn (2008). The bottom row shows the average feature attribution. The scale for all three sets of tables is shown on the right-hand-side of the figure. For LO and IG, only tokens occurring in at least two classes are included.

Although the extent of characterization and inference that can be made from a small subset of words is limited, interesting comparisons can be made in multiple directions. One is between models, seeing the different tokens that the respective

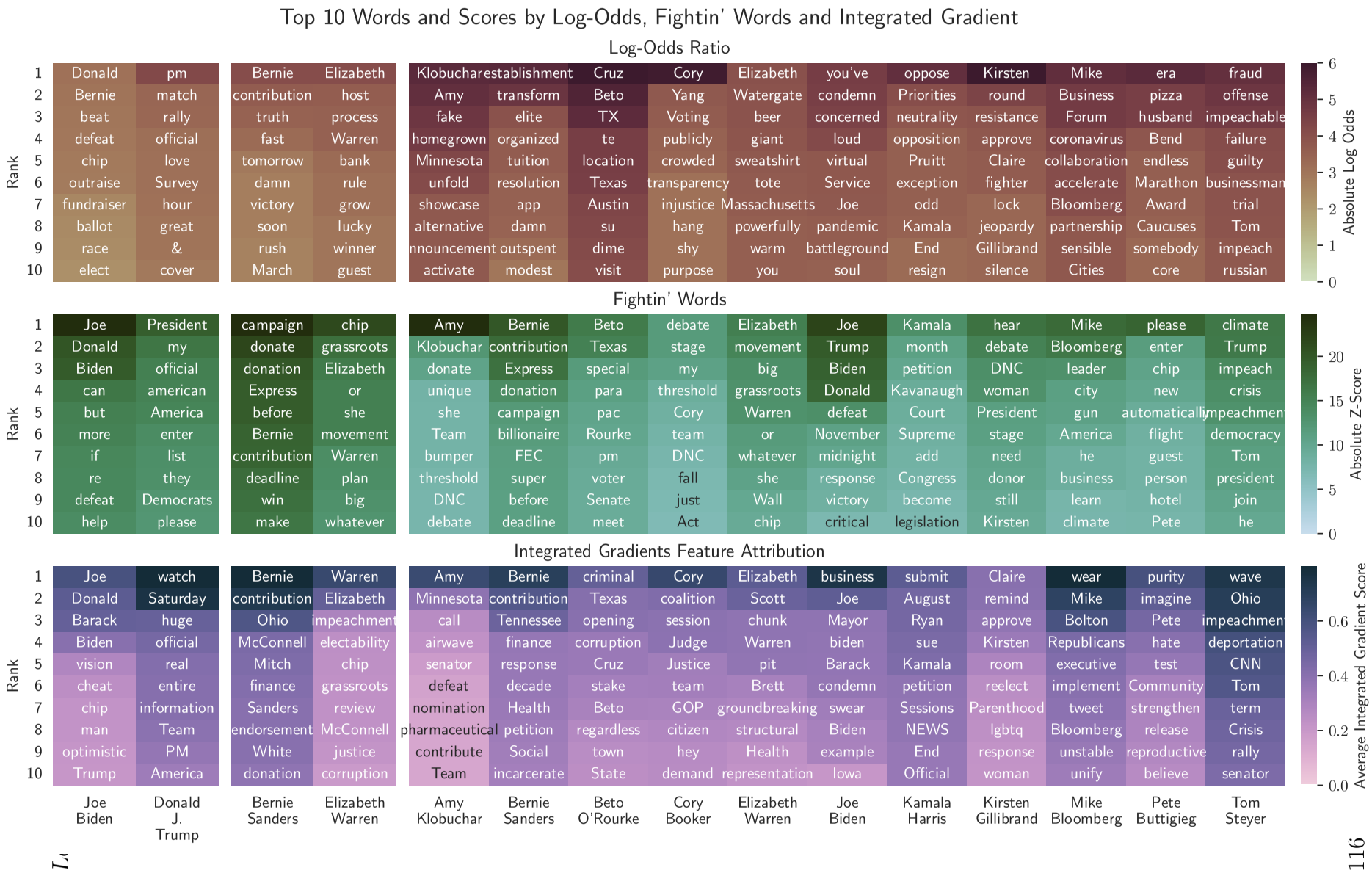

Figure 2.2: Top 10 tokens by candidate by comparison for Log-Odds, Fightin' Words and Integrated Gradient.

models emphasise. The other is within the same model, comparing how changing the reference category affects the top 10 tokens of a campaign.

Each model emphasizes different aspects. LO is a simple, normalized and symmetric ranking of tokens occurring very frequently in one class but not the other. We can see that `Donald` is the token with the highest LO ratio for Joe Biden when comparing Biden to Trump, followed by `Bernie`, `beat`, `defeat` and so on. In contrast, the top LO tokens for Trump when compared to Biden are `pm`, `match`, `rally`, `official` and `love`. We might infer from this that compared to Trump's advertisements, many of Biden's advertisements are about defeating his opponents, whereas Trump advertises events/deadlines (such as rallies, funding deadlines etc.).

In contrast, FW can be interpreted as the increased likelihood of using a particular term given speaker class, relative to a baseline probability of using any word. This hints at a similar interpretation for Trump, highlighting usage of the tokens `my`, `official`, `american`, but shows a semantically less interpretable set of tokens for Biden: `can`, `but`, `more`, `if`. A possible interpretation here is that Trump advertisements are more likely to use declarative statements, in which case qualifying verbs and hypothetical conjunctions are less likely to be used relative to Biden advertisements.

IG emphasizes a similar pattern overall with more resemblance to LO in interpretation. Top Biden tokens include names (`Joe`, `Donald`, `Barack`, `Biden`) whereas Trump tokens include language advertising events or other activities (`watch`, `Saturday`, `PM`, `information`). Across all three rankings, Trump advertisements are associated

with adjectives pertaining to a semantic field of legitimacy and grandeur (`official`, `great`, `huge`, `real`), which appears to reflect Trump's style.

Comparing Sanders and Warren, Sanders' advertisements are more associated with campaigning and fundraising language (`contribution`, `victory`, `donation`, `donate deadline`, `finance`, `endorsement`) and Warren's advertisements are more associated with specific issues and a grassroots movement (`grassroots`, `movement`, `justice`, `corruption`, `impeachment`). This pattern holds less when the comparison includes all other democratic primary candidates. Sanders' advertisements are then more associated with his signature issues and style, with tokens such as (`establishment`, `transform`, `elite`, `tuition`, `damn`, `billionaire`, `Health`), while tokens relating to fundraising rank somewhat higher again for Warren (`beer`, `tote`).

The extent of inference that can be made from these tokens without the context in which they occur is limited. Nevertheless, it appears that all three measures highlight *prima facie* sensible characterizations of each campaign. It is also clear that from these maps that what "characterizes" a campaign very much depends on what that comparison is being made against.

Attribution and Rationales

Although feature attribution based on IG has provided plausible lexical rankings in this instance, what can we say about their performance more generally? In this section I discuss the stability, reliability and interpretation of IG in context.

Using MC Dropout, for each document I sample 30 draws from the posterior feature attribution distribution. Figure 2.3 compares these distributions for a Biden advertisement for the two different classifiers. The blue circles show the per-token attribution from the classifier differentiating Trump and Biden, and the orange diamonds show the same for the Democratic primary. Each point shows the estimated attribution score with bootstrapped ($n = 1000$) 95 percent confidence intervals for each token (small horizontal lines).

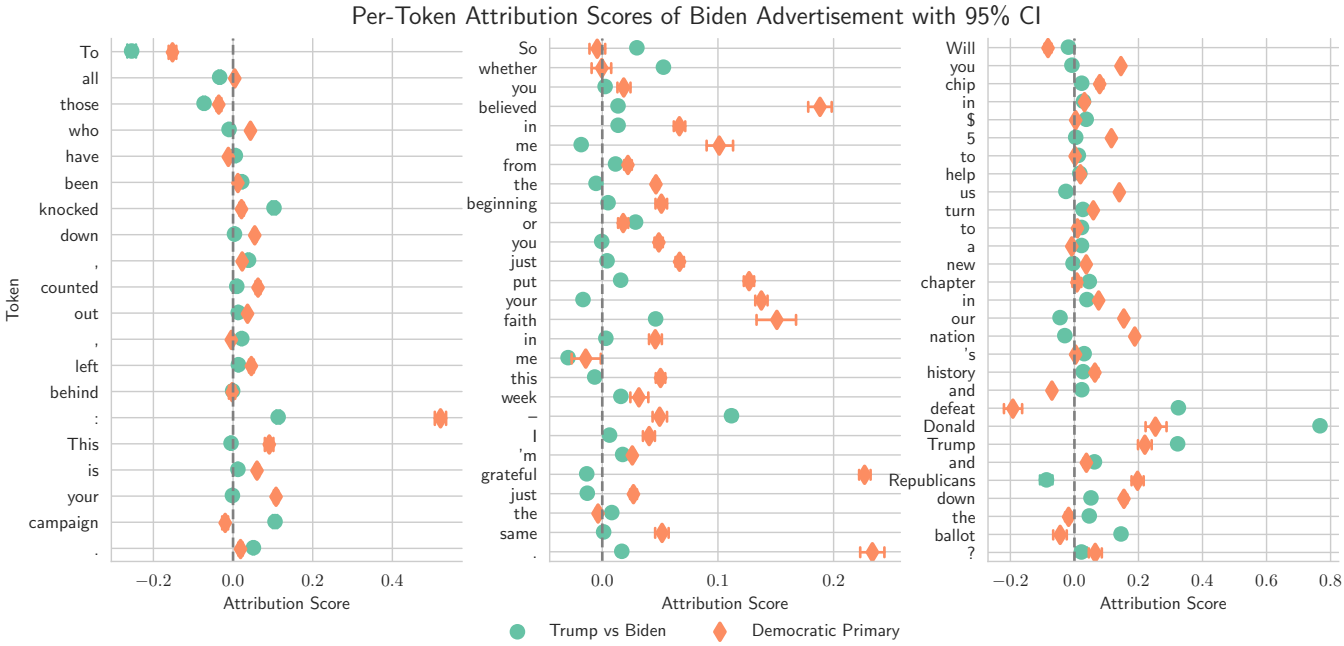

Figure 2.3: Per-Token Feature Attribution (bootstrapped 95% CI). *Note changing x-axis scale.*

The confidence intervals indicate that the uncertainty due to the stability of the attribution strategies does for all but the lowest salience tokens is not significantly different from zero. Combined with rationale strategies for selecting only the highest-salience regions, this instability does not change the interpretation of which regions

are salient. To account for marginal differences in salience, I use the average score from 30 draws to determine the rationale extraction. But are these high-attribution phrases actually sufficient to differentiate Biden advertisements from other campaigns?

I adapt the FRESH validation strategy of Jain et al. (2020) to test whether this is the case. Remember that for an explanation to be *faithful* it must actually reflect the information used by the model to come to its decision (Lipton 2018). This is equivalent to our application, where we are interested in the "logic" of the classifier because we want to know what features it uses to differentiate speakers. Therefore we need to confirm that our feature selection strategy highlights inputs that the model actually determines to be differentiating.

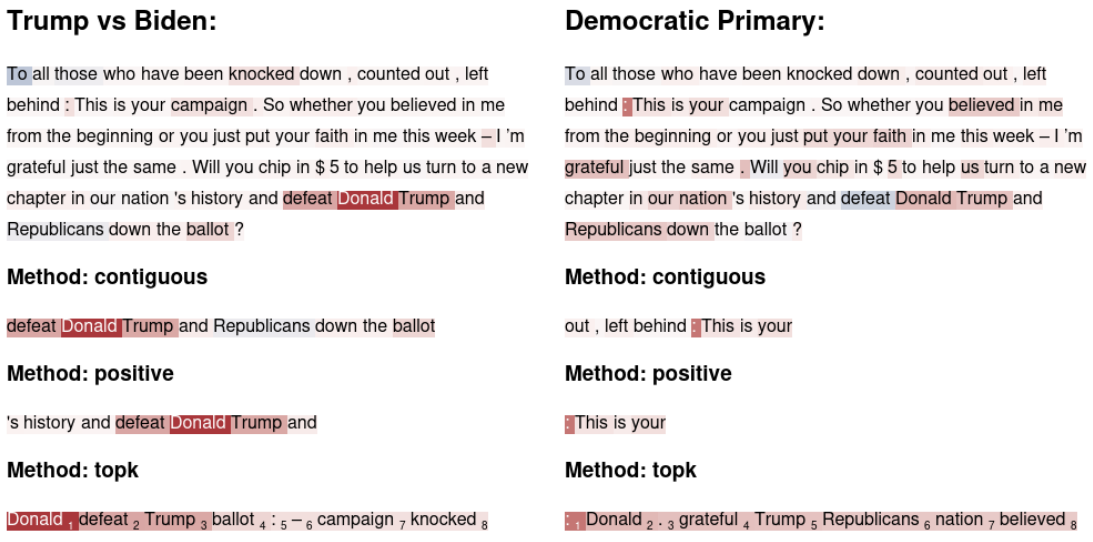

Figure 2.4: Rationale Generation.

Figure 2.4 shows the text of the same Biden advertisement, shaded by mean attribution score over 30 draws for each model. Darker shades indicate a stronger score,

and red indicates the positive class (Biden) whereas blue indicates the negative class (not-Biden). These attribution heat maps help visualize the three rationale extraction strategies. Two are from Jain et al. (2020) (**contiguous**, **topk**), and one is an original approach (**positive**). The contiguous and topk strategies extract fixed proportions of the document as a rationale; contiguous selects the length k region with the highest sum attribution score, whereas topk chooses k tokens with the highest scores. Because my aim is to flexibly identify variable-width spans for characterization, **positive** chooses the contiguously positive segment with the highest average attribution score.[12]

The length of the rationale is an informational trade-off. Shorter rationales are more parsimonious, but are more likely to exclude informative spans. For the **topk** and **contiguous**, I set rationale length to one-tenth the length of the document, as in Jain et al. (2020). For **positive**, the average rationale length is 19.3 percent of the document. Figure 2.5 shows the distribution of rationale size, broken down by comparison group. Rationales for the Democratic Primary were on average the smallest proportion of the text, but the distribution also has the longest tail and smallest inter-quartile range.

Jain et al. (2020) state that to fulfil the *faithfulness* criterion, models trained on the rationales should also be able to accurately predict labels. Intuitively, this shows that extracted rationales contain enough signal that a model that only sees the extracted text would be able to infer the correct label, without leveraging any

[12]Jain et al. (2020) also train a model (`BERT`) to predict rationales, and then use this to generate new rationales. I omit this approach because of the very low accuracy achieved by the model during training.

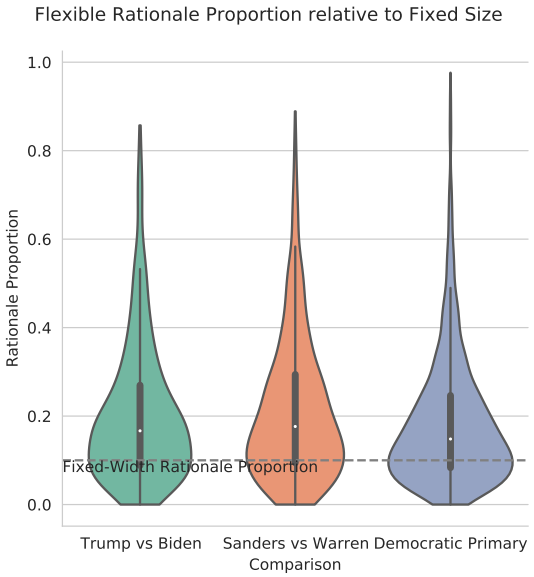

Figure 2.5: Distribution of Rationale Sizes per Comparison Group.

other information available to the original model. In our case, this would show that highlighted portions of the text are relevant for differentiating the speaker.

To measure the "recall" of a rationale extraction strategy, I train a model on a dataset that contains all text except the rationales. To the extent that the accuracy of the model trained on rationales is higher than the model trained on the *text-minus-rationale*, we can be sure that the strategy has extracted a greater proportion of the identifying information. Note that we should not expect the *text-minus-rationale* model to have an accuracy of zero with fixed-length rationales; there is no way to know *a priori* the proportion of tokens in an advertisement that are indicative of its speaker.

Figure 2.6 shows accuracy of the classifiers trained on rationales or text-minus-rationale for each of the three strategies, and compares them to the accuracy of the base model with the full texts. The interpretation for all models save the model

trained on the **topk** rationales have a lower accuracy than the base model is due to the smaller amount of information left in a rationale versus the full text. In comparison, consider the difficulty a trained human coder would face labelling the rationales in Figure 2.4. Although `defeat Donald Trump` is sufficient to distinguish Trump and Biden, most coders would be hard-pressed to label the Democratic primary rationales: `out, left behind: This is your`, `:This is your` and `: Donald . grateful Trump Republicans nation believed`. Furthermore, the drop in accuracy for all strategies shows that excluding these high-attribution score tokens greatly increases the difficulty of correctly classifying the advertisement.

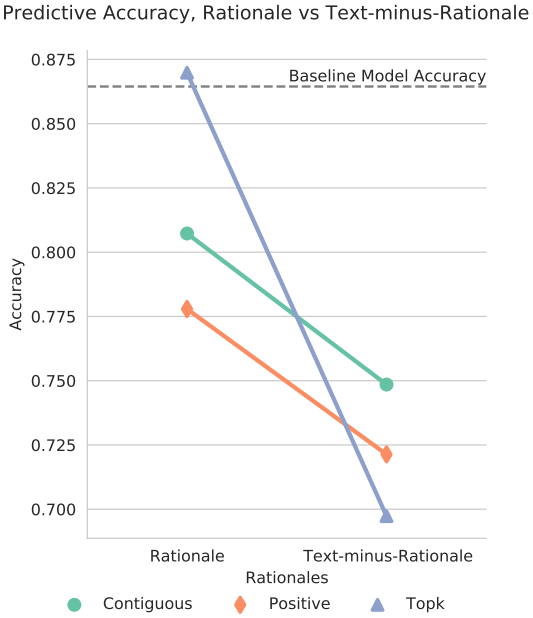

Figure 2.6: Accuracy of Models trained on Rationale versus Text excluding Rationale.

Interpreting Annotation

Having shown that IG functions both as a lexical selection strategy and a method for explicating the patterns identified by transformer-based classifiers, I demonstrate one further utility of this tool for our question. Annotating documents with token-level attribution scores is a helpful aid for qualitative text analysis strategies such as content analysis, by incorporating information on corpus-level patterns into individual documents.

I first discuss the Biden advertisements shown in Figure 2.3 and 2.4. Firstly, there appears to be a mixture of "high", "medium" and "low" attribution regions. For the Trump versus Biden classifier, there is a single phrase with a high attribution score: `defeat Donald Trump` in the third sentence. In contrast, the Democratic primary classifier assigns high scores to `Donald Trump` but a negative score to `defeat`, and its highest score is on the punctuation mark `:` in the first sentence. The Democratic primary likewise assigns significant scores to several ranges of the text that the Trump versus Biden model does not: `you believed in me`, `you just put your faith in`, `in our nation`.

Consider the phrase `defeat Donald Trump` in the third sentence (right-hand panel, Figure 2.3). The substantive interpretation of the attribution scores is that while the phrase `Donald Trump` is indicative of a Biden advertisement, the token `defeat` is only associated with Biden when compared to Trump advertisements, and indicative of non-Biden advertisements when comparing to other Democratic candidates. Additionally, the token `Donald` is a strong indicator of a Biden advertisement when

compared to Trump, whereas this is not the case when compared to other Democratic candidates.

It is unsurprising that the phrase `defeat Donald Trump` indicates that it is the advertisement from Donald Trump's opponent. Comparing Biden to the other democratic primary candidates, we see instead the phrases `: This is your campaign`, `put your faith`, `I'm grateful`, `our nation` and `Donald Trump and Republicans` highlighted. The former four phrases point to a rhetorical style possibly pertaining to a semantic field of faith-based morality (collectivism, faith, humility) being employed by Biden's campaign, and the latter may indicate that most Biden advertisements make reference to the opponents to be defeated.

I apply a similar analysis to the Sanders and Warren campaigns. Figure 2.7 contains the same attribution heat maps as Figure 2.4, but compares three advertisements each from the Sanders and Warren campaigns. The left column shows texts highlighted by attribution score from a classifier trained to differentiate Sanders and Warren, while the right column shows the same for the eleven Democratic Primary candidates.

In the first Sanders advertisement, the phrases `super PAC` and `wealthy` are salient when comparing to all primary candidates, but not when comparing only to Warren. This suggests that within the eleven Democratic primary candidates, Sanders and Warren both campaign on platforms of opposing wealth inequality, whereas other primary candidates do not make it their signature issue.

In the second Sanders advertisement, several tokens reverse attribution, including `Justice`. This suggests that the theme of "Justice" is more associated with other

Figure 2.7: Annotated Sanders and Warren advertisements. Red highlighting indicates positive IG score and blue indicates negative. Darker scores indicate larger absolute values.

primary candidates than either Sanders or Warren. Finally, the often referenced phrase `I am once again asking for your financial support` is scored higher when comparing to the primary candidates, whereas `donation` is no longer salient. This indicates that in terms of fundraising language, `donation` is a strong predictor of Sanders when compared to Warren, but this is no longer the case when comparing to primary candidates.

Similar inferences can be made looking at the three Warren advertisements. The first pair tells us that the words `wealthy` and `chip (in)` are characteristic of Warren fundraising when comparing to the larger group of eleven primary candidates, but not when only comparing to Sanders. The second pair shows a stronger focus onto the phrase `grassroots movement`, and a flip on the phrase `Sign our petition`. The third shows that a similar pattern for the phrase `structural change`, as characterizing of Warren compared to the other ten primary candidates.

Two patterns concerns substantive regions. One is the salience on `Tax/tax` in the second advertisement: when compared with all Democratic candidates, it is one of the few salient tokens, whereas when concerned with Sanders it loses salience. This suggests that this call for taxation, which characterizes Warren with respect to the broader set of Democratic primary candidates, does not differentiate her from Sanders. Of equal interest are the regions that are *not* highlighted. Some of the regions corresponding to more substantive claims: `unrig the system` (second advertisement) `I will not only beat Trump - I'll start making real change` (third advertisement) do not distinguish her from other Democratic primary candidates.

There are limitations to what we can learn from this strategy. Although we have shown IG attribution is a faithful way of identifying *which* information that models use to classify texts, and that this information is provably differentiating between classes, it does not tell us *why* a given token or span is salient. All of the above generalizations combined the evidence from the model with domain-specific knowledge about the election and the campaigns. However, this method is not meant to eliminate the need for domain-specific knowledge, but rather to direct the attention of a researcher to which pieces of information are salient *in context*. As with the other text-as-data tools, that salience must still be linked to the theoretical estimand through argumentation, but the advantage of this method is that the link between the empirical estimand (salient phrases) and the source data is more direct.

Conclusion

The application of text-as-data within computational social science is shaped by the methods and tools available. Where estimators have been based on various simplifying assumptions about language, researchers accordingly restrict their claims to what the model can justify. When representing a corpus as a matrix of word counts, the inferences that can be made from models built on this operationalization of text are based on observations about changes in the patterns of word counts.

Although part of the appeal of DL approaches driving their recent success in NLP is their high predictive accuracy, an equally important appeal of these tools is their modularity and flexibility. Contextual embedding models (of which `BERT` is one

example) are a tool that allows us to model tokens within sequences, and make predictions or inferences between these levels. Although the flexibility of DL tools permits a new approach to language and a new set of estimators, estimands, and questions, this flexibility also introduces a complexity that has limited the utility of these tools in domains where explainability is key. This creates demand for research bridging this explainability gap, and showing applications of DL tools validated in the epistemology of the respective target domains.

I show that DL-based language models combined with feature attribution methods provide a way to describe the linguistic differences between political campaigns that can be justified in terms of empirical sufficiency. Although high predictive accuracy does not tell us *what* makes two campaigns differ, it does provide sufficient evidence that they are differentiable. Although the scores provided by IG cannot be provided in terms of marginal effects, by training new models that are alternately shown high-attribution or low-attribution regions, we can again infer that the scores do correspond to tokens that are informative to differentiating the campaigns. We can conduct statistical tests on these scores using uncertainty measures generated with MC Dropout, and make claims about the statistical significance of token importances.

Finally, because we are able to fit models with entire documents as a unit of analysis, we can construct instance-specific estimands for linguistic phenomena occurring only once in our corpus. This ability to superimpose corpus-level patterns into individual documents brings quantitative and qualitative approaches to text closer, allowing researchers to ask questions about specific phrases in individual documents. This approach is not limited to identifying differentiating in text either; as research moves

in a multimodal direction, this work provides the theoretical foundations for applying the same saliency approaches to image classification models. Future work can also be aimed at constructing annotation tools to augment qualitative text analysis with the advances transfer learning and DL have brought to NLP.

References

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. "Sanity checks for saliency maps." *Advances in Neural Information Processing Systems* 31.

Aggarwal, Charu C et al. 2018. "Neural networks and deep learning."

Atanasova, Pepa, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. "A Diagnostic Study of Explainability Techniques for Text Classification." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3256–74. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.263.

Barabas, Jason, and Jennifer Jerit. 2009. "Estimating the Causal Effects of Media Coverage on Policy-Specific Knowledge." *American Journal of Political Science* 53 (1): 73–89. http://www.jstor.org/stable/25193868.

Barnes, Lucy, and Timothy Hicks. 2018. "Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy." *American Journal of Political Science* 62 (2): 340–54. https://doi.org/10.1111/ajps.12346.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chong, Dennis, and James N Druckman. 2007. "Framing public opinion in competitive democracies." *American Political Science Review* 101 (4): 637–55.

Clark, Tom S., and Benjamin E. Lauderdale. 2012. "The Genealogy of Law." *Political Analysis* 20 (3): 329–50. http://www.jstor.org/stable/23260321.

Denny, Matthew J, and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." *Political*

Analysis 26 (2): 168–89.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

DeYoung, Jay, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. "ERASER: A Benchmark to Evaluate Rationalized NLP Models." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4443–58. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.408.

Diermeier, Daniel, Jean-FranÇOis Godbout, Bei Yu, and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42 (1): 31–55. http://www.jstor.org/stable/41485863.

Edelson, Laura, Tobias Lauinger, and Damon McCoy. 2020. "A security analysis of the Facebook ad library." In *2020 IEEE Symposium on Security and Privacy (SP)*, 661–78. IEEE.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. "How to make causal inferences using texts." *arXiv Preprint arXiv:1802.02163*.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki. 2020. "Keyword assisted topic models." *arXiv Preprint arXiv:2004.05964*.

Ettinger, Allyson. 2020. "What BERT is not: Lessons from a new suite of psy-

cholinguistic diagnostics for language models." *Transactions of the Association for Computational Linguistics* 8: 34–48.

Fong, Christian, and Justin Grimmer. 2019. "Causal inference with latent treatments." *American Journal of Political Science*.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. "Political advertising online and offline." *American Political Science Review* 115 (1): 130–49.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In *international conference on machine learning*, 1050–59. PMLR.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115 (16): E3635–44.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy. 2019. "Measuring group differences in high-dimensional choices: method and application to congressional speech." *Econometrica* 87 (4): 1307–40.

Grimmer, Justin. 2013. "Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57 (3): 624–42. https://doi.org/10.1111/ajps.12000.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Hayes, Danny, and Jennifer L. Lawless. 2015. "As Local News Goes, So Goes Citizen Engagement: Media, Knowledge, and Participation in US House Elections." *The Journal of Politics* 77 (2): 447–62. https://doi.org/10.1086/679749.

Hofman, Jake M, Amit Sharma, and Duncan J Watts. 2017. "Prediction and explanation in social systems." *Science* 355 (6324): 486–88.

Jain, Sarthak, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. "Learning to Faithfully Rationalize by Construction." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4459–73. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.409.

John, Peter, and Will Jennings. 2010. "Punctuations and Turning Points in British Politics: The Policy Agenda of the Queen's Speech, 1940–2005." *British Journal of Political Science* 40 (3): 561–86. https://doi.org/10.1017/S0007123409990068.

Lall, Ranjit, and Thomas Robinson. 2021. "The MIDAS touch: accurate and scalable missing-data imputation with deep learning." *Political Analysis*, 1–18.

Lapinski, John S. 2008. "Policy Substance and Performance in American Lawmaking, 1877–1994." *American Journal of Political Science* 52 (2): 235–51. https://doi.org/10.1111/j.1540-5907.2008.00310.x.

Lauderdale, Benjamin E, and Alexander Herzog. 2016. "Measuring political positions from legislative speech." *Political Analysis* 24 (3): 374–94.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97 (2): 311–31.

Lei, Tao, Regina Barzilay, and Tommi S. Jaakkola. 2016. "Rationalizing Neural

Predictions." In *EMNLP*, 107–17. http://aclweb.org/anthology/D/D16/D16-1011.pdf.

Li, Zhe, Boqing Gong, and Tianbao Yang. 2016. "Improved dropout for shallow and deep learning." *Advances in Neural Information Processing Systems* 29.

Lin, Yongjie, Yi Chern Tan, and Robert Frank. 2019. "Open Sesame: Getting inside BERT's Linguistic Knowledge." In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 241–53. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4825.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23 (1). https://doi.org/10.3390/e23010018.

Lipton, Zachary C. 2018. "The Mythos of Model Interpretability." *Commun. ACM* 61 (10): 36–43. https://doi.org/10.1145/3233231.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A robustly optimized bert pretraining approach." *arXiv Preprint arXiv:1907.11692*.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16 (4): 356–71. https://doi.org/10.1093/pan/mpn004.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. "What is your estimand? Defining the target quantity connects statistical evidence to theory." *American Sociological Review* 86 (3): 532–65.

Mathur, Arunesh, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. 2020. "Manipulative tactics are the norm in

political emails: Evidence from 100K emails from the 2020 U.S. election cycle." https://electionemails2020.org.

Monroe, Burt L, Michael P Colaresi, and Kevin M Quinn. 2008. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16 (4): 372–403.

Peterson, Andrew, and Arthur Spirling. 2018. "Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems." *Political Analysis* 26 (1): 120–28.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084.

Rheault, Ludovic, and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28 (1): 112–33.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Roberts, Margaret E, Brandon M Stewart, and Richard A Nielsen. 2020. "Adjusting for confounding with text matching." *American Journal of Political Science* 64 (4): 887–903.

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis*, 1–25.

Rodriguez, Pedro L, Arthur Spirling, and Brandon M Stewart. 2020. "Embedding Regression: Models for Context-Specific Description and Inference in Political Science."

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A primer in bertology: What we know about how bert works." *Transactions of the Association for Computational Linguistics* 8: 842–66.

Rush, Alexander. 2018. "The Annotated Transformer." In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 52–60. Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2509.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *ArXiv* abs/1910.01108.

Skansi, Sandro. 2018. *Introduction to Deep Learning: from logical calculus to artificial intelligence*. Springer.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22. http://www.jstor.org/stable/25193842.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "Mpnet: Masked and permuted pre-training for language understanding." *Advances in Neural Information Processing Systems* 33: 16857–67.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research* 15 (1): 1929–58.

Stanig, Piero. 2015. "Regulation of Speech and Media Coverage of Corruption: An

Empirical Analysis of the Mexican Press." *American Journal of Political Science* 59 (1): 175–93. https://doi.org/10.1111/ajps.12110.

Storks, Shane, and Joyce Chai. 2021. "Beyond the Tip of the Iceberg: Assessing Coherence of Text Classifiers." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3169–77. Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.272.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic attribution for deep networks." In *International conference on machine learning*, 3319–28. PMLR.

Torres, Michelle, and Francisco Cantú. 2022. "Learning to see: Convolutional neural networks for the analysis of social science data." *Political Analysis* 30 (1): 113–31.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in Neural Information Processing Systems* 30.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." *Advances in Neural Information Processing Systems* 33: 5776–88.

Wiegreffe, Sarah, and Yuval Pinter. 2019. "Attention is not not Explanation." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 11–20. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002.

Yu, Mo, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. "Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4094–4103. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1420.

Appendix 2

Integrated Gradients Explained

In this section I provide a technical explanation of integrated gradients (IG) as a feature importance metric.

As noted in the main body of the article, IG is essentially similar to understanding a linear model in terms of its coefficients, but makes a few adjustments that are specific to the behaviour of deep learning models.

First, a quick recap on multivariate linear models and how we interpret them. This part should be familiar to most political scientists, but is worth revisiting to clarify how we think about interpreting models. Suppose we have the following linear model that expresses y as a function of x_1 and x_2, $F(x_1, x_2)$. (For all that follows, I mainly focus on $F(\cdot)$ instead of y to emphasize that we are trying to interpret a *model*.)

$$F(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_0$$

Given some data, we estimate (constant) values for coefficients β_1 and β_2 that minimise the prediction error for that data. We then typically use a model like this for two tasks: prediction and explanation. For prediction, we calculate a predicted value of the outcome given some values x_1 and x_2 by "plugging in" these values into the $F(x_1, x_2)$. For explanation, we look at the change in the predicted outcome for a one-unit increase of a single input x_i (in this example, $i \in \{1, 2\}$), holding the

other input constant. The question we are trying to answer here is "how much does the predicted value of y change if I vary a single input x_i?" Using terminology of calculus, this value is the partial derivative of the output with respect to a single output: $\delta F/dx_i$.

In the case of a linear model (specifically in the sense of having no polynomial terms), the coefficient on x_i,β_i, is the partial derivative of the output with respect to an input: $\delta F/dx_i = \beta_i$. Moreover, the partial derivatives for a linear model are constant and will always be equal to the coefficients.

However, if we were to include a polynomial term such as the model below, the partial derivative of the output with respect to inputs would not be constant:

$$F(X_1, X_2) = \beta_1 x_1^2 + \beta_2 x_2 + \beta_0$$

For this model, the partial derivative of F with respect to the inputs are $\delta F/dx_1 = 2\beta_1 x_1$ and $\delta F/dx_2 = \beta_2$. The importance of this is that the relative amount that x_1 "contributes" to a given prediction of the model varies in the space of x_1, x_2. Expressing these relative contributions as ratios, the relative contribution of x_1 and x_2 for $F(0, x_2)$ is $0 : 1$, $F(1, x_2)$ is explainable as the ratio $\beta_1 : \beta_2$, $F(2, x_2)$ is $2\beta_1 : beta_2$, an so on.

That the partial derivative of x_1 varies complicates the question of "how much does x_1 matter for predicting y?", because the answer becomes "it depends (on the value of x_1)". Sometimes it matters a lot (when x_1 is large), and sometimes it does not

matter at all. Fortunately, this is still somewhat of a non-issue because we can fully describe this behaviour in functional terms, i.e. $2\beta_1 x_1 : \beta_2$.

We can apply this same logic of the ratios of partial derivatives to interpret the behavior and predictions of any model, including neural networks. For a given set of inputs, the relative importance of each input can be expressed in terms of the ratios of their partial derivatives at that point. However, using this strategy to calculate point-estimations of feature importance is ineffective for neural networks because of the discontinuous activation functions used at many nodes and the fact that most models are trained to saturation. The consequence of this is that there are many localized regions for which inputs have zero importance. The alternative of trying to describe the global behaviour of partial derivatives (such as the functional expression at the end of the previous paragraph) is likewise challenging because even if this functional form is tractable, its extreme complexity will make it as uninterpretable as the model itself.

The alternative proposed in Sundararajan, Taly, and Yan (2017) (and elsewhere) is to calculate some substantively motivated local, but not point, estimate of the ratio of importances. They argue for calculating the path integral from some substantively motivated baseline to the input of interest. They further argue that the only valid path integral is the straightline path. The three panels of Figure 2.8 help illustrate what this path integral is, and why it needs to be the straightline path. In this figure, we consider the function $F(X_1, X_2) = min(0.3x_1^2 + 1.5x_2, 1)$. The min operation has been added to this function to emulate the discontinuous behavior of neural networks, which typically have floor or ceiling functions in intermediary layers.

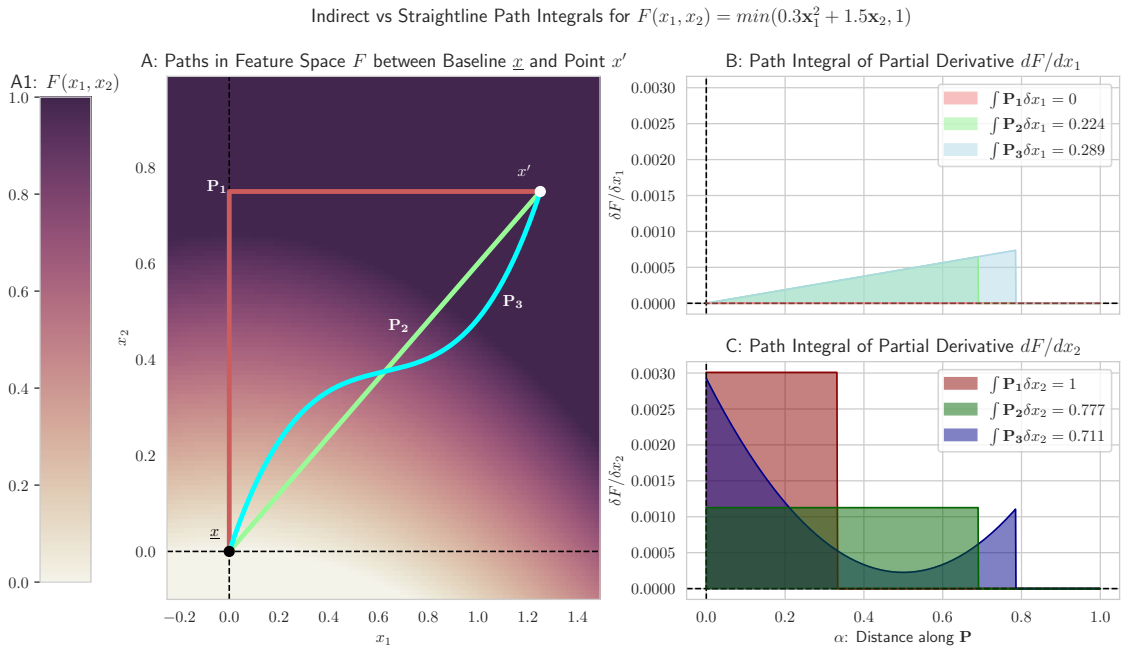

Figure 2.8: Integrated Gradients Explained

Panel A shows the value of $F(x_1, x_2)$ for combinations of x_1 and x_2, with a darker shade indicating a higher value of $F(x_1, x_2)$ according to the color map to the left. Because of the *min* operator in F, the dark region to the top and right of the figure is "capped" at 1. The three lines labelled P_1, P_2 and P_3 show three paths (i.e. sequences of points) through the feature space from a baseline $\underline{x}$ at $(0, 0)$ to x' at $(1.25, 0.75)$.

We can approximate the partial derivatives of $F(x_1, x_2)$ with respect to its inputs along the path by "moving" along the path and then noting how F, x_1 and x_2 change. For instance, P_1 initially moves directly upwards 0.75 units along x_2. Along this vertical part, each tiny movement along P_1 only increases x_2 while x_1 remains constant. Thus for this portion of P_1, the partial path integral of x_1, $\int P_1 \delta x_1$ is zero. Furthermore, note that once P_1 "turns right" and moves 1.25 units in x_1, F

is constant because of the ceiling effect of the *min* operator. Therefore the integral along this portion of the path is still zero. This behaviour is shown in Panels B and C: in the bottom panel, the large red box corresponds to the integral over the partial derivative of $F(x_1, x_2)$ with respect to x_1 along P_1. Because there is no point along P_1 where F increases in x_2, the path integral with respect to x_2 is zero.

Now consider P_2. Each movement along this path increases both x_1 and x_2 by proportional amounts. We see the path integral increasing in x_1 because of the polynomial term in F on x_1, while it remains constant for x_2. For the splined path P_3, the path spends slightly longer in a non-flat region of F, thus the shaded region under the partial derivative stretches further along α, the length of the path.

Sundararajan, Taly, and Yan (2017) argue that the straightline path, P_2 is the only implementation invariant path integral; we see from this exercise that this is true. P_1 would tell us that x_2 is the only important feature, whereas we can distort P_3 arbitrarily to vary the relative sums of the partial path integrals with respect to the two inputs.

Needles in a Haystack: A Semantic Role-Based Approach to Selecting Observations in Text Data

Two recent works highlight the importance of selection of observations as a critical initial step in the analysis of large text datasets. Whereas these papers develop methods for formulating queries that better capture a concept of interest in the data, I present an approach to improving document selection through the use of richer representations of text data. I put forward an alternative that combines linguistic theory formalizing semantic structures in language, called semantic roles, with computational tools for exposing this structure. I explain and illustrate the utility of a semantic role-based approach to observation selection in text data with an application studying partisan animus and the use of the threat in a corpus of campaigning emails from the 2020 United States election cycle. My findings show that the ability to filter text on an abstracted representation permits the investigation of questions and phenomena that are not possible with existing tools. Additionally, in contrast to distributional representations that are increasingly popular in political

science text-as-data, the structured approach to text taken in this paper produces selection strategies that are transparent, predictable, and therefore less prone to proliferating the biases exhibited by distributional models of semantics.

Introduction

Several recent works in political science methodology bring attention to the crucial importance of an early step in most analyses of text data: selection of a relevant set of documents from a larger corpus on the basis of the content of the texts. This step is especially important where the majority of available documents in the corpus are not relevant to the research question, such as studies using newspapers or parliamentary speech data. Although subset selection is not unique to the text-as-data setting, it deserves particular attention because of the non-trivial operationalization decisions that must be made before filtering can be applied.

Existing methods for content-based subset selection in text datasets are predominantly token-based (keywords) or document-based (supervised classification). In this paper I present an alternative approach to text subset selection using semantic role labelling (SRL). This approach retains the transparency and predictability of keyword-based approaches, while also allowing the researcher to specify more narrowly the role–action, agent, or patient–a word occupies in a sentence. This enables the researcher to abstract from the precise phrasing of a query to its underlying semantic structure and brings the query and target closer together in meaning.

I demonstrate the application of this tool with an application to studying the content and strategies of email-based campaigning in the 2020 United States (US) election. Motivated by literature on studying negative messaging (e.g. Ridout and Franz 2008), partisan affect (Iyengar and Krupenkin 2018) and the use of fear, threat (Jerit 2004) and urgency (Mathur et al. 2020) in political campaigning, I study instances of Democratic and Republican candidates talking about defeating their opponent or

the threat of being defeated by their opponent. I provide a comparison of how such a task can be attempted with keywords-only and keywords-plus-supervised classifier approaches.

In addition to providing a useful tool for selecting observations in text data, I attempt in this paper to motivate broader consideration about representations of text for quantitative analyses. I highlight the opportunities and dangers of applying complex models for meaning in language to social science questions, and consider how to develop sophisticated approaches that are nevertheless transparent and predictable.

Theory and Literature

Recent works in text-as-data methodology highlight the critical importance of document selection (D'Orazio et al. 2014; King, Lam, and Roberts 2017). This step follows question definition, and is analogous to the construction of a sample from a population of interest in non-text research (Grimmer, Roberts, and Stewart 2022, 35). Existing methods for selecting on text data can be divided into two categories. One focuses on the development of a list of keywords to filter observations, and the other consists of building a set of examples of relevant and irrelevant texts, and then training a machine learning classifier to extrapolate these labels to a larger dataset. In this section I provide an overview the objectives of researchers when filtering text data, the tools available, and the challenges faced when applying them.

Existing Methods and Challenges

An increasing number of political science studies leverage rich sources of information contained in text datasets. However, it is not always necessary when analyzing a text dataset to select observations on the basis of automated analysis of text, and there may be reasons to prefer these other approaches when they are available.

In some cases, inclusion in the sample can be determined on the basis of document covariates instead of content. Both Barberá et al. (2019) and Pan and Siegel (2020) use document metadata to construct multiple subcorpora of tweets to represent various groups of interest. In their study of legislator responsiveness to voter priorities on Twitter in the US, Barberá et al. (2019) collect an enormous corpus consisting of tens of millions of tweets using a mixture of strategies. For legislators, they collect *all* available tweets from members of Congress during the period of interest based on a list of handles. From the public, they construct four samples: a random sample of the users residing in the US using a random number generator; a random sample of users who followed at least one of five major media outlets; a random sample of users who followed three or more Republican members of Congress and no Democrat; and the vice versa for Democratic-following Twitter users. Inclusion in these samples is therefore based either on authorship traits or author behavior traits (i.e. following). The advantage of selection on non-text covariates versus content is that the space of possible values is typically constrained, and the implications of filtering on these values are more transparent.

Often, text datasets do not have pre-defined covariates that capture the population of interest. One common setting is news media data. Although a common strat-

egy is to select text news data by outlet, date and section–with outlet representing particular audiences or perspectives and date period regarding some event of interest. Further filtering is typically necessary because the content of news is highly heterogeneous. Newspapers and news websites include a multitude of sections, from political or business news to horoscopes, sports and celebrity news. If the researcher is concerned with coverage of a narrow topic or issue, the additional noise in the data due to including all content from a news outlet will make it hard to identify relevant phenomena.

Where the researcher has the requisite resources, one option is to employ human coders to filter a larger corpus on the basis of text content. In their study linking citizen knowledge of political issues with media coverage, Jerit, Barabas, and Bolsen (2006) use human coders to label the full text transcripts of three US media outlets for six-week periods prior to several waves of surveys as relevant to one of their political issues of interest, and include an article in the sample if intercoder reliability is sufficiently high. The advantage of using manual coders versus automated content analysis is human natural language understanding ability, and the ability to interrogate or explain individual labels.

Where human coders are not an option, a common strategy is the use of *keywords*. Keywords are lists of one or more words that are used to apply a filtering inclusion/exclusion logic depending on whether a text does/does not contain one of these words. I include in this category the use of *regular expressions*, such as `econom[a-z]+`, which can be expanded into the list of all words beginning with the letters "econom": economy, economics, economical, economist, and so on. Many interfaces for accessing text data permit queries using a limited subset of regular

expressions, such as Twitter and NewsAPI.

Keyword strategies are a popular and simple option frequently used by political science scholars. Some studies use single keywords: in a study of news coverage of austerity, Barnes and Hicks (2018) choose articles that contain the word "deficit", while Baum (2013), who in his study of the influence of foreign press on public support for the invasion of Iraq, chooses articles that contain the word "Iraq". Others employ multiple keywords, such as Blumenau and Lauderdale (2018), who identify crisis-related votes in the European Parliament by searching for several crisis-related phrases (2018, 471) in legislative summaries, and Weaver (2019) who tests hypotheses on the historical de-legitimation of lynching using newspaper coverage near lynching events and selects articles "using the keywords and phrases pertaining to lynching and different arguments and narratives used to justify or criticize the practice" (2019, 301).

However, King, Lam, and Roberts (2017) show that human keyword selection is extremely unreliable. Running an experiment with mostly undergraduate political science majors, they show that that even for salient and narrow topics such as Obamacare or the Boston Marathon Bombings, there is high variability and low overlap in the keywords chosen by participants. The authors note that this result can be explained by findings in psychology that the cue of remembering a few keywords "strongly *inhibits* the recall of the rest of the set, even though you would recognise them if revealed" (King, Lam, and Roberts 2017, 4).

One solution to this cognitive limitation is to present researchers a list of words to choose from, which limits the need for the researcher to recall and store large

amounts of information by instead making a series of mostly independent decisions. This is one advantage of the strategy of D'Orazio et al. (2014), who present a method for selecting text observations using a combination of manual labelling and supervised machine learning. In order to choose relevant observations from the militarized interstate disputes dataset (MID4), the authors argue for the following two-stage procedure. In the first inductive step, the researcher (or a team of human coders) labels a subset of the corpus as relevant/irrelevant and trains a supervised machine learning algorithm[1] on these samples. In a second transductive step, in order to mitigate the effect of time-specific confounders, the authors fit a series of time-specific SVMs by sampling pre-labelled samples from the results of the first-stage classifier with a fixed ratio of positives to negatives.

However, a challenge for this strategy is dealing with within-text confounders. In their framework for causal inference of latent treatments via text-based interventions, Fong and Grimmer (2019) note that texts contain many latent properties, and estimating the effects of these properties is likely to be confounded by other unmeasured latent traits. Although we are not attempting to estimate unbiased effects, training a classifier to label documents as relevant faces a similar challenge. Relevance is one of many latent properties of a text, but when we train a classifier we cannot ensure that its decision rule will be based solely on the relevance property. Thus the rule learned by the classifier may instead capture any latent property that positive samples have in common, making its extrapolation behavior unpredictable.

[1] D'Orazio et al. (2014) specifically argue that researchers should use a support vector machine (SVM) for text classification based on the finding that SVMs are suited to the high dimensionality and extreme sparseness of a document-term matrix.

This problem is particularly likely to occur when the corpus contains highly heterogeneous texts. Although the standardized format of records such as MID4 corpus used by D'Orazio et al. (2014) is likely to have a higher degree of consistency in structure and language, domains such as newspapers, speech transcripts and especially social media are likely to have multiple higher-order latent sources of variance that may obscure relevancy, including author-specific effects (Lauderdale and Herzog 2016). Given the relatively opaque nature of the decision rules of machine learning classifiers (or in the case of an SVM, the high dimensionality of the feature space through which it draws a hyperplane), there is no straightforward way to determine whether a classifier has learned to discriminate documents solely on the basis of relevancy.

The solution offered by King, Lam, and Roberts (2017) may offer a best-of-both-worlds. Specifically designed to address the unreliability of human-selected keywords lists, they offer the following strategy for selecting text observations. The first step is similar to D'Orazio et al. (2014); the researcher manually classifies a subset of documents as relevant/irrelevant, then trains a diverse set of algorithmic classifiers to extrapolate these labels to an unseen set of texts. However, in the second step, a second model mines keywords from the set of relevant documents and the researcher manually selects from these. This step can simply consist of identifying the words that only occur in the relevant set, then using personal judgment to choose words from this list. These two steps can be done iteratively based on the sets of returned documents for the chosen keywords until the researcher is satisfied with the keyword list and selected documents.

In addition to providing a solution to the cognitive challenges of building keyword lists, the use of a Boolean filtering logic for selection has the advantage of being a

transparent and interpretable decision rule. This method is particularly advantageous when the researcher does not have access to the full text of the documents prior to selecting them, such as when using an API with a request limit.

Nevertheless, even computer-assisted keyword-based approaches are limited. The goal when constructing a sample is to include documents that contain or describe a phenomenon or concept of interest, and to exclude those that do not. Keyword-based strategies work by leveraging the fact that particular words tend to occur in mentions of these target concepts–this is the logic operationalized explicitly in King, Lam, and Roberts (2017).

However, because the link between keywords and concepts is based on co-occurrence[2] and not equivalence, the overlap between a word and the mention of a target concept is not guaranteed. In particular, because words often have multiple meanings, senses and uses (polysemy), keyword-based strategy are restricted by an implicit trade-off of choosing words that are likely to co-occur with the target concept while avoiding words that co-occur with other, irrelevant concepts.

Ultimately, a fundamental challenge of a keyword-based strategy is that it is only a proxy for what the researcher is actually searching for. Barnes and Hicks (2018) are interested in describing the language that newspapers use when reporting on fiscal policy (2018, 346). This will overlap with articles containing the keyword "deficit", but may *systematically* exclude discussions of fiscal policy that do not contain that word. The exclusion can introduce biases both in the types of fiscal policy issues considered and the register used by the author when discussing these issues. Weaver

[2]Unless the target itself is a specific word, such as in Rodman (2019) where the author investigates the changing semantic context of the word "equality" over a long period.

(2019) notes the "imperfect mapping between keywords and contextual meaning" (2019, 301), and chooses keywords "to minimize false positives, though this may result in more false negatives (discussion of lynching using other words)" (2019, SI-D, 61).

Authors using keyword-based strategies for searching are invariably familiar with the trade-offs this approach provides, and frame downstream analyses with the caveats that their data collection strategy requires. These examples should not be taken as a criticism of their strategy or results, but an example of the challenges researchers face and a motivation for having more tools for this difficult step. In sum, there are several desirable qualities for a strategy for observation selection on text content. There should be a clear link between the target concept, query and decision rule in order to help identify false positives due to confounding. A strategy also benefits from having an inductive, "discovery" component (Grimmer, Roberts, and Stewart 2021) because the link between the target concept and the query may not be clear from the outset and may need to be adjusted as additional edge cases are found.

Semantic Roles

The alternative strategy I present supplements existing methods by reconsidering the representation of text data for subset selection. My aim is to enable researchers to build more complex linguistic features into their queries, in order to match their target concept or phenomenon more closely. In this section I introduce the linguistic theory behind semantic roles, and demonstrate how they are used to systematize the extraction of meaning from text.

One challenge of dealing with text data is that the same idea or event may be expressed in a large number of distinct ways. Take for example the following sentence, which can be expressed in several semantically equivalent ways:

(1) *Russia invaded Ukraine on 24 February 2022.*

This can be rearranged into the passive voice:

(2) *Ukraine was invaded by Russia on 24 February 2022.*

Put the temporal clause first:

(3) *On 24 February 2022, Russia invaded Ukraine.*

And even substitute one or more words with synonyms:

(4) *Russian forces attacked Ukraine earlier this year.*

Linguists have developed a framework for abstracting from the specific formulation of a sentence to an underlying predicate structure through a process called Semantic Role Labelling (SRL). SRL is most easily explained with examples.[3] The predicate structure of the sentence (1) above would be as follows:

```
[ARG0: Russia] [V: invaded] [ARG1: Ukraine] [ARGM-TMP: on
24 February 2022]. [4]
```

[3]For an in-depth introduction to semantic role labelling, see Chapter 19 of *Speech and Language Processing*, 3rd Edition by Dan Jurafsky and James H. Martin.

[4]I manually constructed these parses and compared my answers to the results of the online SRL tool available at `http://en.sempar.ims.uni-stuttgart.de/parse`.

Parsing begins with identifying the predicate of the sentence, which is almost always a verb. The predicate of the sentence is "invade", marked by `V`. Each verb predicate has an associated set of arguments, compiled in detail in the publicly available PropBank (Johansson and Nugues 2008). Although the definitions of the arguments varies between verbs, they are labelled in a consistent way that enables comparison. PropBank lists the verb `invade` as having two arguments: `ARG0`, the invader, which is Russia, and `ARG1`, the place invaded, which is Ukraine. In general, `ARG0` represents a proto-agent (the grammatical subject, or do-er) and if it is a transitive verb `ARG1` represents a proto-patient (the thing that the doing is done to). The above parse also contains the tag `ARGM-TMP`, which is a temporal argument that indicates *when* the action takes place.

The power of SRL comes in the fact that the first three sentences above have identical parses. Given that sentences (2) and (3) are represented as:

```
[ARG1: Ukraine] was [V: invaded] [ARG1: by Russia] [ARGM-TMP:
on 24 February 2022].
```

```
[ARGM-TMP: On 24 February 2022], [ARG0: Russia] [V: invaded]
[ARG1: Ukraine].
```

Representing the sentences as the set {`V`, `ARG0`, `ARG1`, `[...]`}, all three sentences are identical: {invade, Russia, Ukraine, (24 February 2022)}. For the present analysis, I ignore arguments "beyond" `ARG1`. This is primarily because the use of `ARG2` to `ARG5` is inconsistent.[5]

[5]For example, `ARG2` can be the benefactive (the entity affected by an action being done to the patient), the instrument (the thing used do the action), the end-state (the state the patient arrives by having been acted on), and so on.

Dealing with synonyms requires some additional steps. The fourth sentence is parsed as follows:

```
[ARG0: Russian forces] [V: attacked] [ARG1: Ukraine]
[ARGM-TMP: earlier this year].
```

This can be represented as: {attack, Russian forces, Ukraine, (earlier this year)}. In their application of SRL to the automated extraction of narratives from text, Ash, Gauthier, and Widmer (2021) provide a pipeline for identifying this as a sufficiently equivalent set to {invade, Russia, Ukraine},

Synonyms and Distributed Representations

Because the space of verbs is small, the difference in verb can be addressed in part through the use of synset relations.[6] A synset is a set of one or more words that are considered semantically equivalent. In addition to direct synonyms, I use hypernym and hyponym relations to identify the set of verbs with more and less specific meanings for a given verb (in contrast because their goal is to reduce the feature space, Ash, Gauthier, and Widmer 2021 use a lowest common ancestor logic). In this case, "invade" is a hyponym of "attack", and "attack" is a hypernym of "invade". Synset relations are particularly useful and appropriate in the case of verbs because precision is important–minute differences in verb can radically change the meaning of a sentence.

[6]Synset relations can be found online at https://wordnet.princeton.edu/("About WordNet," n.d.), and are implemented in Python using the `nltk` library (Bird, Klein, and Loper 2009).

The space of the arguments is considerably larger (possibly infinite), making a pre-labelled approach less feasible. One of the key innovations in Ash, Gauthier, and Widmer (2021) is to use word embeddings and unsupervised clustering to solve this problem. Given a parsed corpus, `ARG0` and `ARG1` are represented by the average `GloVe` embedding (Pennington, Socher, and Manning 2014) of the nouns, verbs and adjectives of the labelled span. The set of all `ARG0` and `ARG1` embeddings is then clustered with the k-means algorithm, and the clusters are labelled according to the most common embedding within the cluster. In the resulting simplified SRL representation of the corpus, all `ARG0` and `ARG1` are replaced with their corresponding cluster label.

The logic upholding this strategy is that words with similar distributed representations (i.e. word embeddings) have similar meanings (semantic values). Combined with thorough validation (which the authors perform), the technique allows for the efficient aggregation of SRL sets and rich summary of a large corpus. However, the application of embeddings for reducing various entities to a single cluster label can lead to problematic equivalences, especially in applications involving social groups[7].

Much of the progress in the automated extraction of meaning from text has relied on a distributional approach to semantics (Harris 1954; Boleda 2019 provides recent discussion and review), which leverages an empirical regularity that words occurring in similar contexts tend to have similar meanings. Thus many natural language processing (NLP) models (e.g. embedding models such as `Word2Vec`, `GloVe` and

[7]Given their extensive validation and transparency, I speculate that this was less of an issue for the congressional speech corpus used by Ash, Gauthier, and Widmer (2021).

`BERT`) are models for describing the complex relationship between a word and its context. However, these are models for describing how language is *used* in the corpus the model is fitted on, which is a product of the social context that produced the corpus. This distinction is consequential because meaning and association as provided by these models encodes the biases of the data it is trained on (and social context that produced that data), including problematic stereotypes on the basis of race, gender, profession and so on (Caliskan, Bryson, and Narayanan 2017).

When using automated methods to group the space of arguments under a set of discrete labels, relying on models utilizing a distributional approach to semantics transfers encoded biases to our analysis, specifically by equating entities on the basis of contextual association. The concrete implications of this behavior can be demonstrated with our running example. For instance, using the GloVe model employed by Ash, Gauthier, and Widmer (2021), `Russia` and `Russian forces` have a cosine similarity of 0.64 whereas `Russia` and `Ukraine` have a cosine similarity of 0.77. This result persists with the more sophisticated embedding model I use in this paper, which gives the scores 0.55 and 0.80 respectively. Additionally, this does not seem to be due to comparing countries versus a country and its military; `Russia` and `Japan` have a cosine similarity of 0.62 using GloVe. Clearly, grouping Russia and Ukraine into the same cluster while leaving Russian forces separate is deeply problematic, but it also would lead to erroneous conclusions if we used these embeddings to reduce and cluster the space of labels.

Given that our application is selection and not corpus summary, the alternative I prefer is manual labelling. In the example application, by filtering sentences by `V` first, the total number of spans that needs to be manually labelled is greatly re-

duced. Computationally-assisted options include, for instance, the use of named entity recognition (NER) to recognize specific actors (such as politicians), then merging this with available datasets on actor characteristics (such as party, from Ballotpedia).

Whereas recent developments in document selection discussed in the previous section (D'Orazio et al. 2014; King, Lam, and Roberts 2017) develop complexity and specificity of a query to fit the data, the aim is to leverage a richer representation of text data that allows us to "fit the data to the query". By specifying not only keywords, but the role they occupy in a sentence, using a semantic role-based approach allows researchers to match on the functional structures of the language that describes or produces the target concepts and phenomena.

Finally, although the steps after selection are largely beyond the scope of this paper, the {action, agent, patient} representation of sentences in a corpus allows for systematic exploration of substantive patterns. One can study the roles that a particular entity tends to occupy–for instance, in the corpus of campaigning emails analyzed later in this paper, "immigrants" and related entities occupy `ARG1` more frequently than `ARG0`, showing that they are discussed as the subject of an action than a group with agency. This in of itself is suggestive of the way that immigrants are discussed in the context of campaigning; not as the intended audience or a group with entity, but as a group that either the candidate or voter will do something to (whether help, harm or otherwise). Further details of this analysis are included in the appendix.

Data and Application

I demonstrate the semantic role-based search with an application to studying the use of animus (i.e. negative attitudes towards the opposing party) and fear of losing as a fundraising and mobilizational tool in a corpus of political campaigning emails from the 2020 US election. This study draws on two themes in existing scholarship on the content of political campaigning. The first relates to the study of partisan affect. Researchers have found that partisan animus has had an increasing impact on political participation in the past two decades (Iyengar and Krupenkin 2018). Accordingly, I expect to find evidence of political campaigns motivating their supporters through oppositional language, based on "defeating" the enemy (the opposing party).

Another strategy employed by political campaigns is the use of fear, threat and urgency (Jerit 2004; Mathur et al. 2020) as a means to engage emotional responses to mobilizational appeals. These strategies take various shapes and forms, including the infamous advertisement run by Lyndon B. Johnson in 1964, that associated his opponent Barry Goldwater with the threat of nuclear annihilation (Jerit 2004, 566). As partisan affect increases, the stakes associated with the opposing party coming into government become higher. I therefore also search for parties using the threat of the opposing party winning the election as a mobilizational and fundraising tool.

I study this question using a corpus consisting of $169,152$ political emails collected by Mathur et al. (2020) between 2 December 2019 and 11 November 2020 from $2,569$ political campaigns. Senders include "candidates in federal and state races as well as Political Action Committees (PACs), Super PACs, political parties and other political organizations" (Mathur et al. 2020, 2). Along with metadata on

the sender including office sought, office level, party affiliation, and other covariates, the dataset includes the subject and body text with identifying information and externally directed URLs removed.[8]

Several features of the body text of campaigning emails make it a challenging corpus to filter observations on. There is typically a significant portion of shared text between emails, in the form of boilerplate: "Dear <name>, <greeting>" at the beginning and "<complimentary close>, <sender>, <legal disclaimer>" at the end. Therefore the phenomenon of interest is often at the sub-document level, in the main body text of the email. Within this, Mathur et al. (2020) find that the sentences in emails display topical heterogeneity, typically starting with a closer focus to the subject line of the email and moving towards a fundraising message (2020, fig. 3). For most applications, this results in an increase in the ratio of noise to signal for each observation. Additionally, this dataset is highly imbalanced on party affiliation, with $135,475$ emails from Democratic Party-affiliated campaigns and only $33,677$ emails from Republican Party-affiliated campaigns.

Within this corpus, I search for emails where the candidate mentions either defeating the opposing party, or their own party being defeated. This task is meant to be the first step in a content analysis of how increasing partisan affect manifests in campaigning and fundraising, either through animus or threat of loss.

I compare three approaches to the task. The first uses the semantic role representation of the dataset to directly identify sentences mentioning defeating the opponent. To demonstrate that a semantic role approach provides something that cannot be

[8]Details on the exact collection method as well as access to the full dataset are available at `https://electionemails2020.org/`.

done with existing methods, I attempt the same task using a keyword-based approach and a combined keywords-plus-classifier approach. Although this task naturally lends itself to a semantic role-based approach, it is also the case that research questions are inevitably shaped by the tools available. I argue that it makes sense to ask a question like this for substantive reasons listed above, but also that it has not already been studied because the tools were not available.

I use the same initial pre-processing for all three approaches. Each email is split into sentences using a pre-trained sentence segmentation model, providing a total of $1,126,149$ unique sentences across the $169,152$ emails. I then filter out sentences that only appear in primary-level races because the opponent in these contests are from the same party as the candidate. The final number of sentences for searching $851,056$.

For the semantic role approach, I apply an `BERT`-based SRL model (Devlin et al. 2019; Shi and Lin 2019) from OpenIE via AllenNLP (Gardner et al. 2018), and store the list of unique parses per sentence. In order to simplify the subsequent step, I reduce the spans labelled with roles `ARG0` and `ARG1` by parsing the dependency tree to identify the head noun phrase, and extract and lemmatize this.[9] From the resulting representation of the dataset, I retain all parses that contain a complete {`V`, `ARG0`, `ARG1`} set. Using this representation of the corpus, I identify statements from campaigns that refer to defeating the opponent. I search for all statements in campaign emails with "defeat" or its synonym as the verb (hereafter denoted "defeat*"), and a Democratic or Republican entity as `ARG1`.

[9]The differences between my pipeline and Ash, Gauthier, and Widmer (2021) are explained in more detail in the corpus.

For the keyword-based approach, I select all sentences that contain "defeat", a set of synonyms unlikely to induce false positives, and keywords referring to Democratic and Republicans, the party leaders, or the positions typically associated with these groups. I then compare the overlap between the sentences selected by a keyword-based approach with an semantic role-based approach.

For the combined keyword and supervised learning approach, I pretend that my goal is to identify the same set of sentences where the action is defeat and the patient is the opposing party, and therefore treat inclusion is the semantic role-based sample as the "true" label. For simplicity, I focus only on the case of defeating Democrats. I select an initial "plausible sample" using a keyword-based strategy to select plausible sentences that contain the phenomenon of interest, and supplement it with a random sample of other sentences. I then "manually label" these observations according to whether they were in the semantic role-based sample, and use these observations to train a series of machine learning classifiers. Within each class of classifier, I use the fitted model with the highest recall on the training data in a 5-fold cross validation to identify the set of candidate matches in the rest of the dataset, and then reveal the label of these. The method is then assessed against the semantic role approach based on the proportion of positive samples "uncovered" and the total number of samples that would need to be "manually labelled".

For each of the comparisons, I conduct an error analysis to identify and describe the reasons that the alternatives select different observations from the semantic role-based approach. For the keyword-based approach, this consists of a manual analysis of a random sample of sentences selected by one approach and not the other. For the combined approach, I also consider the feature importances on the supervised

classifiers to partially characterize the latent logic they use to label observations.

Results

Semantic Role Approach

Having prepared the semantic role representation of the corpus, the first step is to identify verbs that have the same meaning or are used in the same way as "defeat" (e.g. beat). I look up the relevant sense of the verb "defeat" in WordNet (`defeat.v.01`: win a victory over) and then collect the list of the lemma form of all of its synonyms, hyponyms and hypernyms. The final list is:

> `get the better of`, `overcome`, `defeat`, `beat`, `beat out`, `crush`, `shell`, `trounce`, `vanquish`, `conquer`, `demolish`, `destroy`, `down`, `lurch`, `skunk`, `nose`, `overrun`, `rout`, `rout out`, `expel`, `survive`, `pull through`, `pull round`, `come through`, `make it`, `upset`, `wallop`

I note that there are a number verbs unlikely to appear in the corpus (e.g. "trounce") as well as words that would return false positives if used as a noun (e.g. "upset"). I retain the latter because the semantic role search allows me to restrict my search to sentences where the token is being used as a verb, and retain the former because they are not wrong *per se* and should not affect the number of false positives or negatives.

In parses with one of the above tokens as the verb, there are a total of $7,924$ unique values for `ARG0` and `ARG1`. I manually label these into six categories: ambiguous ($n = 3,930$), Republican ($n = 2,223$), Democrat ($n = 1,750$), Independent ($n = 11$), Libertarian ($n = 7$), and N/A[10] ($n = 2$). A span is coded as Republican if it referred to the Republican party (e.g. Republican Party, GOP), a Republican politician, a group likely to be referring to Republicans in context (e.g. conservatives, Trump crony), or the ideology, designs or action of a known Republican (e.g. Cornyn's hateful, bigoted agenda). Spans were coded as Democratic in the same way, with the addition that epithets like "radical marxist" were also coded as Democrat on the basis that Republicans frequently refer to Democrats in this manner. If it is ambiguous who or what a span could refer to, it is coded as ambiguous. This included spans such as "extremist", as this is an epithet used by both parties.

Of the sentences with defeat* as `V` and a span classified as Democrat as `ARG1`, $2,226$ are from Republican-affiliated campaigns and 949 are from Democratic-affiliated campaigns. In contrast, for sentences with `ARG1` classified as Republican, $8,090$ are from Democratic-affiliated campaigns, 470 are from Republican-affiliated campaigns. A random sampling of two parsed sentences from each of these four groups is listed in Table 3.1.

Several patterns stand out in the random samples. Firstly, partisan animus appears to be paired with fundraising. The examples are either asking the recipient for money to win, or to fight against the donors backing the opponent. As for campaigns referring to the their own defeat, the threat is either in terms of the outcome (e.g. the opponent taking control of the legislature) or the threat posed by the strategy of

[10]Spans containing only whitespace or other non-word characters were labelled N/A.

the opponent. Overall, the semantic role-based strategy for selecting these samples performs in a transparent and predictable way, meaning that there are no false positives (including among the broader set of results not listed in this paper). I discuss false negatives in the conclusion.

Table 3.1: Eight sentences from subsets chosen with semantic role-based filtering (randomly selected, two per group). Note: `[...]` indicates output has been truncated for presentation purposes.

Subset	Campaign (Seat)	Parsed Text
Dem: defeat Rep	Michigan Democrats, DSCC, Jeanne Shaheen (NH), Barbara Bollier (KS), Lulu Seikaly (TX3), Sara Gideon (ME), MJ Hegar (TX), Gary Peters (MI), Theresa Greenfield (IA)	`Will you help [ARG0: Mark Kelly] [V: overcome] [ARG1: Mitch McConnell 's mega donors] and flip the Senate ?`
Dem: defeat Rep	Lorena Garcia (CO)	`Pitch in 100 , 250 , 1,000 or whatever you can afford today to ensure that [ARG0: we] [V: beat] [ARG1: Gardner] [ARGM-TMP: in November] !`
Rep: defeat Dem	Tracey Mann (KS1), Jacob LaTurner (KS2)	`Here are some ways you can help [ARG0: us] [V: defeat] [ARG1: our opponent and the coastal liberals who are funding her] .`
Rep: defeat Dem	John Cummings (NY14)	`That 's why NYC Police and every resident of New York needs you to step up right now with a critical donation of 15 , 25 , 55 or even more right now to [V: defeat] [ARG1: AOC] .`
Dem: defeat Dem	Julia Brownley (CA26)	`So [ARG0: Republicans] are making historic moves to [V: DEFEAT] [ARG1: Democrats like Julia] and end our Democratic Majority for good .`
Dem: defeat Dem	Raja Krishnamoorthi (IL8)	`But my team just alerted me that [ARG0: Republicans] are panicking and dumped 10 million on ads to [V: destroy] [ARG1: Democrats ' chances of winning] .`
Rep: defeat Rep	Andy Biggs (AZ5)	`And with [ARG0: the Democrats] spending millions to [V: defeat] [ARG1: me and every Republican in Arizona] , your contribution ca n't wait ! [...]`
Rep: defeat Rep	Kevin McCarthy (CA23)	`[ARG0: The radical left] is pulling all of the stops to [V: destroy] [ARG1: President Trump] and obtain power .`

Comparison: Keywords Only

The keyword-based strategy is similar to the semantic role-based one. I begin by selecting a smaller set of synonyms for "defeat" in order to reduce the number of false positives due to polysemy:

> `defeat`, `overcome`, `defeat`, `beat`, `crush`, `demolish`, `destroy`, `overrun`, `rout`, `expel`

Given that I would not have the list of candidate spans to manually label, I instead try to choose a set of entities representing Republicans and Democrat following the same logic as the manual coding exercise. For Democrats, this includes:

> `democrat`, `biden`, `harris`, `pelosi`, `schumer`, `sanders`, `obama`, `liberal`, `progressiv`, `socialis`, `communis`, `marxis`, `dnc`

and for Republicans this includes:

> `republican`, `trump`, `pence`, `mcconnell`, `mccarthy`, `cruz`, `nunes`, `conservativ`, `right-wing`, `gop`, `rnc`

These keywords are then used separately to match the lowercased text of each sentence in the corpus. The final returned samples are the intersection of the defeat* query result and the Democrat*/Republican* query result. I present the overlap between these results using the keyword and semantic role-based methods in Figure 3.1, paying particular attention to when the two methods diverge.

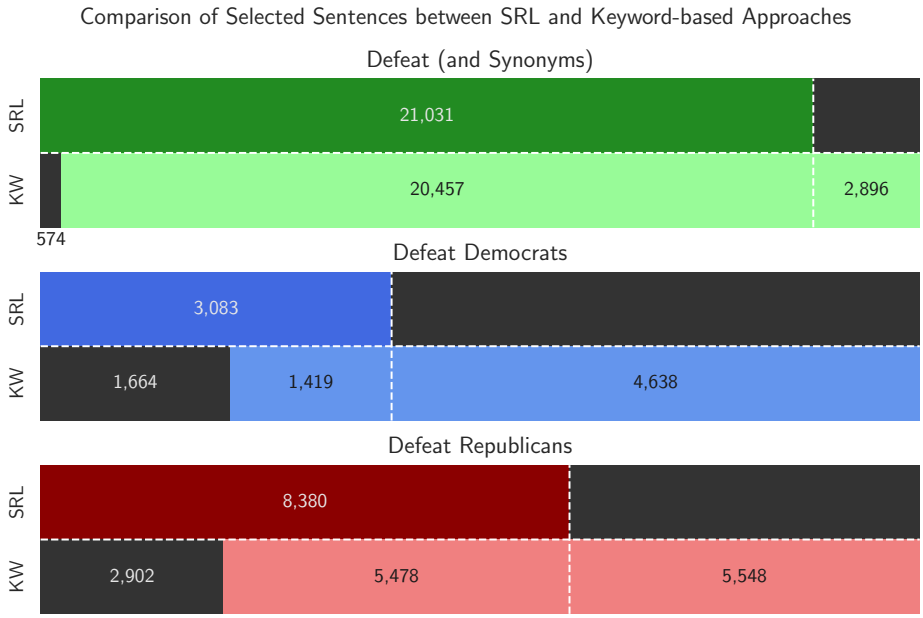

Figure 3.1: Comparison of returned sets between semantic role-based and keyword-based approaches.

Figure 3.1 is similar to a Euler (or Venn) diagram. Within each pair of bars, the upper bar represents the samples returned by a semantic role-based query, and the lower bar represents the samples returned by the keyword-based query. The top pair of bars compares the samples returned when querying only on the verb, the middle pair compares defeat* + Democrat*, and the bottom pair compares defeat* + Republican*. The dark areas represent non-overlap, and the numbers indicate the number of samples within that subset.

The top pair of bars shows that there is not a large difference in returned samples when the query is only looking at the verb. Of a total of $21,031$ sentences returned by the semantic role-based query, $20,457$ also are in the keyword-based result. There are $2,896$ sentences not captured by the keyword-based strategy, and it also returns

an additional 574 sentences not in the semantic role result. Differences appear to be due to not lemmatizing verbs for the keyword-based strategy (e.g. "It's common knowledge what Doug Jones accomplished and overcame in Alabama, but what I don't think most people realize is how he did it.") or a keyword being used as a noun (e.g. "If I miss it, I'll face a brutal defeat.").

Greater differences occur when I combine defeat* with Democrat* or Republican*. For Democrats, the keyword-based strategy "misses" over half of the sentences that the semantic role-based approach retrieves ($1,664$ out of $3,083$), whereas for Republicans this rate is lower ($2,902$ out of $8,380$). Many of these "false negatives" appear to be not providing individual candidate names as keywords–e.g."I will be that Congressman when I defeat AOC this November.", "If you can, we are going to elect Nicole Galloway, defeat Mike Parson, and flip Missouri blue.". This can be mitigated by using a list of candidates contesting races in 2020 as keywords, and removing first names shared by candidates from both parties. "False positives" from the keyword strategy include phrases where the Republican or Democrat keyword is the agent, such as "We can't let Donald Trump's hand picked pawn use horrible, racist tactics to defeat Ammar" being returned in by the defeat Republican keyword query.

The keyword-based strategy also detects some matches that are missed by the semantic role-based strategy, such as the following example:

> "Here's the plan If just 3 more Democrats from 13903 rush 10 before midnight, we'll OUTRAISE our Republican opponent after our debate,

> WIN this Senate race, and hand Mitch McConnell the most humiliating defeat of his career." (Anthony Brindisi, D-NY22)

Here, although "defeat" is not the verb, "hand Mitch McConnell […] defeat" is arguably relevant to to the kinds of sentences that I want to identify if I am studying instances of campaigns saying they will defeat each other or be defeated. I discuss the issue of false negatives further below.

Comparison: Keywords plus Classifiers

In this comparison task, the sentences returned by the semantic role-based strategy are treated as the true labels, and the goal of the classifier is to identify as many of these as possible for a subsequent "manual" coding. To simplify the analysis, I only consider sentences where the `ARG1` is a Democrat.

I begin by "creating" training data. I use the keyword-based strategy to select a set of observations ($n = 6,057$) that I "manually" label by revealing whether they were included in the semantic role result. This is meant to emulate a researcher identifying likely matches with a keyword-based strategy, and then reading each and coding it according to whether the verb is a synonym of defeat and the subject (`ARG1`) is a Democrat. In order to increase the number of negative samples, I also draw a random sample $2,000$ additional sentences not already in the training data and "manually" code these as well.

The classifier is therefore trained on $8,057$ observations with $1,424$ positive cases, and must assign a positive label to as many of the remaining $1,659$ positive cases in

the unlabelled data as possible ($n = 842,999$). In this hypothetical exercise, I would "manually" code all positively labelled observations, thus the cost of false positives would be additional time spent coding.

I train two classes of machine learning classifier on the Term Frequency-Inverse Document Frequency (tf-idf) matrix representation of the training data: Random Forest (RF) and a support vector machine (SVM). The choice of these two classifiers is motivated by the use of tree-based models in political science research (Montgomery and Olivella 2018), the argument by D'Orazio et al. (2014) that SVMs are ideal for the high dimensionality and sparsity of text matrices, and because both models have well-established feature importance metrics. For each classifier, I train five models using 5-fold cross validation. From these five, I prioritize recall and then weighted F1 because the objective is to detect as many of the positive cases as possible. Based on the cross-validated training statistics, the SVM outperformed the RF: for RF the highest recall is 0.246 ($F1_{weighted} = 0.828$) while for SVM the highest recall is 0.512 ($F1_{weighted} = 0.873$).

I then use the best model from each class to extrapolate labels to the remainder of the cases. The results are summarized in a confusion matrix (Table 3.2). The columns indicate which model made the classification (SVM or RF), and whether it included the observation in the sample (1) or not (0). The rows indicate the "true" label; whether the semantic role-based method included the observation in the sample. Each cell contains the number of observations belonging to this combination of predicted and actual labels.

Both models perform exceedingly well in terms of raw accuracy; RF has an accuracy

Table 3.2: Confusion matrix for random forest and support vector machine classifiers.

	RF: 0	RF: 1	SVC: 0	SVC: 1
True: 0	840,788	552	838,087	3,253
True: 1	1,404	255	949	710

of 0.998 and SVM has an accuracy of 0.995. However, our interest is in recall and the rate of false negatives. Columns "SVM: 1" and "RF: 1" show the number of observations the model would have flagged for manual coding and the number of these flagged samples that would have been true positives, and row "True: 1" shows us the number of false negatives and true positives.

Based on these results, the Random Forest model would require manually coding fewer observations ($n = 807$) but would only detect 15.3% ($1 - \frac{1,404}{1659} = 0.153$) of the positive cases in the unlabelled dataset. In contrast, the SVM would require manually coding more observations ($n = 3,963$), but would also detect 42.9% of positive cases. Based on this analysis, we might prefer to use SVM for this kind of task.[11]

As noted, a challenge of a supervised machine learning classifier strategy is that the researcher cannot specify which latent features (in this case, meaning not explicitly in the set of features given to the model) the classifier should use to differentiate positive and negative cases. However, inspecting the most important features for each of the classifiers can give us some sense of the underlying logic of each classifier.

[11] Although I would not so far as to claim that SVM is better at detecting semantic role structures than RF, as this paper does not conduct sufficiently thorough analysis to show that this holds generally.

Table 3.3: Top five tokens for support vector machine by coefficient.

Token	SVM Coefficients
`defeat`	3.71
`crush`	3.15
`beat`	2.63
`defeating`	2.59
`beaten`	2.15

Support vector machines classify samples by constructing the hyperplane through the feature space that best partitions the space into homogeneously-labelled regions. The coefficients on the primal model are analogous to the "weights" of this hyperplane corresponding to features in the original space. Table 3.3 shows the tokens (features on the tf-idf matrix) with the five highest importances. Notably, all tokens are either "defeat", its synonyms or its conjugation. This indicates that the SVM "learned" to differentiate sentences most strongly on the verb.

Table 3.4: Top five tokens for random forest classifier by normalized Gini feature importance.

Token	Feature Importance (Gini)
`defeat`	0.0291
`democrats`	0.0123
`help`	0.0119
`socialist`	0.0113
`liberal`	0.0107

For the RF, we can use the normalized mean Gini decrease metric as a measure of the relative importance of input features.[12] Table 3.4 shows the top five tokens. In contrast to the SVM, the most important features of the RF are ones that seem substantively plausible for the task: "defeat" and "democrats" are the two most important tokens. The inclusion of "help" reflects a construction we see in many of the positively labelled samples seen earlier: "help us defeat the democrats (with your donation)".

I also consider sentences that both models flagged as false negatives and false positives. Table 3.5 lists two randomly drawn false negatives (FN) and two randomly drawn false positives (FP).[13] The false negatives appear to suffer from a similar challenge to the keyword-based strategy; not recognizing named entities as belonging to the Democratic party. This can likewise be ameliorated with the strategy suggested above. In contrast, the false positives exhibit semantically linked claims to defeating the Democrats; protecting Republicans and stopping Democrats (from winning). The plausibility of these results suggests that the iterative approach of King, Lam, and Roberts (2017) could be applied to semantic roles as well, to inductively identify the semantic roles that best match the latent concept of interest.

[12]Thesis-specific footnote: an explanation of the interpretation of these scores is contained in the Appendix 1D.

[13]Note on presentation: the use of fixed-width font in Table 3.1 indicated that the outputs were parses of the SRL model. In Table 3.5, I present the unprocessed original text of the emails, and therefore present it in regular font.

Table 3.5: Two false negatives and two false positives. Note: [...] indicates output has been truncated for presentation purposes.

Prediction	Text
FN	If our grassroots team to steps up, we'll have what we need to beat Pete.
FN	If we are going to defeat entrenched incumbent Eliot Engel and send Jamaal to Congress, we need everyone to pitch in and help.
FP	Stand with President Trump DONATE HERE to help me protect the President from the Radical Left
FP	Fran Flynn WE MUST STOP THEM! Please contribute 25, 50, [...]

Conclusion

As the quantities of readily available data for studying political campaigns and other social phenomena continues to increase, methods for selecting subsets will continue to grow in importance, and novel tools can broaden the scope of the questions we ask. Formalizations of language from linguistic theory offer to help us systematize the phenomena that we are searching for, and computational automations for detecting these structures provides the scalability needed for analyses of large corpora.

The analysis and comparisons of the previous section highlight that the semantic role approach is complementary to existing methodological work on observation selection in text data. The advantage of filtering the SRL representation of a corpus is that it permits querying at a "higher" level of linguistic complexity, based on the relations between entities. Nevertheless, constructing the queries is subject to the same (if not stronger) cognitive limitations that motivate the work of King,

Lam, and Roberts (2017). The patterns that sophisticated models detect play an important role in the *discovery* stage of research (Grimmer, Roberts, and Stewart 2021; Grimmer, Roberts, and Stewart 2022), and careful validation of results can mitigate the harmful or erroneous inferences.

More broadly, as with any classification task, there is a trade-off between transparency and complexity when choosing a method for selecting observations on the basis of text content. A more transparent inclusion/exclusion rule reduces the need for validating the behavior of the classifier because its decisions are predictable from the data. This is the strength of keyword-based approaches; we can succinctly describe the set of returned documents as "the set of documents containing (one or more of) the keyword(s)". The semantic role-based approach is also highly transparent, with a clear link between the decision rule and selected texts. However, the additional complexity of employing an automated tool for labelling semantic rules increases the need for validation, because itself is prone to error, and errors early on in a pipeline propagate to the rest of the analysis (Ash, Gauthier, and Widmer 2021). This increased requirement should not diminish the usefulness of a method, as the metric for tools is not the extent to which they can *replace* human coders, but instead augment or amplify them (Grimmer and Stewart 2013).

This work also serves to call attention to the complex and heterogeneous nature of the notion of "relevancy" and observation selection. In some cases relevancy is binary. In cases where the objective is to identify texts containing a particular linguistic phenomenon, such as King, Lam, and Roberts (2017) who track the words used by Chinese netizens to evade sponsors, then relevant samples are those that contain those words, and thus a keyword approach is appropriate. In others, such

as when tracking ideational flows via similarity of language (Grimmer 2010; Barnes and Jacobs-Harukawa 2022), relevancy is a continuous and multidimensional concept with a variety of operationalizations. Texts can be relevant because they pertain to the same topic, or because they argue for the same outcome, or because they take the same stance on an issue.

Looking ahead, this work calls for work on the ways in which we *represent* text data, itself an active field in natural language processing and artificial intelligence. Next steps in this agenda include building task-specific corpora for political science domains and questions. The biases exhibited in the computational approaches to semantics that I discuss in this paper are amplified by models, but originate in the training data. With careful corpus curation and validation, it may be possible to train language models with predictable forms of bias that are relevant for political science applications.

References

"About WordNet." n.d. Princeton University; Princeton University.

Ash, Elliott, Germain Gauthier, and Philine Widmer. 2021. "Text semantics capture political and economic narratives." *arXiv Preprint arXiv:2108.01720*.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data." *American Political Science Review* 113 (4): 883–901. https://doi.org/10.1017/S0003055419000352.

Barnes, Lucy, and Timothy Hicks. 2018. "Making Austerity Popular: The Media and Mass Attitudes Towards Fiscal Policy." *American Journal of Political Science* 62 (2): 340–54. https://doi.org/10.1111/ajps.12346.

Barnes, Lucy, and Musashi Jacobs-Harukawa. 2022. "Linking Economic Ideas and Narratives Across Corpora." European Political Science Association; Conference Paper.

Baum, Matthew A. 2013. "The Iraq Coalition of the Willing and (Politically) Able: Party Systems, the Press, and Public Influence on Foreign Policy." *American Journal of Political Science* 57 (2): 442–58. https://doi.org/10.1111/j.1540-5907.2012.00627.x.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Blumenau, Jack, and Benjamin E. Lauderdale. 2018. "Never Let a Good Crisis Go to Waste: Agenda Setting and Legislative Voting in Response to the EU Crisis."

The Journal of Politics 80 (2): 462–78. https://doi.org/10.1086/694543.

Boleda, Gemma. 2019. "Distributional semantics and linguistic theory." *arXiv Preprint arXiv:1905.01896*.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356 (6334): 183–86.

D'Orazio, Vito, Steven T. Landis, Glenn Palmer, and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22 (2): 224–42. http://www.jstor.org/stable/24573223.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Fong, Christian, and Justin Grimmer. 2019. "Causal inference with latent treatments." *American Journal of Political Science*.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. "Allennlp: A deep semantic natural language processing platform." *arXiv Preprint arXiv:1803.07640*.

Grimmer, Justin. 2010. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18 (1): 1–35.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24 (1): 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921.

Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Harris, Zellig S. 1954. "Distributional structure." *Word* 10 (2-3): 146–62.

Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39 (S1): 201–18. https://doi.org/10.1111/pops.12487.

Jerit, Jennifer. 2004. "Survival of the fittest: Rhetoric during the course of an election campaign." *Political Psychology* 25 (4): 563–75.

Jerit, Jennifer, Jason Barabas, and Toby Bolsen. 2006. "Citizens, Knowledge, and the Information Environment." *American Journal of Political Science* 50 (2): 266–82. http://www.jstor.org/stable/3694272.

Johansson, Richard, and Pierre Nugues. 2008. "Dependency-based semantic role labeling of PropBank." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 69–78.

King, Gary, Patrick Lam, and Margaret E Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61 (4): 971–88.

Lauderdale, Benjamin E, and Alexander Herzog. 2016. "Measuring political positions from legislative speech." *Political Analysis* 24 (3): 374–94.

Mathur, Arunesh, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. 2020. "Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle." https://electionemails2020.org.

Montgomery, Jacob M., and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–44. https://doi.org/10.1111/ajps.12361.

Pan, Jennifer, and Alexandra A. Siegel. 2020. "How Saudi Crackdowns Fail to Silence Online Dissent." *American Political Science Review* 114 (1): 109–25. https://doi.org/10.1017/S0003055419000650.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–43. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv Preprint arXiv:1908.10084*.

Ridout, Travis N, and Michael Franz. 2008. "Evaluating measures of campaign tone." *Political Communication* 25 (2): 158–79.

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis*, 1–25.

Shi, Peng, and Jimmy Lin. 2019. "Simple bert models for relation extraction and semantic role labeling." *arXiv Preprint arXiv:1904.05255*.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. "MPNet: Masked and Permuted Pre-training for Language Understanding." In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 33:16857–67. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

Weaver, Michael. 2019. "'Judge Lynch' in the Court of Public Opinion: Publicity and the De-legitimation of Lynching." *American Political Science Review* 113 (2): 293–310. https://doi.org/10.1017/S0003055418000886.

Appendix 3

Comparison to Ash, Gauthier, and Widmer (2021)

Ash, Gauthier, and Widmer (2021) present "an unsupervised method to quantify latent narrative structures in text documents". Below is a simplified step-by-step technical summary of their pipeline here. Beginning with a collection of text documents:

1. Split corpus into sentences using SpaCy's sentence boundary detection.
2. Apply the OpenIE semantic role labeller (SRL) from AllenNLP to sentences. Store as list of parses per sentence.
3. Simplify spans labelled with roles `ARG0`, `ARG1` and `ARG2` by extracting nouns, verbs, adjectives and adverbs.
4. Embed simplified `ARG0`, `ARG1` and `ARG2` spans using the Arora et al weighted word vector embeddings.
5. Cluster the embeddings using an (unweighted) `kmeans` model and Euclidean distance.
6. Label clusters based on highest frequency span assigned to cluster.
7. Reduce `V` role using most common synonym logic based on WordNet entries.
8. Filter parses to retain only complete $\langle ARG0, V, ARG1 \rangle$ narrative tuples.
9. (Plot 100-250 most common tuples in an interactive network graph.)

This pipeline combines a large number of statistical NLP models—sentence-boundary detection, SRL, POS-tagging, phrase embedding, and unsupervised

clustering—to produce descriptively rich, interpretable and useful summaries of the *claims* in a corpus. As the authors point out, the application of these narrative tuples is not limited to the creating summaries of the narratives present in Congressional speech, but can be used much more widely as an operationalization strategy, which is exactly what I do in this paper.

My pipeline modifies the choice of models, span simplification, and removes the clustering step entirely for reasons detailed in the main body of the paper.

1. Split corpus into sentences using SpaCy's sentence boundary detection.
2. Apply the OpenIE semantic role labeller (SRL) from AllenNLP to sentences. Store as list of parses per sentence.
3. Simplify spans labelled with roles `ARG0`, `ARG1` and `ARG2` by parsing the dependency tree to identify the head noun phrase, then extract and lemmatize this.
4. Embed simplified `ARG0`, `ARG1` and `ARG2` spans with a semantic similarity sentence transformer.
5. Use WordNet to convert `V` to its lemma.
6. Filter parses to retain only complete $\langle V, ARG0, ARG1 \rangle$ role sets.
7. Search sets.

Pre-Processing

In general, pre-processing is a trade-off between making observations comparable and losing information. My main aim is to annotate the existing text with addi-

tional information to facilitate its search. The most "destructive" step is the third, which includes stopword and punctuation removal and lemmatization. I justify these choices in two ways: firstly, I retain the non-lemmatized versions so that the eventual presentation of texts is original, and secondly the queries are lemmatized to help simplify matching.

Additional Analysis: Immigration

In a preliminary analysis, I combine SRL with semantic embeddings similar to Ash, Gauthier, and Widmer (2021) as a means of extracting descriptive *narratives* about immigration from the dataset. Table 3.6 shows the ten spans labelled as `ARG0` or `ARG1` with the highest cosine similarity as given by a general-purpose semantic similarity model (Reimers and Gurevych 2019; Song et al. 2020) to the prompt `immigrants`, along with the frequency with which they were labelled as a given role in the corpus. Note that the texts in the table have been parsed, lemmatized and filtered for stopwords in order to facilitate the counting of similar spans.

Table 3.6: Ten closest `ARG0` and `ARG1` spans to `immigrants`.

Rank	Processed Span	Similarity	ARG0 Freq	ARG1 Freq
1	`immigrant people`	0.879208	0	1
2	`Immigration`	0.794525	3	0
3	`immigration`	0.794525	6	52
4	`immigrant`	0.783041	67	123
5	`Refugees`	0.77411	0	1

Rank	Processed Span	Similarity	ARG0 Freq	ARG1 Freq
6	`undocumented Americans`	0.761743	5	7
7	`immigrant community`	0.76072	4	9
8	`immigrant those who they`	0.750108	0	1
9	`undocumented people`	0.734184	0	2
10	`immigrant community people`	0.733681	0	2

Although the first four entries of the table are relatively uninteresting and can be matched with a simple regular expression, the top ten also includes `Refugees`, `undocumented Americans`, and `undocumented people`. Although all of these entities are hyponyms of the query word (i.e. a more specific sense of the same concept), they are likely and suitable synonyms for studying the USA context.

In order to consider further examples that do not directly contain the same word stem as the prompt, Table 3.7 displays the ten most similar spans to `immigrants` that do not contain the stem `immig*`. These results display a result that exemplifies why a distributional approach to semantics in social science contexts must be considered critically: the model associates Mexicans and Latin Americans with immigration. Although it is very much not the case that `Mexicans` and `Latinos` are ontologically equivalent to `immigrants`, the embedding model (and most other NLP models) learn semantics based on the idea that words used in similar contexts are similar in meaning. In this case, this co-occurrence reflects the fact that speakers may be referring to immigrants when using the term "Mexican" or "Latino". If our interest is in how language is used, then it may make sense to include candi-

date matches in the downstream analysis, but then it becomes crucial to interpret this method as identifying the keyword and *relevant* entities, instead of semantically *equivalent* entities. Finally, it should be noted that this table also contains some unambiguous false positives: the inclusion of "New Mexicans", which refers to denizens of the state of New Mexico, and "AMERICANS", which does not refer to immigrants in the sense meant in this analysis. It may be the case that the latter is included as an antonym.

An interesting pattern appears in these tables. With the exception of these two false positives, entities relating to `immigrants` occupy `ARG1` much more frequently, which indicates that they are being referred to less often as agents and more often the subject of actions. This finding suggests that references to immigrants in campaigning emails are usually in the context of doing something *to* immigrants (which could be helping or harming or otherwise), but the lower rate of agency assignment is itself indicative of how immigrants are being discussed.

Table 3.7: `ARG0` and `ARG1` spans to `immigrants` that do not contain the stem `immig*`.

Rank	Processed Span	Similarity	ARG0 Freq	ARG1 Freq
13	`Mexicans`	0.722313	0	4
15	`migrant`	0.715708	4	6
16	`Mexican Americans`	0.714172	0	2
17	`Latinos`	0.712983	9	18
18	`undocumented young people`	0.708584	1	0
19	`undocumented folk`	0.707429	1	3
21	`New Mexicans`	0.703349	63	52

Rank	Processed Span	Similarity	ARG0 Freq	ARG1 Freq
27	undocumented New Yorkers	0.691626	0	2
30	Hispanics	0.690476	4	6
31	AMERICANS	0.689596	8	5

I manually select twenty entities from the top fifty, which I list in the appendix (under Lists). These match to 713 emails, which is the sample that this method returns. However, instead of using the full texts of emails, table 3.8 shows five randomly drawn narratives containing entities similar to the query `immigrants`. This shows the qualitatively rich representation provided by SRL sets at a level that can still be aggregated. A full table containing 384 unique and complete {action, agent, patient} sets is included in the supplementary materials.

Table 3.8: Five randomly drawn complete narratives containing an entity similar to `immigrants`.

ARG0	V	ARG1	Frequency
immigrant	pursue	american Dream	1
that	welcome	immigrant	1
ICE	release	immigrant	2
Trump	mistreat	immigrant	1
migrant	face	indefinite lock in solitary confinement	1

Conclusions

Although most of the findings and innovations of the papers in this thesis are methodological, the substantive and normative motivations for studying political campaigning remain. Having shown in the first paper that targeting can have an effect, I now ask the more crucial question; what can be done about targeting? In a follow-up study, I propose to employ the same template for simulating a targeted campaign, but additionally test two messages priming respondents to the targeting. One group receives a "bespoke" disclosure, informing them that the advertisement they are viewing has been tailored to be relevant to their interests. Another group receives a "warning" disclosure, informing them that the advertisement they are viewing has been chosen because it is more likely to persuade them. If the effect of targeting is weakened by the stronger disclosure, then we find an empirical basis for arguing that existing requirements are ineffective, and stronger regulation is required to preserve voter autonomy. The goal of this agenda is to inform empirically-guided regulation to reduce the harms associated with this technology.

An important methodological implication of the first paper is that targeting works because there is sufficient heterogeneity between the effect of advertisements on

individuals. This result requires further study in two directions; developing tools for describing and summarizing this heterogeneity to characterize the groups for which particular treatments have an effect, and developing a framework for identifying individual-level effects and detecting when a null result is due to over-aggregation. The experimental design also offers opportunities for further applications where the allocation rule needs to be complex and dynamically updating, such as a more efficient template for adaptive treatment designs.

The next steps for the second paper is the development of a hybrid computationally-augmented qualitative content analysis, with applications to both enhancing human intuition and discovering model bias. The method can also be applied to further applications, such as characterizing speaker and party style in parliamentary speech or identifying salient characteristics in political image and video data. More broadly the paper shows the utility of techniques for model interpretation from explainable machine learning, and offers a way forward for using non-linear models to understand the relationships between empirical phenomena.

The third paper emphasizes the importance of validation and transparency when using complex models for political science research, and demonstrates that theoretically-motivated approaches from linguistics have much to offer when trying to operationalize text data. Focusing on the task of document selection, the next step for this research is to combine the semantic role-based approach with the inductive lexicon building methods of King, Lam, and Roberts (2017). In terms of contributions political science methodologists can make to the broader discipline, there is a clear agenda for developing tools to curate domain-specific corpora that can be used to train large language models that exhibit predictable and non-harmful

forms of bias.

All three of these papers are influenced by, and contribute to a new interdisciplinary paradigm of computational social science research. I bring methods from machine learning and deep learning in order to flexibly describe, measure, and infer patterns in the complex and heterogeneous social processes that we study. The aim throughout is to confront the complexity of real-world phenomena with an inductive, empirically-guided approach.